l'université nantes angers le mans
COMMUNAUTÉ D'UNIVERSITÉS ET ÉTABLISSEMENTS

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Lambert PÉPIN

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le label de l'Université de Nantes Angers Le Mans

École doctorale : Sciences et technologies de l'information, et mathématiques (ED 503)

Discipline : Informatique et applications, section CNU 27
Unité de recherche : Laboratoire d'informatique de Nantes-Atlantique (LINA)

Soutenue le 21 octobre 2015

Fouille exploratoire de messages publiés sur Twitter pour l'aide à la décision

JURY

Rapporteurs :	**M^me^ Bénédicte LE GRAND**, Professeur des universités, Université Paris 1 **M^me^ Rim FAIZ**, Professeur des universités, Université de Carthage, Tunisie
Examinateur :	**M. Imed KACEM**, Professeur des universités, Université de Lorraine
Invité :	**M. Philippe SUIGNARD**, Ingénieur chercheur expert, EDF R&D
Directrice de thèse :	**M^me^ Pascale KUNTZ**, Professeur des universités, Université de Nantes
Co-encadrants de thèse :	**M. Fabrice GUILLET**, Professeur des universités, Université de Nantes **M. Julien BLANCHARD**, Maître de conférences, Université de Nantes

Remerciements

Avant tout, je tiens à remercier les membres du jury de m'avoir fait l'honneur de leur participation à ma soutenance de thèse. Je remercie Madame Bénédicte Le Grand et Madame Rim Faiz d'avoir accepté de rapporter cette thèse et je remercie Monsieur Imed Kacem d'avoir accepté d'être examinateur lors de cette soutenance. J'espère que vos conseils et encouragements seront aussi utiles à mon avenir que vos questions et remarques l'ont été pour ces travaux.

Cette thèse CIFRE s'est déroulée en partenariat avec le Laboratoire d'Informatique de Nantes Atlantique et la direction Recherche et Développement d'EDF Clamart, et je tiens à remercier tout autant mes encadrants académiques et industriels. Je remercie sincèrement ma directrice de thèse, Madame Pascale Kuntz, mon co-directeur de thèse Fabrice Guillet et mon co-encadrant Julien Blanchard pour leur accompagnement qui a su prouver tout au long de cette thèse que les kilomètres n'arrêtent pas les connaissances. J'adresse également mes plus sincères remerciements à mon responsable industriel, Monsieur Philippe Suignard, qui m'a fait confiance depuis le premier jour et sans qui cette thèse n'aurait jamais eu lieu. Vos compétences scientifiques et techniques ainsi que vos conseils théoriques et pratiques ont été d'irremplaçables atouts pour mener à bien ces travaux de recherche. Pour votre confiance, vos regards critiques, vos encouragements et votre disponibilité, je vous adresse ma plus sincère reconnaissance.

Je tiens également à remercier chaleureusement les membres des deux équipes au sein desquelles j'ai été amené à travailler pendant ces trois ans. Un grand merci aux membres du département Innovation Commerciale, Analyse des Marchés et de leur Environnement de la R&D d'EDF pour ces échanges très riches, sur des sujets toujours plus variés. J'ai eu la chance de travailler dans une ambiance enrichissante et détendue à laquelle vous n'êtes certainement pas étrangers. Je vous remercie tous pour ces échanges formels et informels, pour ces moments de détente et de réflexion : Chloé, Sophie, Vanesha, Anthony, Kevin, Louis, Mathieu, Rémi, Romain, Sylvain.

J'adresse une pensée particulière à Yvon Haradji dont l'expérience nous a permis de naviguer dans les méandres de l'administration et à son doctorant Thomas Huraux qui m'a accompagné sur le pont jour après jour.

Un grand merci également à l'équipe Data User Knowledge du LINA avec qui

j'ai eu la chance d'avoir des échanges qui, même s'ils étaient plus rares, n'étaient pas moins intéressants. Je vous remercie pour l'accueil que vous m'avez réservé au labo et en dehors. Je remercie tout particulièrement Fabien Picarougne, Anthony Coutant et Nicolas Greffard qui ont toujours répondu présent.

Je remercie du fond du cœur mes amis, et en particulier ceux que je ne vois pas assez : le Kartier, les sudistes, le quintette de tête et les clubbeurs bien sûr. Merci d'être là et merci de m'avoir accompagné à chaque étape de ma vie.

Enfin, je remercie ma famille, proche et éloignée ; cette thèse est un peu la votre. Pour finir, je remercie mes princesses, Balou, numéro bis et surtout Audrey pour son éternel soutien.

Table des matières

Liste des tableaux

Table des figures

Liste des symboles

$\mathcal{D}$	L'ensemble des tweets étudiés
$\mathcal{T}$	La période de temps correspondante
$\mathcal{T}_1, \ldots, \mathcal{T}_n$	Les périodes de 24h
$\mathbf{d} \in \mathcal{D}$	Un tweet
$t \in d$	Un terme
$\mathcal{V}$	Le vocabulaire correspondant à $\mathcal{D}$
$\mathcal{P}$	L'ensemble des motifs
$\mathbf{P} \in \mathcal{P}$	Un motif ($\mathbf{P} \subseteq \mathcal{V}$)
L_k	L'ensemble des motifs de longueur k
C_k	L'ensemble des motifs candidats de longueur k
f	La fréquence d'un motif
$\vec{ev}(t)$	Le vecteur d'évolution du terme t
s	La fonction de score
$N(t, i)$	Le nombre de tweets contenant le terme t publiés pendant la période $\mathcal{T}_i$
τ	Le coefficient de Kendall
M_S	La matrice de co-évolution (de taille $\mid \mathcal{V} \mid$)
K	Le nombre de thèmes
α	L'hyper-paramètre (pseudo-observations) pour les documents
β	L'hyper-paramètre (pseudo-observations) pour les thèmes
$\mathcal{W}$	L'ensemble des occurrences des termes dans les documents
$w \in \mathcal{W}$	Un mot dans un document (une instance particulière d'un terme $t \in \mathcal{V}$)

$k \in K$ Un thème

z Le thème associé à une instance particulière d'un mot dans un document

$\boldsymbol{\theta_d} = \{p(z = k \mid m = d)\}_{k \in \mathcal{K}}$ La distribution sur les thèmes du tweet d

$\boldsymbol{\phi_k} = \{p(w = t \mid z = k)\}_{t \in \mathcal{V}}$ La distribution sur les termes du thème k

$\Theta = \{\boldsymbol{\theta_d}\}_{d \in \mathcal{D}}$ L'ensemble des distributions des documents sur les thèmes

$\Phi = \{\boldsymbol{\phi_k}\}_{k \in K}$ L'ensemble des distributions des thèmes sur les termes

$n_k^{(t)}$ Le nombre d'instances du terme t associées au thème k

$n_d^{(k)}$ Le nombre d'instances du thème k associées au tweet d

$\mathcal{M}_i = (\mathcal{D}_i, \mathcal{V}_i, K, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$ Le modèle entraîné pour la période i

$\mathcal{V}_i^*$ Le vocabulaire représentatif de $\mathcal{M}_i$

$\mathcal{U}_{i,i+1}$ Le vocabulaire commun aux modèles $\mathcal{M}_i$ et $\mathcal{M}_{i+1}$

$\triangleright$ La relation "évolue en"

1

Introduction

The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it. It is a very useful assumption under many circumstances, but one too easily forgets that it is false.

—Alan Turing

La massification des usages des médias sociaux impacte aujourd'hui les stratégies décisionnelles des entreprises fortement attachées à la relation client. Des études en intelligence décisionnelle (e.g. [PBS11]) pointent la nécessité d'introduire des processus d'analyse du contenu de ces médias dans les nouveaux systèmes d'aide à la décision pour aider les décideurs à mieux cerner les tendances et les opinions des consommateurs. De fait, l'intérêt pour ces informations stimule l'émergence d'une nouvelle génération de systèmes de gestion de la relation client ("Customer Management Systems") [AAN+13].

Le groupe EDF, dans lequel s'est effectuée cette thèse, est un exemple paradigmatique de cette évolution. Dans un contexte de prix bas sur le marché de l'énergie avec une concurrence accrue et une pression réglementaire renforcée, la fidélisation de ses clients est un enjeu majeur. Elle repose en amont sur une compréhension fine de leurs comportements. C'est dans cet environnement d'amélioration continue de la relation client du groupe que le Domaine Analyse et Connaissance Client (DACC) de la Direction Commerce, soutenu par le département ICAME (Innovation Commerciale, Analyse des Marchés et de leur Environnement) de la Direction Recherche et Développement, analyse les préoccupations de ses clients. La compréhension du langage naturel est au cœur des besoins de ces équipes qui analysent déjà

de nombreuses données issues des réclamations clients, des réponses aux questions ouvertes des enquêtes de satisfaction ou des données saisies par les conseillers. Toutefois, de nouveaux canaux prennent de plus en plus d'importance et de nouvelles approches doivent être développées pour analyser ce que dit le client, à la fois sur des outils conversationnels interactifs (e.g. la conseillère virtuelle, Laura, présente sur le site Bleu Ciel d'EDF[1]), par mail et sur les médias sociaux grand public.

1 Extraction d'informations "pertinentes" pour l'entreprise sur Twitter

Dans cette thèse, nous nous sommes focalisés sur les informations concernant le groupe EDF diffusées sur Twitter qui est dans le palmarès des dix sites les plus utilisés du Web[2] et qui joue un rôle important dans la diffusion des opinions. Cependant, comme l'ont noté Gundecha et Liu [GL12] dans un panorama sur la fouille de média sociaux, "*analyser les opinions qui circulent à propos d'une entreprise [...] est un nouveau défi : les données issues des médias sociaux sont volumineuses, bruitées, distribuées, non-structurées et dynamiques*". L'importance et la difficulté de la problématique ont conduit au développement de très nombreux travaux ces dernières années et des efforts significatifs ont notamment porté sur les questions de complexité de traitement de l'information liées au passage à l'échelle des données manipulées. A titre d'illustration, les derniers chiffres officiels font état de 500 millions de tweets publiés chaque jour[3]. Si l'efficacité des algorithmes s'est effectivement accrue, les informations utiles extraites concernent essentiellement les opinions les plus populaires à un instant donné. Cet effet est particulièrement marqué pour Twitter dont la dynamique temporelle des thématiques les plus populaires est constituée d'une succession de pics (voir figure 1.1). Cependant, en sus de l'information saillante, les entreprises ont besoin également d'analyser l'évolution des sujets qui les concernent au cours du temps de façon à détecter des "signaux faibles" qui apparaissent à des intervalles plus irréguliers. Ils représentent des fragments d'information à haute valeur ajoutée, cachés dans la masse des flux de messages, dont l'importance croissante a été récemment soulignée par des experts en intelligence stratégique [HMS14].

La problématique générale n'est pas récente. Bien avant Twitter, elle était déjà présente chez H.P. Luhn, souvent considéré comme l'inventeur du terme "Business Intelligence", qui dans son article précurseur "A Business Intelligence System" [Luh58] cherchait à fournir des services d'alerte pour les scientifiques et les ingé-

[1] https ://particulier.edf.fr/fr/accueil.html
[2] http ://www.alexa.com/topsites
[3] business.twitter.com/fr/basics/learn-twitter

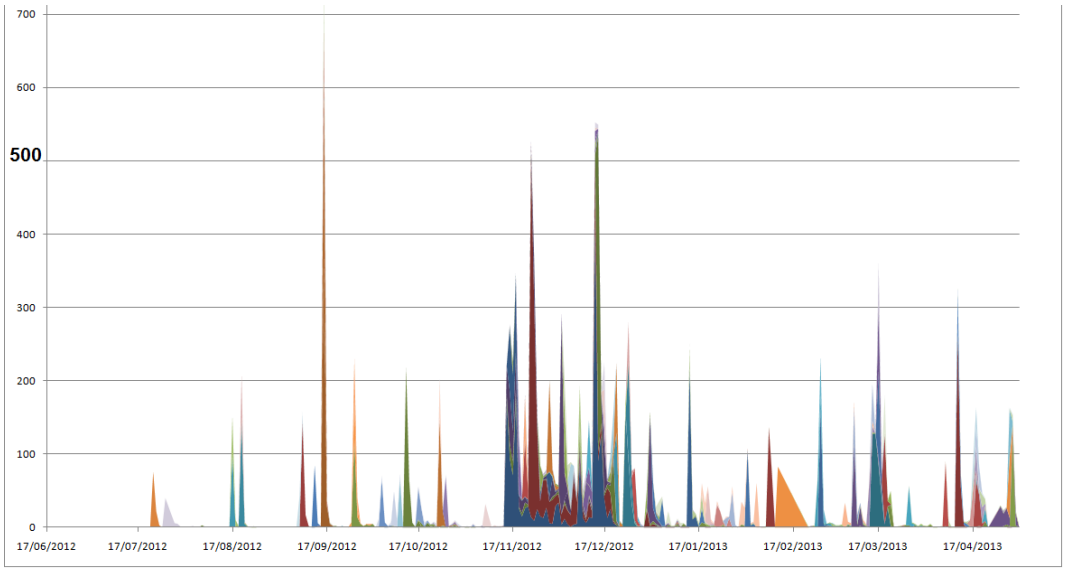


FIGURE 1.1 – Graphique en aires empilées représentant la répartition dans le temps du nombre de retweets des 255 tweets les plus populaires. Chaque pic correspond à un tweet et ses retweets calculés par regroupement hiérarchique tronqué (voir section 3.4).

nieurs, non pas à cette époque sur les opinions des clients, mais sur la littérature scientifique et technique. Mais les données d'aujourd'hui ont profondément changé de nature. Les données de Twitter sont des textes informels, courts et bruités, de 140 caractères maximum, qui contiennent de nombreuses fautes d'orthographe et de multiples innovations linguistiques [NGKA11]. S'ajoutent aux termes, des éléments spécifiques : 26% des tweets contiennent une URL, 17% un #hashtag (label inclus dans le tweet pour organiser les conversations) et 55% une @mention (référence à un utilisateur pour porter le tweet à son attention) [CWT13]. En augmentant le nombre d'unités syntaxiques distinctes, ces particularités renforcent l'aspect parcimonieux de la représentation des documents et il est alors nécessaire de questionner la pertinence des approches classiques tant la faible co-occurrence des termes dans les documents limite l'inférence de connaissances sémantiques [RDL10, HD10]. De plus, contrairement aux sources classiques organisées, tels que les journaux, la télévision et la radio, les tweets couvrent un très large panel de sujets créés pour des actions variées : bavardages, partages d'informations, commentaires d'événements, conversations entre amis par des acteurs hétérogènes [JSFT07]. En outre, ici, l'information ne suit pas un fil unidirectionnel (du producteur au consommateur) et les utilisateurs, qui s'appuient sur des plateformes interactives pour créer, partager et modifier du contenu auto-généré, peuvent être à la fois producteurs et consommateurs. Cette spécificité favorise la publication de messages moins formatés, couvrant un éventail de sujets plus variés [HTK13, PGMJ11]. La grande variabilité dans le vocabulaire employé peut être mise en avant par l'observation du nombre de mots nouveaux par jour, c'est-à-dire, du nombre de mots jamais employés jusque là dans

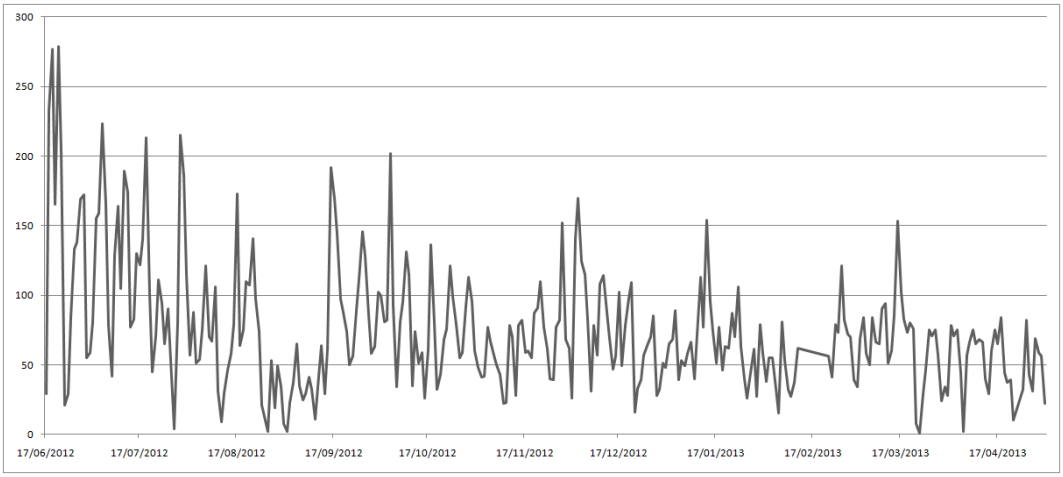


FIGURE 1.2 – Évolution du nombre de mots nouveaux par jour. Même si la tendance générale est à la baisse, le vocabulaire continue à se renouveler.

le corpus. Comme l'illustre la figure 1.2, même si la tendance générale est à la baisse, le vocabulaire continue à se renouveler. En conséquence, les méthodes d'extraction de thématiques cherchant à produire des connaissances exploitables doivent faire le tri entre les informations importantes et le bruit sémantique.

Une autre particularité, bien connue des données actuelles, est le volume toujours croissant disponible à l'analyse. Accompagnant la popularité croissante de Twitter, les volumes de données produits battent régulièrement de nouveaux records[4]. Les approches traditionnelles n'ont pas été conçues pour traiter un flux de données aussi conséquent et la capacité à monter en charge représente maintenant un enjeu majeur pour les méthodes souhaitant modéliser le flux Twitter.

Enfin, à ces caractéristiques, s'ajoute la dimension temporelle, puisque l'information partagée est étroitement liée à la date de sa publication et l'intérêt du public évolue très rapidement. En effet, Twitter a été conçu pour analyser, stocker et mettre à disposition les tweets en temps-réel [BGL+12] et les "Trending Topics" (sujets populaires définis selon un algorithme interne à Twitter) changent très rapidement : 73% des Trending Topics ont une seule période d'activité et 31% des périodes ne durent qu'une journée [KLPM10]. Par ailleurs, une étude récente a montré que les hashtags accumulaient, en moyenne, plus de 50% de leurs occurrences dans les deux premières heures suivant leur publication [KCLC13]. Dans ce contexte de forte nouveauté, le nombre de thématiques détectées peut rapidement devenir ingérable tant un analyste humain sera dans l'incapacité de traiter l'ensemble des résultats. Ainsi, les méthodes d'extraction de thématiques sont amenées à sélectionner avec soin le niveau de granularité des informations produites tout en se concentrant sur l'évolution temporelle des thématiques afin de proposer la bonne quantité d'informations à l'analyste. Outre la volumétrie mentionnée précédemment, la spécificité de ces

[4] blog.twitter.com/2013/new-tweets-per-second-record-and-how

données renouvelle des questions majeures de traitement de l'information :

- Quel est le degré de granularité des données pertinent pour l'extraction de "connaissances sémantiques" utiles pour la prise de décisions ?
- Quelle est l'échelle temporelle pertinente pour la détection de ces connaissances et de leur évolution ?

Pour résumer, les approches d'extraction de thématiques développées pour l'analyse des données Twitter doivent faire face à quatre difficultés majeures :

1. le bruit syntactique
2. le bruit sémantique
3. le volume des données
4. la forte nouveauté

Sans modèle préalable sur les données de Twitter concernant le groupe EDF, nous nous sommes initialement placés dans le cadre de la fouille exploratoire de données en analysant les combinaisons de termes fréquents (motifs fréquents) puis la co-évolution des termes. Les premiers résultats ont permis d'extraire quelques « pépites » non triviales qui ont validé l'hypothèse de l'existence de « signaux faibles » cachés dans la masse. Mais ces approches se sont heurtées à l'explosion combinatoire des combinaisons de termes. Et le contournement du problème par le recours à des filtres basés sur des mesures de qualité ou des fonctions de score a conduit à l'extraction de grandes tendances et la perte des fragments d'information potentiellement plus riches. Cette démarche exploratoire, menée pour l'interprétation des résultats en collaboration étroite avec un expert de l'entreprise, nous a aidés à mieux cerner le degré de granularité pertinent pour l'analyse.

La deuxième partie de notre travail a donc porté sur l'analyse de l'évolution des thèmes ("topics") associée aux tweets qui se place à un niveau d'abstraction intermédiaire entre les termes individuels et les documents dans leur ensemble. Des premières expérimentations sur quelques périodes temporelles pré-fixées ont confirmé la pertinence de l'approche "Latente Dirichlet Allocation" (notée LDA dans la suite) [BNJ03]. Toutefois, il ne s'agit pas seulement d'extraire à chaque intervalle de temps les thèmes les plus populaires mais d'analyser l'évolution des relations entre ces thèmes pour détecter notamment des comportements « surprenants » eu égard aux connaissances d'un expert. Les difficultés de modélisation des connaissances expertes sont bien connues en ingénierie des connaissances (problèmes de disponibilité, difficulté de restitution verbale de concepts, etc) et nous n'y avons pas échappé. Ainsi, au lieu d'aborder de front une modélisation préalable

difficilement envisageable, nous avons choisi d'intégrer l'expert dans le processus de découverte via un support visuel interactif.

L'intérêt de la visualisation pour faciliter le traitement des informations complexes était déjà reconnu par des statisticiens au XIXème siècle [Che78] et l'essor aujourd'hui de la visualisation de l'information [KMSZ09] s'ancre dans une longue tradition de valorisation des données (e.g. la "graphique" de Bertin [BER67] ou l'analyse de données exploratoire de Tukey [Tuk77]). Le récent panorama de T. Schreck et D. Keim [SK13] illustre le potentiel de la fouille visuelle ("visual analytics") pour l'analyse de données des médias sociaux. Stimulés par cette voie de recherche prometteuse, nous avons développé un processus de visualisation interactif de l'évolution temporelle des relations entre les thèmes qui combine un modèle dynamique d'extraction de thèmes (Dynamic-Lda) et une restitution visuelle basée sur un modèle de graphe inspiré par les diagrammes de Sankey [San96]. Les diagrammes de Sankey ont été proposés à l'origine pour visualiser des flots d'énergie puis étendus à la visualisation de relations dynamiques dans des systèmes. Leur appropriation en entreprise est favorisée par leur popularité dans les environnements industriels. Un prototype intégrant ces travaux a servi de preuve de concept et nous a permis de mener des expérimentations sur des données réelles concernant le groupe EDF.

2 Plan du manuscrit

La suite de ce manuscrit est organisé en six chapitres :

Le **chapitre 2** présente un état de l'art des méthodes d'extraction de thématiques appliquées aux données Twitter. Les approches sont organisées autour de quatre tâches : la détection et le résumé d'événements, et la détection de nouvelles tendances et le suivi de tendances. Ces approches sont examinées en regard des difficultés majeures caractéristiques de l'analyse de tweets : le bruit sémantique, le volume des données et la forte nouveauté. Une synthèse décrit pour chaque tâche l'adéquation des méthodes proposées avec les spécificités des données Twitter.

Le **chapitre 3** décrit plusieurs approches d'apprentissage non supervisé mises en place en collaboration avec un expert de l'entreprise dans un contexte de fouille exploratoire. Puisque nous ne disposions pas de connaissances *a priori* sur les données manipulées, nous avons mis en place différentes approches dans le but d'évaluer le type d'informations valorisables qu'elles contiennent. Nous présentons tout d'abord le processus de collecte et de pré-traitements, puis nous décrivons deux approches classiques : le clustering de textes et l'extraction de motifs fréquents. Pour finir, nous présentons une approche originale pour l'analyse de la co-évolution des termes dans un corpus dynamique.

Le **chapitre 4** est consacré à l'extraction de thématiques par un modèle LDA. Nous décrivons dans un premier temps le processus générateur du modèle classique et une méthode d'estimation des paramètres. Dans un second temps, nous présentons une variante adaptée aux corpus dynamiques permettant de suivre l'évolution temporelle des thèmes extraits et nous comparons empiriquement sept mesures de divergences entre distributions de probabilités utilisées pour détecter des continuités dans leur évolution.

Dans le **chapitre 5**, nous présentons un outil de fouille interactif, développé en partenariat avec un ingénieur de recherche du Laboratoire d'Informatique de Nantes Atlantique, permettant à un analyste de visualiser et d'explorer l'évolution temporelle de thématiques. Ce chapitre débute par un bref état de l'art des méthodes de visualisation de données textuelles, puis il décrit notre outil de visualisation basé sur les diagrammes de Sankey ainsi que les interactions proposées à l'analyste. Par la suite, nous comparons les visualisations produites selon différentes mesures de divergence, et nous argumentons sur le choix de la mesure la plus adaptée. Finalement, puisque le nombre de thèmes, défini *a priori* pour chaque modèle LDA, apparaît comme étant un paramètre crucial de notre méthodologie, nous proposons d'intégrer le modèle non-paramétrique "Hierarchical Dirichlet Process" (HDP), qui estime automatiquement le nombre de thèmes à partir des données.

Finalement, le **chapitre 6** présente les conclusions et perspectives et clôt ce manuscrit.

2

Extraction de thématiques dans Twitter : un état de l'art

Questions that pertain to the foundations of mathematics, although treated by many in recent times, still lack a satisfactory solution. Ambiguity of language is philosophy's main source of problems. That is why it is of the utmost importance to examine attentively the very words we use.

—Giuseppe Peano

Sommaire

1 Introduction

L'extraction de thématiques ("Topic Modeling") est une tâche consistant à découvrir (automatiquement) la structure sémantique sous-jacente d'un ensemble de documents [BL09]. Popularisée par l'initiative Topic Detection & Tracking [ACD+98] de la DARPA[1], l'extraction de thématiques était appliquée à l'origine aux articles de presse dans un contexte de recherche d'information [FDGM99]. Les thématiques extraites ont depuis montré tout leur intérêt pour résumer, filtrer et organiser une collection de documents [CZYZ12]. Toutefois, même si l'extraction automatique du contenu sémantique de documents non-structurés offre une formidable opportunité pour extraire des informations à forte valeur ajoutée du flux Twitter, l'application directe des méthodes historiques est rendue particulièrement complexe en raison des spécificités du contenu des tweets.

Dans ce chapitre, nous proposons une organisation de la littérature en quatre grandes tâches. Dans la section 2.2, nous commençons par justifier de l'intérêt de cette organisation en montrant la diversité des définitions proposées pour les notions d'événements et de tendances. Nous décrivons, par la suite, les approches de détection d'événements dans la section 2.3, de résumé d'événements dans la section 2.4, de détection de nouvelles tendances dans la section 2.5 et de suivi de tendances dans la section 2.6.

2 Événements et tendances dans un flux Twitter

L'analyse des messages publiés sur Twitter répond à différents objectifs qui sont généralement exprimés au travers de deux concepts sous-jacents : les "événements" (events) pour l'analyse statique et les "tendances" (trends) pour l'analyse dynamique.

De nombreuses définitions du concept d'événement ont été proposées dans la littérature [DJY11, REC+12, MMJ13], la plus complète étant certainement celle de Dou et al. [DWS+12]. Dans ces travaux, les auteurs combinent plusieurs paradigmes et proposent de décrire un événement à travers quatre attributs : le sujet, le lieu, la date et les protagonistes, correspondant aux quatre questions classiques de l'enquêteur : qui ? quoi ? où ? et quand ? Par ailleurs, il est également nécessaire de distinguer les événements en deux types en fonction de leur origine (endogène ou exogène), car, même si cela est parfois sous-entendu, la plupart des méthodes se concentrent sur l'analyse d'événements, dit du "monde réel". Les événements de ce type sont motivés par des stimulus externes tels qu'un débat public, une confrontation sportive ou la sortie d'un nouveau produit et pour eux, la définition précédente est parfaitement adaptée. Ils diffèrent des événements, dits en-ligne (parfois ap-

[1] www.darpa.mil/

pelés "Twitter memes"), qui trouvent leur source au sein du réseau d'utilisateurs (endogène) et ne rentrent pas toujours dans le cadre de cette définition. Pour eux les concepts de lieu, de date ou de personnalité ne sont pas toujours adaptés, par exemple, dans le cas du "meme" #FollowFriday qui encourage les utilisateurs à publier tous les vendredis une liste de comptes à suivre, le concept de lieu ne paraît pas adapté. La réalité de cette distinction est mise en avant par les travaux de Cui et al. [CZL+12] qui proposent un classifieur à deux classes : "monde-réel" et "Twitter meme" permettant de séparer les hashtags uniquement sur la base du nombre de mots qu'ils contiennent et leurs positions dans les tweets. En fonction des besoins opérationnels, il est alors nécessaire de prendre en compte cette différence de nature, car les approches proposées dans la littérature sont souvent inadaptées à l'étude des événements "en-ligne". Enfin, dans le cadre de l'analyse des signaux faibles, il peut également être intéressant de distinguer les événements dont la portée est globale, qui bénéficient d'une large audience et dont les effets sont facilement détectables, et les événements locaux, qui peuvent être sous-représentés dans les médias traditionnels malgré un fort intérêt au sein de la communauté Twitter [NBG11].

En complément, l'étude des tendances, s'intéresse à l'évolution temporelle des thématiques et est particulièrement adaptée aux flux de textes dynamiques. Cette évolution peut correspondre à plusieurs facteurs. Tout d'abord, l'intérêt du public pour une thématique peut évoluer au cours du temps menant à l'apparition ou la disparition de thématiques. Ensuite, plusieurs thématiques peuvent fusionner lorsque leurs sujets se rejoignent (par exemple lorsque deux équipes sont amenées à s'affronter dans une compétition sportive) ou à l'inverse une thématique peut se scinder en sous-thématiques (par exemple lorsque plusieurs volets d'une affaire judiciaire sont discutés). Finalement, le vocabulaire employé pour traiter une thématique peut également évoluer au cours du temps même si le sujet de fond reste le même. C'est par exemple le cas lorsque le public s'approprie les éléments de communication d'un candidat à une élection lors des débats qui enflamment les médias sociaux.

De façon à mettre en regard la spécificité des données et les besoins des analystes, nous proposons dans ce chapitre, une organisation de la littérature en quatre grandes tâches de l'extraction de thématiques :

1. la détection d'événements
2. le résumé d'événements
3. la détection (précoce) de nouvelles tendances
4. l'analyse des tendances

Globalement, les travaux s'intéressant à l'étude des événements sur Twitter (détection et résumé) cherchent à situer les événements dans le temps et dans l'espace : les événements détectés correspondent à des instantanés du réseau pris à

un moment donné. Plus précisément, ces travaux se divisent en deux groupes : les premiers concernent la détection d'événements et cherchent à sélectionner les documents correspondant à une période et un sujet précis alors que les seconds s'apparentent au résumé d'information : étant donné un ensemble de documents pertinents, ils cherchent à répondre aux quatre questions de l'investigation (qui ? quoi ? où ? quand ?) en détectant les personnalités impliquées, en identifiant la période et le lieu de l'événement et en modélisant le sujet abordé. La détection de nouvelles tendances cherche à séparer les thématiques émergentes de l'ensemble des thématiques connues dans des délais s'approchant du temps réel et met l'accent sur la capacité des méthodes à passer à l'échelle afin de pouvoir traiter l'ensemble du flux. Et l'analyse des tendances cherche à suivre dans le temps les thématiques détectées et à modéliser l'évolution de l'intérêt des utilisateurs.

3 Détection d'événements

La détection d'événements peut être considérée comme la première étape de l'analyse des événements : son objectif est de sélectionner les documents pertinents parmi un flux de tweets et de cibler les termes discriminants. La seconde étape, qui consiste à valoriser le contenu de ces documents, n'est en général pas adressée par les travaux de ce domaine, ce qui justifie la distinction entre détection d'événements et résumé d'événements.

Les travaux en détection d'événements sur Twitter peuvent être divisés en deux catégories : les premiers se concentrent sur le contenu des documents et appliquent, entre autres, des méthodes de reconnaissance d'entités-nommées pour sélectionner les tweets en fonction des noms de lieux ou de personnalités qu'ils contiennent ("document-pivot"). Les seconds tirent parti de la grande réactivité du système et des utilisateurs en appliquant des méthodes de détection de pics à des signaux construits à partir des tweets pour détecter des périodes d'effervescence sur un sujet donné ("feature-pivot").

3.1 Reconnaissance d'entités nommées

Les méthodes de reconnaissance d'entités nommées ("Named-Entity Recognition" - NER) [JM00] permettent, grâce à des techniques linguistiques basées sur la grammaire ou à des modèles statistiques, d'extraire des références à des personnalités, des organisations ou des lieux et d'identifier les parties informatives d'un tweet. Par exemple, l'outil EDDI [BSH+10] permet d'organiser le fil d'actualité d'un utilisateur en thématiques construites à partir des phrases nominales contenues dans les tweets. Dans un premier temps, les phrases nominales sont extraites par un outil

d'analyse morpho-syntaxique[2] puis elles sont envoyées à un moteur de recherche externe[3] pour enrichir les tweets avec des termes populaires trouvés sur le web. En effet, les documents pertinents contiennent des termes connexes qui sont extraits à l'aide d'une mesure de pondération de type TF-IDF [Jon72]. Ces termes servent ensuite à labelliser les tweets avec des thématiques parfois sous-entendues. Cette approche permet de filtrer les flux de textes, souvent volumineux, reçus par les utilisateurs en se concentrant sur des marqueurs sémantiques particuliers. De manière similaire, l'approche STED [HCZ+13] cible les motifs de la forme "entité-nommée + verbe d'action" pour transférer des événements identifiés manuellement dans des articles de presse vers les tweets. Les motifs extraits des articles sont utilisés pour filtrer les tweets traitant du même événement et ceux-ci sont ensuite regroupés pour former les thématiques. Cette approche semi-supervisée permet de détecter des événements sur Twitter à condition que l'utilisateur soit capable de fournir des documents relatifs à un centre d'intérêt particulier. Toutefois son application est fortement limitée lorsque celui-ci n'a pas de connaissance *a priori* sur les données.

C'est justement pour résoudre ce problème que le système TwiCal [REC+12] a été proposé. En effet, celui-ci ajoute la détection d'expressions temporelles à l'approche précédente pour détecter automatiquement les événements dans les tweets. Les termes désignant une période temporelle, tels que "lundi" ou "demain", sont mis en correspondance avec leurs dates effectives puis les événements significatifs sont détectés en mesurant la co-occurrence dans les tweets des entités nommées et des expressions temporelles. Les résultats de cette approche reposent sur la présence d'expressions temporelles dans les données, et la collecte se fait par mots clés du type "aujourd'hui" ou "demain". Cette contrainte importante limite toutefois la généralité de l'approche qui se retrouve cantonnée à un type de données particulier.

3.2 Détections de pics

Les algorithmes de détection de pics offrent une approche complémentaire pour traiter les données Twitter sujettes au bruit sémantique. En effet, ces algorithmes permettent de localiser les événements dans le temps et se concentrent ainsi sur les périodes pertinentes du flux. Les dates de début et de fin des pics fournissent des bornes temporelles aux événements détectés et permettent de cibler les documents publiés pendant ces périodes. Dans le contexte de la détection d'événements, les algorithmes de détection de pics sont souvent appliqués à des signaux construits à partir de l'évolution temporelle du nombre d'occurrences d'unités syntaxiques telles que les mots, les N-grammes, etc.... Les pics sont ensuite regroupés pour former des thématiques. Dans ce cadre, les approches pour regrouper les pics peuvent se

[2] nlp.stanford.edu/software
[3] developer.yahoo.com/search/boss

diviser en deux catégories : "document-pivot" et "feature-pivot" [FYYL05]. Les premières se basent sur la co-occurrence des unités syntaxiques dans les documents et permettent de distinguer différents événements ayant lieu en même temps. Les secondes se basent sur la forme des pics et distinguent des événements pouvant utiliser le même vocabulaire mais se déroulant à des moments différents.

Le système Twevent [LSD12] détecte les pics dans le nombre d'occurrences d'unités syntaxiques appelées "cohesive segments" correspondant à des N-grammes statistiquement très probables [LWH+12]. La probabilité d'apparition des tweets est modélisée par une distribution binomiale et les segments saillants sont définis comme ceux présentant une probabilité anormalement élevée. Le système regroupe les pics détectés avec l'algorithme des k-plus proches voisins [JP73] en calculant la similarité entre chaque paire de segments saillants à partir de leurs co-occurrences dans les tweets (document-pivot). Finalement, la valeur informative ("newsworthiness") des événements est estimée à partir de leur probabilité d'apparition en tant que lien ("anchor text") dans un article de la Wikipédia. Ce mécanisme, faisant appel à des ressources externes, permet au système Twevent de traiter le bruit sémantique en filtrant les événements parasites, mais en contre-partie, cela limite son intérêt dans un contexte de forte nouveauté car aucun des segments saillants correspondant à un événement inconnu ne sera présent dans un article de la Wikipédia.

Toutefois, l'utilisation de ressources externes est coûteuse en temps de calcul, et des approches alternatives ont été proposées. En partant de l'observation que la moitié des segments détectés par Twevent sont des bigrammes [LSD12], le système "Event from Tweets" (ET) [PK13] réduit le temps de calcul en appliquant directement l'algorithme des k-plus proches voisins aux bigrammes. Ainsi il n'est pas nécessaire d'avoir recours à un historique de co-occurrences. De plus, ET propose de supprimer le recours à la Wikipédia en triant les événements détectés dans l'ordre décroissant du nombre de segments qui les composent. Dans cette approche, la réduction du temps de calcul se fait au détriment de la sensibilité au bruit sémantique : en effet, les événements triviaux ne sont plus filtrés, ce qui oblige l'analyste à parcourir un nombre plus important de résultats, ne présentant parfois pas d'intérêt opérationnel. En conséquence, une autre alternative a été proposée par Weng et Lee [WL11]. Leur système EDCoW propose de détecter des événements dans Twitter en appliquant une méthode avancée d'analyse du signal. Afin de réduire les coûts en espace et en temps, une décomposition en ondelettes [D+92] est appliquée aux signaux construits à partir de l'évolution temporelle du nombre d'occurrences des termes. Les termes triviaux sont filtrés en analysant l'auto-corrélation des signaux correspondant et la corrélation croisée ("cross-correlation") entre chaque paire de signaux restant est calculée afin de former une matrice de contingence. Dans un second temps, les événements sont construits en appliquant à cette matrice un algorithme de partitionnement basé sur la modularité (feature-pivot)[New06]. Fina-

lement, un score de pertinence est attribué à chaque événement en fonction de la corrélation croisée des mots qui le composent. Cette approche entièrement automatique et développée dans le but d'analyser un volume important de tweets présente l'avantage de détecter des événements à forte valeur informative qui correspondent aux descriptions des médias traditionnels.

À la suite d'un algorithme de détection de pics des méthodes de partitionnement peuvent être appliquées pour organiser les événements. Ces méthodes sont généralement automatiques. Toutefois, dans un contexte fortement bruité tel que Twitter, il peut être intéressant d'intégrer l'expertise humaine de façon à améliorer la qualité des groupes construits, et donc des événements détectés. Dans ce contexte, le système TwitterStand [SST+09] propose de transformer Twitter en un agrégateur de contenu dont les contributeurs ne sont pas connus à l'avance. Une liste d'utilisateurs dont les centres d'intérêt incluent les sujets d'actualité (comptes officiels des journaux et des chaînes de télévision, blogueurs influents) est mise à jour de manière dynamique en supprimant les utilisateurs inactifs et en ajoutant de nouveaux contributeurs. Les tweets de ces utilisateurs sont rassemblés et un classifieur Bayésien naïf [Mit97] est entraîné sur un jeu de données annoté afin d'assurer la valeur informative des événements détectés en sélectionnant les tweets traitant de sujets d'actualité. Dans un second temps, un algorithme de regroupement dynamique (Incremental Clustering) [DHS12] est appliqué pour regrouper les tweets en événements par une variante de la similarité cosinus [SKK+00]. Cette approche ne parcourt chaque tweet qu'une fois de manière à limiter la complexité algorithmique et le recours à des contributeurs officiels permet de cibler des tweets à forte valeur informative.

Une approche similaire est proposée dans RW-Event [BNG11] qui applique un algorithme de regroupement dynamique aux tweets en comparant (en termes de similarité cosinus) chaque nouveau tweet aux clusters existants. Si aucune similarité n'excède le seuil prédéfini, un nouveau cluster est créé. Afin d'accélérer le procédé, le système ne compare les nouveaux documents qu'aux centroïdes des clusters existants. Finalement, un classifieur SVM [WF05] basé sur des propriétés temporelles, sociales et sémantiques sépare les clusters correspondant à un événement du monde réel de ceux correspondant à du bruit. Puisque le nombre de thématiques présentes dans le corpus est inconnu, le nombre de clusters créés par cette approche n'est pas borné. Une limite de cette approche est l'explosion potentielle du nombre de comparaisons nécessaires à la classification d'un tweet.

3.3 Discussion

Dans le tableau 2.1, nous résumons les méthodes de détection d'événements en mettant l'accent sur trois propriétés clés dans le contexte de Twitter : le bruit séman-

tique, le volume des données et le contexte de forte nouveauté. Elles sont présentées par catégories, dans l'ordre de citation dans ce chapitre. Bien que les méthodes de reconnaissance d'entités nommées permettent de sélectionner des tweets pertinents en ciblant des termes informatifs et s'accommodent ainsi du bruit sémantique inhérent à Twitter, les nombreuses étapes de traitement additionnelles et les appels récurrents à des sources externes sont trop coûteux en temps pour permettre une généralisation des approches à de gros volumes de données. Néanmoins, même s'il semble contradictoire de vouloir consolider les événements par le recours à des sources externes tout en maintenant une faible complexité algorithmique, une sélection minutieuse des documents et la compression des données permettent une détection efficace d'événements informatifs. Toutefois, la détection d'événements reste un travail d'analyse de corpus statiques qui ne permet pas de prendre en compte le contexte de forte nouveauté inhérent à Twitter tant aucun suivi n'est effectué et aucune limite n'est mise au nombre d'événements détectés.

Référence	Bruit sémantique	Volume de données	Nouveauté des données
Eddi [BSH+10]	✓	X	X
STED [HCZ+13]	✓	X	X
Twical [REC+12]	✓	X	X
Twevent [LSD12]	✓	X	X
ET [PK13]	X	✓	X
EDCoW [WL11]	✓	✓	X
TwitterStand [SST+09]	✓	✓	X
RW-Event [BNG11]	✓	✓	X

TABLE 2.1 – Comparaison des méthodes de détection d'événements dans le contexte de données Twitter.

4 Résumé d'événements

Dans cette section, nous supposons qu'un événement a déjà été identifié et que l'ensemble des tweets s'y rapportant a déjà été rassemblé, par exemple par une des approches précédentes. L'objectif des méthodes présentées ici est de fournir un résumé de l'événement étudié en modélisant le sujet traité, en détectant les personnalités impliquées et en localisant l'événement dans le temps et dans l'espace. Pour ce faire, différents types de résumés ont été proposés dans la littérature : un sous-ensemble de tweets, un tableau de bord contenant les quatre éléments constitutifs d'un événement ou encore une liste (ordonnée) de termes représentatifs (mots, bigrammes, etc…). De plus, puisque les méthodes de résumé d'événements ciblent souvent des événements endogènes (dit du monde-réel), il n'est pas rare que des

ressources supplémentaires soient disponibles (retranscription de discours[4], articles de presse, sites web, etc...). Dans la suite, nous distinguons deux cas : le cas où l'information disponible est réduite aux tweets et le cas où des ressources externes sont utilisées.

4.1 Résumé à partir des tweets

La grande majorité des approches proposées pour résumer un ensemble de tweets relatif à un événement est basée sur le modèle Latent Dirichlet Allocation (LDA) [BNJ03] ou l'une de ses variantes. Nous donnons une présentation détaillée du modèle LDA dans le chapitre 4. Rappelons juste que LDA est un modèle probabiliste génératif à variable latente (les thèmes) dont les paramètres sont estimés à partir des données. Ce modèle représente chaque document par une distribution de probabilités sur les thèmes et chaque thème par une distribution de probabilités sur les termes du vocabulaire.

Plusieurs extensions de l'approche originale ont été proposées, notamment afin de prendre en compte la dimension temporelle. En particulier, le modèle Gaussian Decay Topic model (GDTM) [CA13] ajoute une fonction d'oubli au prior de Dirichlet (la distribution *a priori*) de LDA pour exploiter la corrélation temporelle entre tweets. Tout d'abord, ce modèle analyse les phrases nominales afin d'améliorer l'interprétabilité des résultats et ainsi limiter l'impact du bruit sémantique. Ensuite, les thèmes calculés pour un tweet héritent des paramètres des thèmes des tweets précédents afin de suivre le glissement de concepts ("concept drift") dû au contexte de forte nouveauté. De façon à présenter des résultats concis, le résumé est constitué d'une liste de thèmes, et pour chacun d'eux, de l'ensemble des tweets les plus représentatifs. Néanmoins, même si l'évaluation montre de bons résultats, la complexité accrue du modèle limite son application à des jeux de données de taille modeste (de 10 000 à 250 000 tweets).

En complément des thèmes produits par LDA, il s'agit souvent pour résumer un événement est de répondre spécifiquement à chacune des questions qui ? quoi ? où ? et quand ? Dou et al. [DWS+12] proposent l'outil Leadline qui offre une interface graphique pour explorer un événement sous plusieurs angles : étant donné un ensemble de tweets, Leadline construit un modèle LDA pour extraire le contenu sémantique et répondre à la question "quoi ?". Il construit ensuite un signal pour chaque thème en fonction des dates de publications des tweets associés et applique une méthode de détection de pics [Mon85] à ces signaux pour localiser l'événement dans le temps. Finalement, un algorithme d'extraction d'entités nommées[5] est appliqué pour détecter les personnalités et les lieux et répondre aux questions "qui ?"

[4] www.realclearpolitics.com/transcript_speeches

[5] Alias-i. lingpipe 4.1.0. http ://alias-i.com/lingpipe, 2012

et "où ?". Le processus est complété par une interface graphique interactive représentant ces éléments sous la forme d'un "stream graphe" (aires empilées distribuées autour d'un axe central), d'une carte et d'un nuage de mots.

Une représentation similaire est proposée par Marcus et al. [MBB+11] dans Twitinfo. Contrairement à Leadline, cette approche n'utilise pas LDA pour modéliser le contenu sémantique de l'événement et se concentre sur la détection de pics en l'appliquant au signal modélisant le volume total de tweets. Étant donné un mot clé relatif à un événement, Twitinfo commence par récupérer auprès de Twitter tous les tweets contenant ce mot puis détecte les périodes de forte activité en appliquant un algorithme de détection de pics inspiré par l'algorithme d'anti-congestion du protocole de transport réseau TCP[6]. Chaque pic correspond à un sous-événement borné dans le temps (ce qui répond à la question Quand ?) labellisé avec les termes discriminants de la période (ce qui répond à la question Quoi ?). Afin de compléter l'analyse, une interface graphique est proposée ; elle permet en outre d'afficher les tweets les plus pertinents et leur polarité (positive ou négative), ainsi qu'une carte localisant les tweets (ce qui répond à la question Où ?).

4.2 Utilisation de ressources externes

L'utilisation de ressources externes permet d'enrichir le résumé d'un événement avec des informations provenant d'une source dont la taille n'est pas limitée (retranscriptions de discours, articles de presse ou sites web). Ces informations additionnelles permettent d'améliorer le traitement du bruit sémantique en valorisant le contenu des messages courts. Toutefois, une difficulté est d'aligner les tweets relatifs à un événement et les parties correspondantes (phrases, paragraphes) des ressources externes.

Par exemple, l'outil d'analyse et de visualisation d'événements Vox Civitas [DNKS10] a été développé pour enrichir une vidéo avec un ensemble dynamique de tweets évoluant en fonction du contenu de la vidéo. Étant donnés une vidéo et un ensemble de tweets correspondant au même événement, Vox Civitas commence par calculer la similarité entre un tweet et la partie de la vidéo durant laquelle il a été posté via la similarité cosinus entre vecteur de mots. Cette procédure simple permet de résumer un événement connu par un ensemble de tweets. De plus, dans l'intention de satisfaire aux besoins journalistiques, une interface graphique propose de visualiser l'évolution temporelle de deux indicateurs analytiques : la popularité sur les réseaux sociaux pour couvrir les "moments décisifs" et une analyse des sentiments pour favoriser les moments "surprenants" et "inhabituels".

Afin de tirer parti de l'expressivité des modèles probabilistes dans le cadre de données provenant de plusieurs sources, Hu et al. [HJWK12] proposent ET-LDA,

[6] tools.ietf.org/html/rfc2988

un modèle permettant de construire des thèmes simultanément dans la retranscription d'un événement et dans les tweets correspondants. Cette variante de LDA effectue deux tâches en parallèle : elle segmente la retranscription de l'événement et estime deux types de thèmes au sein des tweets. ET-LDA définit des thèmes généraux décrivant l'ensemble de l'événement et des thèmes spécifiques assignés à un paragraphe particulier de la retranscription. Suite à cela, les paragraphes consécutifs assignés au même thème sont regroupés pour former des "segments" et chaque tweet est assigné soit à un segment spécifique (et à son thème) soit à un thème général. Le résumé de l'événement est alors composé, à la fois, de tweets *épisodiques* qui sont généralement publiés pendant l'événement et correspondent à des thèmes spécifiques et des tweets *réguliers* qui ont un intérêt plus général.

Une approche similaire est proposée par le "cross-collection topic-aspect model" (ccTAM) [GLD12] qui consiste à modéliser simultanément les thèmes d'un ensemble d'articles de presse et d'un ensemble de tweets. Ce modèle combine le "cross-collection-LDA model" (cc-LDA) [PG09] et le "topic-aspect model" (TAM) [PG10] afin d'identifier des paires phrase-tweet. Le modèle cc-LDA est développé pour modéliser les thèmes dans des documents provenant de différentes sources alors que le modèle ATM ajoute un thème "de fond" servant à représenter l'ensemble des documents. En combinant ces deux approches, le modèle ccTAM propose de construire trois types de thèmes (toujours sous la forme de distributions de probabilités) : les thèmes "de fond", les thèmes communs aux deux collections et les thèmes spécifiques à une collection. Finalement, un algorithme d'ordonnancement ("ranking") de graphe biparti est appliqué au graphe tweet-phrase pour générer deux résumés, un pour les articles, un pour les tweets, qui sont complémentaires à la fois au niveau général et au niveau de la phrase.

Une troisième approche, le modèle de Dual Latent Dirichlet Allocation (DLDA) [JLZ+11], propose d'extraire des thèmes à la fois dans les tweets et dans le contenu vers lequel pointent les URLs qui apparaissent dans les tweets. En construisant simultanément deux ensembles de thèmes pour les textes courts et les textes longs, DLDA est capable de transférer automatiquement de l'information depuis les sites webs vers les tweets afin de partitionner des textes courts. Deux ensembles de thèmes sont construits pour les sources externes (sites web) et pour la source cible (les tweets) en utilisant des priors asymétriques dans le but de forcer chaque source à être générée par ses propres thèmes. Comme dans le cas d'une application standard du modèle LDA, les documents de chaque source sont résumés par des distributions de probabilités sur les thèmes de cette source qui sont eux-mêmes représentés par des distributions sur le vocabulaire.

4.3 Discussion

Dans le tableau 2.2, nous synthétisons les méthodes de résumé d'événements en mettant l'accent sur les trois propriétés que les méthodes de modélisation de thématiques devraient satisfaire dans le contexte de Twitter. Il apparaît tout d'abord que ces méthodes sont adaptées au bruit sémantique puisqu'elles s'appuient soit sur des unités syntaxiques porteuses de sens, soit sur des documents directement liés à un événement pour produire un résumé à forte valeur informative. Toutefois, ces approches ne semblent pas adaptées à l'analyse d'une part importante du flux de tweets car elles requièrent une sélection manuelle des données pertinentes, ce qui n'est pas toujours faisable, en particulier dans le cas d'événements nouveaux ou inattendus. De plus, les approches s'inspirant du modèle de thèmes LDA, qui, comme lui, supposent que les documents sont interchangeables au sein de l'ensemble du corpus [BNJ03], ne sont pas destinées à suivre l'évolution des événements dans le temps.

Référence	Bruit sémantique	Volume de données	Nouveauté des données
GDTM [CA13]	✓	X	+/-
Leadline [DWS+12]	✓	X	+/-
Twitinfo [MBB+11]	✓	+/-	X
Vox Civitas [DNKS10]	✓	X	X
ET-LDA [HJWK12]	✓	X	X
ccTAM[GLD12]	✓	X	X
DLDA[JLZ+11]	✓	X	X

TABLE 2.2 – Comparaison des méthodes de résumé d'événements dans le contexte de données Twitter.

5 Détection de nouvelles tendances

La dimension temporelle est un élément clé des méthodes de découverte d'événements qui s'appuient sur cette composante pour détecter et résumer les événements. Toutefois, ces méthodes se concentrent sur des événements courts et indépendants et ne modélisent pas l'évolution temporelle des thématiques. Les résumés construits correspondent à des instantanés de Twitter représentatifs d'un événement ciblé et de ses sous-événements mais ils ne permettent pas de suivre le cycle de vie d'un concept particulier. La détection de nouvelles tendances ("Emerging Trend Detection"), également appelée "New Event Detection" ou "First Story Detection" [ALJ00], cherche à identifier l'apparition de plusieurs messages traitant d'un sujet commun qui n'avait jamais été abordé jusque là. En identifiant les tendances avant qu'elles deviennent populaires, les approches proposées permettent de connaître,

en un temps qui s'approche du temps réel, les centres d'intérêt des utilisateurs du service de microblogging. La détection de nouvelles tendances se fait en séparant, dans le flux total de tweets, les nouveaux sujets des informations connues. En conséquence les approches de ce domaine mettent l'accent sur la capacité à passer à l'échelle.

5.1 Capteur sociaux

Un des travaux les plus célèbres de la détection de nouvelles tendances dans les données Twitter est le système Toretter de Sakaki et al. [SOM10]. Ce système d'alerte anti tremblements de terre adapte la théorie des réseaux de capteurs aux utilisateurs de Twitter considérés comme des *capteurs sociaux* afin de détecter les catastrophes naturelles ayant lieu au Japon. Toretter repose sur l'hypothèse que, lors d'un tremblement de terre, les utilisateurs (les "capteurs") vont tweeter en temps réel et ainsi produire un signal d'urgence qui pourra être analysé. Étant donné un ensemble de tweets relatifs à une catastrophe naturelle (obtenu par exemple grâce à une requête sur les termes "tremblement de terre"), le système commence par appliquer un classifieur SVM à deux classes [Joa98] de manière à distinguer les vrais rapports d'incident, tels que "ca tremble maintenant" du bruit ne correspondant pas à un rapport temps réel : "le tremblement de terre d'hier était effrayant". Dans un second temps, le modèle de filtre de Kalman [FHL+03] est appliqué pour estimer l'emplacement de la catastrophe en fonction de la date de publication et de la géolocalisation des tweets. Ce modèle probabiliste est répandu pour estimer l'emplacement d'un capteur dans un réseau et montre de bons résultats une fois transposé au contexte Twitter. De plus, la grande réactivité des utilisateurs et du service permet au système d'envoyer des alertes plus rapidement que les antennes officielles. Toutefois, ce système présente plusieurs limitations : il repose sur une forte implication du public, à la fois du point de vue de la réactivité et de la mise à disposition des informations de géolocalisation. De plus, il s'appuie sur une description manuelle des événements afin de cerner le vocabulaire employé pour rapporter la catastrophe et pour entraîner le classifieur.

5.2 Regroupement dynamique

Dans le contexte de l'analyse de flux de textes, l'approche classique pour détecter de nouvelles tendances consiste à comparer un nouveau document aux documents obtenus jusque là et détecter une nouvelle tendance s'il n'est similaire à aucun des messages précédents [APL98]. Cependant, dans le contexte de Twitter, cette approche rencontre deux principales difficultés. Premièrement, le volume de données et la vitesse à laquelle les tweets sont publiés interdisent la comparaison à l'en-

semble des documents précédents. Deuxièmement, en raison de la grande variété des sujets abordés, le nombre de nouvelles tendances détectées peut croître au-delà des capacités de calcul. Pour aborder ces difficultés, Petrović et al. [POL10] exploitent les propriétés du "locality sensitive hashing" (LSH) [IM98] pour résoudre une version approchée du problème en limitant le nombre de comparaisons à un nombre fixe de "*buckets*". En effet, une des propriétés de ce schéma de hachage est que la probabilité de collision entre deux points augmente avec leur similarité. En d'autres termes, un nouveau tweet aura plus de chances d'être haché dans le même *bucket* qu'un de ses voisins les plus similaires. Cette configuration permet ensuite de limiter le nombre de comparaisons aux documents du *bucket* auquel il a préalablement été assigné. Afin de borner le nombre de comparaisons, le nombre de documents dans chaque *bucket* est maintenu constant en retirant au fur et à mesure les plus anciens. L'évaluation montre que cette approximation qui fonctionne en temps et en espace constants produit des résultats compétitifs en comparaison des méthodes de l'état de l'art.

5.3 Factorisation de matrices

Une deuxième approche très répandue, appelée "Dictionary Learning", a été développée à l'origine dans le domaine du traitement du signal [Mal99] et exploite la propriété de réduction de dimensions de la factorisation de matrices pour représenter les documents par une combinaison de "quelques" atomes : les thèmes. Dans cette approche, les documents sont regroupés par intervalles, et sont considérés comme une séquence de matrices correspondant aux matrices creuses terme$\times$document. Les documents entrants sont détectés comme nouveaux s'il ne peuvent pas être représentés, avec une faible erreur, par les atomes calculés jusque là. L'algorithme NVL-CLUST [KMBS11] alterne entre une phase de détection et une phase d'apprentissage pour calculer la meilleure représentation à chaque intervalle. Une fois détectés, les nouveaux documents sont utilisés pour entraîner un second dictionnaire dont les atomes correspondent aux tendances émergentes. Le nombre de tendances détectées est fixé *a priori* et les thèmes sont représentés par leur distribution sur les termes. Des groupes de documents sont construits en assignant chaque document à l'atome qui le représente le plus. Une évaluation manuelle montre de bons résultats et des nouvelles tendances pertinentes ont été identifiées. Toutefois, la taille du vocabulaire est fixée ce qui est très contraignant dans le contexte de forte nouveauté. De plus, comme le notent les auteurs, une version temps-réel aurait besoin d'optimisations supplémentaires.

Ces remarques ont été intégrées par Saha and Sindhwani qui proposent, online-NMF, une extension du travail précédent prenant en compte le glissement de concept ("concept drift") en enrichissant dynamiquement le vocabulaire lors de l'arrivée de

nouveaux documents [SS12]. Les données sont stockées pendant une courte fenêtre de temps de manière à limiter la taille du modèle et les coûts de calcul. A chaque intervalle, le système ajoute les documents entrants à la matrice contenant l'historique des données et calcule les thèmes via la factorisation. La fonction de coût optimisée durant la factorisation satisfait deux considérations : un pourcentage fixé des thèmes correspond à une évolution des thèmes de l'intervalle précédent et le reste correspond aux nouvelles tendances. Une fois détectées, les nouvelles tendances sont ajoutées aux données historiques afin d'être prises en compte pour la détection de nouveauté dans le nouvel intervalle. Cette procédure ainsi que les optimisations proposées permettent de prendre en compte le volume de données et le contexte de forte nouveauté, même si la présence de nombreux paramètres peut rester contraignante.

5.4 Latent Dirichlet Allocation

Une alternative à la factorisation de matrice est proposée par Lau et al. [LCB12] qui appliquent une variante dynamique de LDA pour analyser l'évolution de documents et extraire des nouvelles tendances. Le corpus est divisé en périodes successives et un modèle LDA est appris pour chacune d'elles. Dans cette variante, un nouveau modèle hérite des paramètres du modèle précédent (distributions des documents et distributions des thèmes) de façon à produire des thèmes comparables. De plus, cela permet de contrôler les coûts de calcul en ne conservant que les documents du dernier intervalle. Pour cela, le vocabulaire est recalculé à chaque mise à jour avec les documents de la fenêtre de temps en cours, ce qui permet également de prendre en compte le glissement de concepts. Le fait que le vocabulaire soit dynamique permet de faire apparaître de nouveaux concepts dans les données Twitter. Une correspondance un-pour-un est maintenue entre les thèmes de deux modèles consécutifs et un thème est dit nouveau s'il n'est pas similaire, en termes de divergence de Jensen-Shannon [Lin91], à un thème précédent. En outre, un document est dit nouveau si son thème majoritaire l'est.

5.5 Modèles de langage

Le système de détection de sujet saillant en temps réel, TopicSketch [XZJ+13] implémente un modèle basé sur le concept de résumé qui est adapté au grand volume de données. TopicSketch met à jour un résumé du flux de données et construit un modèle de détection de nouveauté pouvant potentiellement traiter l'ensemble des tweets publiés. Le modèle proposé s'appuie sur l'accélération de trois quantités : le nombre total de tweets, le nombre d'occurrences de chaque terme individuellement et le nombre d'occurrences des paires de termes. Afin d'optimiser les temps de cal-

cul, une fonction de hachage est employée pour associer les termes à un nombre *significativement* plus petit de "buckets". Ainsi, durant la phase d'apprentissage, tous les termes d'un "bucket" sont identifiés en un "mot" unique, ce qui réduit la taille du vocabulaire. Cette méthode permet de détecter les thèmes présentant un motif saillant marqué, en revanche elle ne détecte pas les nouveautés qui peuvent apparaître lentement. De plus, cette approche basée uniquement sur le signal ne prend pas en compte le bruit sémantique.

5.6 Discussion

Dans le tableau 2.3, nous résumons les approches de détection de nouvelles tendances selon les trois critères déjà mentionnés. Il apparaît que les efforts de développement fournis répondent partiellement aux contraintes de la modélisation de thématiques dans Twitter, même s'il reste quelques limitations. En effet, Toretter nécessite une liste manuelle de termes afin de récupérer les tweets relatifs aux tremblements de terre, ce qui limite l'utilisation du système dans d'autres contextes. Par ailleurs, les modèles utilisant une fonction de hachage comme ceux de Petrović et al. et Xie et al. ne réduisent les temps de calcul qu'au détriment d'une abstraction des termes ou des documents limitant l'interprétabilité des résultats. De manière analogue, dans l'approche de factorisation de matrice proposée par Saha and Sindhwani, les thèmes "anciens" sont retirés des données stockées de manière à de réduire les coûts de calculs. Cela permet la détection de thèmes périodiques, mais limite l'efficacité du système lors d'une analyse à long terme.

Référence	Bruit sémantique	Volume de données	Nouveauté des données
Toretter [SOM10]	X	✓	✓
LSH-FSD[POL10]	+/-	✓	✓
NVL-CLUST[KMBS11]	X	X	X
online-NMF [SS12]	X	✓	✓
online LDA variant[LCB12]	X	X	✓
TopicSketch [XZJ+13]	X	✓	X

TABLE 2.3 – Comparaison des méthodes de détection de nouvelles tendances dans le contexte de données Twitter.

6 Suivi de tendances

Afin que les nouvelles tendances détectées ne restent pas, comme les événements, des instantanés décrivant des discussions entretenues en ligne à un instant donné, il est nécessaire de proposer à l'analyste un suivi permettant de remettre ces informations dans le contexte. Ainsi, les approches de suivi de tendances ("Trend

Analysis") cherchent à modéliser l'évolution temporelle des thèmes détectés. Ces approches trouvent leur origine dans la tâche de "Topic Tracking" du "Topic Detection & Tracking" et sont également appelées "Trend Analysis" [LCB12] ou "News Tracking" [CYL+99]. Les modèles de ce domaine exploitent explicitement les dates de publication des documents pour prendre en compte le contexte de forte nouveauté inhérent à Twitter et proposer un suivi de thématiques qui maintient la carte mentale de l'analyste.

6.1 Extraction de règles d'association

Dans le but de modéliser la dynamique des tweets, la méthode "Transaction-based Rule Change Mining" (TRCM) [AOGS13] propose d'appliquer un algorithme d'extraction de règles d'association temporelles pour détecter la dynamique des hashtags. TRCM détecte des changements dans les règles extraites en fonction de l'évolution des hashtags dans les tweets. Tout d'abord, TRCM divise le corpus en périodes de temps consécutives et extrait les règles d'association entre hashtags pour chaque période. Ensuite, un algorithme classique est appliqué pour mettre en correspondance les règles détectées dans chaque période [SK+01] et ainsi mettre en avant des motifs de la forme "règle émergente", "nouvelle règle", "règle inattendue" et "règle morte". Ce procédé permet de suivre l'évolution des règles au cours du temps et prend en compte les contraintes d'un contexte de forte nouveauté. La similarité entre les règles [LSLL09] dépend simplement du nombre de hashtags apparaissant dans les prémisses et dans les conséquences des règles. Toutefois, l'explosion combinatoire liée au nombre de règles générées et au nombre de comparaisons nécessaires peut être un frein à l'utilisation de cette approche dans un contexte applicatif.

6.2 "Agenda Setting Theory"

En mettant l'accent sur "la compréhension intuitive des mesures estimées", permise par la visualisation sous forme de timeline, le modèle ThemeRiver [XWW+13] propose une méthodologie inspirée par la recherche en "agenda-setting" [MS72]. La théorie de l'agenda-setting repose sur la forte corrélation observée entre les sujets traités dans les médias de masse (appelés "média-agenda") et les sujets de conversation considérés comme importants par la majorité du public (appelés "public-agenda"). Cette corrélation semble suggérer que les médias orientent les conversations du public. La théorie de l'agenda-setting est considérée comme un modèle à "somme nulle" dans lequel les sujets de conversation doivent lutter pour obtenir une couverture médiatique et l'attention du public dans un contexte où ces deux ressources sont limitées. En effet, le temps et la place des médias sont limités, tout

comme la capacité du public à traiter de l'information. Ainsi, l'addition d'un nouveau sujet se fait au détriment des autres. Zhu et al. [Zhu92] ont développé un système d'équations différentielles qui modélise l'évolution de l'intérêt du public pour un sujet particulier entre deux périodes de temps consécutives en fonction de la couverture médiatique. Ce système intègre deux mécanismes : la couverture médiatique d'un sujet donné augmente l'intérêt du public pour ce sujet et à l'inverse, la couverture médiatique des autres sujets diminue l'intérêt pour le premier sujet. L'analyse de k sujets correspond à la mise en place d'un système à k équations dont les paramètres sont estimés par une méthode de régression standard. ThemeRiver étend ce système pour prendre en compte la diversité dans "l'agenda" de différents leaders d'opinion en représentant la couverture médiatique par la somme des couvertures des différents leaders. La plus grande particularité de ThemeRiver dans le contexte de la fouille textuelle de données Twitter est l'analyse de la corrélation entre l'évolution de l'intérêt des leaders d'opinion et l'évolution de la compétitivité des thèmes. Toutefois, l'ensemble des thèmes analysés est sélectionné manuellement, ce qui favorise l'interprétation des résultats mais limite la prise en compte de la nouveauté.

6.3 Regroupement dynamique

Une autre approche pour visualiser les flux de textes dans un contexte de forte nouveauté est proposée par TwitterScope [GHN12]. Cette approche combine un algorithme de regroupement et de positionnement dynamique et un outil de visualisation pour construire des cartes évolutives de tweets. TwitterScope affiche les tweets similaires sous forme de "pays" sur une carte qui est mise à jour régulièrement à chaque fois que de nouveaux tweets arrivent afin de proposer "une vue continue et succincte de l'évolution des centres d'intérêt". Tout d'abord, un algorithme de partitionnement de graphe utilisant une mesure de modularité est appliqué [New06] pour identifier des clusters de tweets et la carte est construite par un algorithme de placement basé sur le "multidimensional scaling" [Kru64] pour placer les tweets similaires proches les uns des autres. Lorsque de nouveaux tweets arrivent, TwitterScope recalcule une nouvelle partition et la vue est mise à jour à la fois en préservant la proximité entre pays et en évitant l'introduction de recouvrements entre pays. Même si TwitterScope propose une vue synthétique et dynamique d'un ensemble de tweets, l'algorithme de regroupement et de positionnement dynamique ne paraît pas entièrement adapté à un flux de données fortement nouvelles. En effet, les clusters calculés entre deux mises à jour pourraient être totalement différents et les cartes créées pourraient ne plus avoir aucun point commun. Ainsi, l'analyste ne serait plus à même de suivre l'évolution temporelle des thématiques.

6.4 Latent Dirichlet Allocation

De nombreux modèles probabilistes ont été proposés pour traiter la nature fortement dynamique des données Twitter. LDA est une approche populaire pour construire des thèmes représentatifs d'un corpus statique et différentes variantes ont été proposées pour modéliser les transitions entre thèmes dans un corpus dynamique. Le modèle "Multi-Faceted Topic Model" (MfTM) [VJLN13] décrit une relation sémantique latente entre termes et entités et capture les caractéristiques temporelles de chaque thème avec une variable dédiée. MfTM étend LDA pour prendre en compte le bruit sémantique en représentant un thème par cinq variables cachées (les "facettes") au lieu d'une. Ces cinq facettes correspondent : aux termes (comme dans LDA), à trois types d'entités nommées, les personnalités, les lieux et les organisations, et aux dates de publication. Chaque date est alors associée à une distribution de probabilités sur les thèmes, et en observant symétriquement le score de chaque thème au cours du temps, MfTM permet d'estimer l'évolution des thèmes. Malgré cette variable spécifique dédiée au temps, MfTM hérite du même inconvénient que le modèle LDA sur lequel il repose : les documents sont supposés interchangeables et cette approche n'est donc pas pleinement adaptée à la forte nouveauté de Twitter.

Une première approche pour résoudre cette limitation est de diviser le corpus en intervalles de temps et de construire des modèles *indépendants* pour chaque intervalle. Les documents ne sont alors plus supposés interchangeables qu'au sein de chaque intervalle. Le suivi de thématiques est assuré en calculant la similarité entre thèmes en fonction de leur distribution de probabilités sur les termes. De manière à proposer une description synthétique des données trouvant un compromis entre la précision et le volume de données présenté à l'analyste, TopicFlow [MSH+13] entraîne un modèle LDA pour chaque intervalle avec un nombre fixé de thèmes. En effet, comme le font remarquer les auteurs, un nombre adapté de thèmes permet d'extraire des connaissances sans encombrer la visualisation. Suite à cela, la similarité entre thèmes est calculée par similarité cosinus entre les vecteurs de mots en se restreignant aux 20 termes les plus représentatifs afin d'accélérer le procédé. Les paires de thèmes dont la similarité est supérieure au seuil fixé sont considérées similaires et sont reliées sur la visualisation.

Une autre façon d'éviter les écueils de l'interchangeabilité des documents est de construire des modèles *dépendants* qui estiment les thèmes en fonction de ceux des modèles précédents. Le modèle "Dynamic Multi-Relational Chinese Restaurant Process" (D-MRelCRP) [LB12] propose par exemple d'introduire une fonction d'oubli afin de modéliser la popularité des thèmes. Par ailleurs, cette approche modélise la relation entre utilisateurs en intégrant la notion de "voisinage" dans les paramètres. Plusieurs définitions de voisinages, telles que les relations sur le réseau ou la localisation géographique, sont combinées de manière à modéliser différentes

influences. Ce processus peut permettre de cibler plusieurs zones d'intérêt et ainsi résoudre en partie le problème du bruit. En complément, l'évolution temporelle de la personnalité de l'utilisateur et des distributions des thèmes est modélisée en intégrant deux composantes dynamiques aux priors de Dirichlet, ce qui permet de suivre l'évolution des thématiques. Toutefois, comme souligné par les auteurs, l'évolution du réseau n'est pas modélisé, ce qui peut limiter l'intérêt des voisinages puisque le nombre d'utilisateurs et les liens qui les relient évoluent au cours du temps.

6.5 Discussion

Dans le tableau 2.4, il apparaît que les coûts de calcul nécessaires à l'estimation des paramètres des modèles proposés ici est une limitation importante pour la modélisation du flux total et peu d'approches sont destinées à traiter le flux complet. De plus, la prise en compte du bruit sémantique semble se faire au détriment de la gestion du contexte de forte nouveauté. Par exemple, le modèle D-MRelCRP de Lakkaraju et al. s'appuie sur les relations entre utilisateurs afin de consolider les thématiques mais l'évolution du réseau n'étant pas modélisée, cette approche ne semble pas adaptée à une analyse sur le long terme. Néanmoins, l'outil TopicFlow de Malik et al. semble intéressant car il propose à l'analyste un nombre contrôlé de thèmes calculés à partir d'un vocabulaire dynamique et évoluant de manière riche : les thèmes peuvent se scinder et fusionner au cours du temps en fonction de leur similarité. Toutefois, plusieurs points peuvent être améliorés puisque cette approche assigne chaque document à son thème le plus représentatif sans prendre en compte la propriété de clustering flou de LDA : en associant à chaque document une distribution sur les thèmes, LDA permet d'assigner un document à plusieurs thèmes avec plus ou moins de force. En sus, le calcul de la similarité entre thèmes établis à partir de vocabulaires différents n'est pas détaillé et pose pourtant problème. Comment calculer la similarité entre vecteurs dans des espaces vectoriels différents ? Puisque cette approche semble adaptée à l'étude de l'évolution temporelle de thématiques dans un contexte de forte nouveauté, nous décrivons dans le chapitre 4 une proposition similaire exploitant la propriété de clustering flou de LDA. Nous décrivons notamment le procédé employé pour comparer des thèmes calculés sur des vocabulaires différents et nous proposons une comparaison des distributions empiriques de sept mesures de divergence entre distributions de probabilités afin de déterminer la mesure la plus adaptée au calcul de similarité entre thèmes.

7 Conclusion

Dans ce chapitre, nous avons proposé une organisation des travaux d'extraction de thématiques dans Twitter basée sur deux concepts essentiels de la fouille de don-

Référence	Bruit sémantique	Volume de données	Nouveauté des données
TRCM [AOGS13]	X	X	+/-
ThemeRiver [XWW+13]	X	X	✓
TwitterScope [GHN12]	X	✓	X
MfTM [VJLN13]	✓	X	X
TopicFlow [MSH+13]	X	X	✓
D-MRelCRP [LB12]	✓	X	X

TABLE 2.4 – Comparaison des méthodes d'analyse de tendances dans le contexte de données Twitter.

nées textuelles dans ce nouveau média : les événements et les tendances. Les événements décrivent des instantanés du réseau, étudiés sans relation les uns avec les autres. Leur analyse se décline en deux sous-tâches, la détection et le résumé d'événements. Ces tâches mettent l'accent sur la richesse de l'information extraite en fournissant une description complète des événements : sujet, lieu, date et personnalités impliquées. Les tendances découlent de l'analyse de la dimension temporelle du flux de texte, et s'intéressent à l'évolution des thématiques. Cette évolution peut correspondre, non seulement, à l'apparition ou à la disparition de thématiques liées aux multiples intérêts du public, mais également, à des glissements sémantiques dans le vocabulaire employé pour les décrire. Ces travaux peuvent être divisés en deux sous-tâches, la détection de nouvelles tendances et le suivi de tendances.

Ces quatre sous-tâches sont mises en regard des difficultés majeures de l'extraction de thématiques dans les données Twitter présentées en introduction : le bruit sémantique lié à la multitude de sujets traités, le volume des données créées par ce média très populaire et le contexte de forte nouveauté caractéristique de cette plateforme de communication en temps réel. Puisque les travaux existants ne traitent pas explicitement de la difficulté liée au bruit syntaxique, cet aspect est étudié séparément dans le chapitre suivant, où il est donné une description détaillée des prétraitements appliqués aux données Twitter.

Cette synthèse bibliographique met en avant le caractère contradictoire de l'analyse des données Twitter, entre traitement de données volumineuses et évoluant rapidement, et nécessité de consolider les thématiques construites afin d'extraire une information pertinente pour l'analyste. L'analyse en temps-réel rend délicate l'application de filtres sémantiques, pourtant nécessaires à la détection de concepts porteurs de sens, et les algorithmes existants peuvent conduire à l'extraction d'une abondance de thématiques sans précisions des relations sous-jacentes et se confrontent à la difficulté du nettoyage du "bruit sémantique".

3

Fouille exploratoire

In mathematics the art of proposing a question must be held of higher value than solving it.

—Georg Cantor

Sommaire

1 Introduction

En parallèle de la réflexion théorique sur la modélisation du problème, nous avons exploré des données réelles dans l'intention de mieux saisir leurs spécificités et de recueillir des retours d'interprétation. Puisque nous ne possédions pas d'informations sur les données *a priori* dans cette tâche d'extraction de connaissances, nous ne disposions pas de méthodes de référence pour conduire l'analyse. Ainsi, nous avons mis en place, en collaboration avec un expert de l'entreprise, plusieurs approches classiques d'apprentissage non-supervisé afin de déterminer au fur et à mesure quels types d'informations valorisables étaient contenus dans ces données spécifiques.

Dans ce chapitre, nous présentons les différentes étapes de cette étude prospective nous ayant amenés à proposer une variante dynamique et non paramétrique du modèle LDA. Tout d'abord, nous présentons, dans la section 3.2, le processus de collecte mis en place en amont et effectué sans interruption en vue de rassembler l'ensemble des tweets concernant l'entreprise dans les conditions fixées par Twitter. Puis nous présentons, dans la section 3.3, les pré-traitements appliqués aux tweets non-structurés extraits des services de Twitter dans le but de les transformer en données exploitables. Suite à cela, nous présentons l'application de deux approches classiques de l'apprentissage non-supervisé : le clustering de texte, dans la section 3.4, et l'extraction de motifs fréquents, dans la section 3.5, permettant de mettre en avant le contexte de forte nouveauté des données Twitter. Pour finir ce chapitre, nous présentons, dans la section 3.6, une approche d'analyse de la co-évolution temporelle des termes développée dans le but de proposer à l'analyste un premier moyen d'étudier le comportement des thématiques sur le long terme.

2 Collecte des données

2.1 Limitations à la création de jeux de données de référence

Twitter a mis en place un accès facilité aux messages et aux données relatives publiés sur son service via différentes APIs[1], mais les utilisateurs, y compris les chercheurs, ne sont pas autorisés à partager des jeux de données via des méthodes automatiques[2]. La seule façon de partager automatiquement des tweets est de distribuer la liste de leurs identifiants uniques afin de permettre à d'autres personnes de rassembler les messages en utilisant les services des APIs. Cette limitation vise à protéger la vie privée des utilisateurs en garantissant que les messages supprimés par un utilisateur ne seront plus distribués. Néanmoins, cela impacte grandement

[1] dev.twitter.com
[2] dev.twitter.com/overview/terms

la communauté des chercheurs car de nombreux jeux de données cités dans la littérature ne sont plus disponibles (Edinburgh Corpus[3], SNAP dataset[4], Twitter FSD Corpus[5]).

Outre les limites techniques fixées par Twitter, telles que la limite du nombre de requêtes effectuées aux APIs ou le court historique de données mis à disposition, le système de droits mis en place pour gérer l'accès aux APIs limite également la mise en place de jeux de données volumineux. En effet, l'accès standard ne permet de récupérer qu'un pourcentage limité de l'ensemble des messages publiés et les accès privilégiés ne sont offerts que dans des conditions strictes[6]. Un petit nombre de chercheurs ont, ou ont eu, accès à une part plus significative du flux afin de constituer de grands jeux de données et mener des études approfondies ([HD10, SOM10]). Cependant, même si les identifiants des tweets correspondants étaient distribués, il serait très difficile de récupérer toutes les données avec les contraintes des APIs.

En sus des limitations techniques, la difficulté à définir des mesures de qualité et des valeurs de références dans la définition des thématiques ont également limité la création de jeux de données de référence. Tout d'abord, l'énorme volume de messages publiés sur Twitter empêche toute analyse manuelle, ce qui contrarie l'établissement de catégories standard pour traiter les thématiques. Une analyse manuelle est parfois proposée pour de petits jeux de données [CA13, DWS+12], mais pour les plus gros, une analyse automatique est nécessaire pour déterminer les valeurs de référence, telles que les labels ou les classes [APM+13, HD10]. Dans ce cas, la comparaison des méthodes revient essentiellement à déterminer quelle approche est la plus similaire à celle ayant servi à établir les valeurs de référence.

Par ailleurs, la grande variété de tâches et de centres d'intérêts suscitée par Twitter limite l'établissement d'évaluations standard. En effet, certains travaux s'intéressent à l'analyse de l'ensemble de l'activité du réseau [GHN12, REC+12] alors que d'autres se concentrent spécifiquement sur l'impact sur les réseaux sociaux d'événements ayant lieu dans le "monde réel" [SOM10, HCZ+13]. Certaines études s'intéressent à la valeur informative des thématiques détectées dans un souci de présentations journalistiques [DNKS10] alors que d'autres se concentrent sur des domaines plus spécifiques tels que les marques ou les célébrités [CALC13]. Les thématiques construites dans différents contextes ne sont pas forcément transposables et les méthodes les ayant détectées ne sont pas nécessairement comparables.

Plus généralement, les thématiques peuvent être décrites à plusieurs niveaux d'abstraction, depuis les grandes tendances générales, jusqu'à l'information la plus pointue, ce qui limite la création de valeurs de référence universelles [LCB12]. Il

[3] homepages.inf.ed.ac.uk/miles/papers/socmed10.pdf
[4] snap.stanford.edu/data/twitter7.html
[5] demeter.inf.ed.ac.uk/cross/docs/fsd_corpus.tar.gz
[6] twittercommunity.com/t/api-firehose/7684

est bien connu que la définition du nombre de thématiques adapté à un corpus particulier reste un problème ouvert et largement sujet à interprétation [MC85].

2.2 Propositions récentes de corpus

Malgré ces limitations, à la fois techniques et méthodologiques, et afin de satisfaire les besoins croissants en moyens d'évaluation des méthodes, deux jeux de données ont été rendus disponibles.

Le premier est un jeu de données non étiquetées qui est composé d'un million de tweets publiés pendant une période d'un mois et contenant une information de géolocalisation. Ce jeu de données a été créé pour la tâche "MC1 - Characterization of an Epidemic Spread" du challende VAST2011 [7] qui s'intéresse à la caractérisation de la diffusion d'une épidémie via l'analyse des tweets. Ce jeu de données a également été utilisé dans le contexte de la détection d'événements [PK13]. Toutefois, en raison de l'absence de valeur de vérité, l'évaluation de la méthode proposée n'a pu se faire que selon le critère de la *précision* puisque le *rappel* aurait nécessité l'énumération de *tous* les événements du corpus.

Le deuxième corpus[8] est un jeu de données volumineux élaboré dans le but d'évaluer les méthodes de détection d'événements [MMJ13]. Il est composé de 120 millions de tweets publiés pendant une période d'un mois, dont 150 000 se sont vu attribuer un jugement d'appartenance à l'un des 500 événements qu'il couvre. Les événements sont construits en combinant deux méthodes d'extraction automatique ([POL10, AS12]) avec le Current Events Portal (CEP)[9] proposé par Wikipédia. Les deux méthodes ont été sélectionnées pour leur capacité à produire des groupes de documents du même niveau de granularité (plutôt que des groupes de termes) et à traiter des gros volumes de données. En combinant ces approches avec la ligne éditoriale de Wikipédia, les événements extraits du CEP sont de bonne qualité et couvrent de nombreux sujets et catégories. Les jugements d'appartenance sont générés manuellement par une évaluation participative mise en place grâce au Mechanical Turk[10] d'Amazon. La tâche d'évaluation proposée aux participants consiste à lire un tweet et à estimer s'il est en relation avec un événement d'actualité. Dans l'affirmative, il est ensuite demandé de décrire brièvement l'événement et de sélectionner la catégorie qui lui correspond. En raison de l'effort mis en place pour constituer manuellement des valeurs de références pour un jeu de données volumineux, ce corpus pourrait devenir une référence.

[7] hcil2.cs.umd.edu/newvarepository/VAST Challenge 2011/challenges/
[8] mir.dcs.gla.ac.uk/resources/
[9] en.wikipedia.org/wiki/Portal :Current_events
[10] www.mturk.com/mturk/welcome

2.3 Création de jeux de données *ad hoc*

Cependant, les difficultés à partager des corpus communs au sein de la communauté de chercheurs conduisent la plupart des travaux du domaine de la modélisation de thématiques appliqués aux données Twitter à créer un jeu de données d'évaluation. Pour cela, il existe deux stratégies : suivre en continu le pourcentage représentatif de Twitter via l'API Streaming[11] ou utiliser l'API REST[12] pour mettre en place un système de requêtes et collecter les messages pertinents. Dans le premier cas, l'API Streaming propose un échantillon représentatif du flux total avec un faible temps de latence [REC+12, GHN12, CDS10] et il a été montré qu'en utilisant plusieurs accès en parallèle, il était possible de récupérer une partie importante du flux [DWS+12]. Dans le second cas, grâce à son système de requêtes, l'API REST permet à une application tierce de collecter tous les tweets pertinents publiés dans un récent passé[13]. Par ailleurs, les critères de recherche proposés permettent n'importe quelle combinaison de termes [CA13, DNKS10, SS12], hashtags [XWW+13], auteurs [GSB+13, VJLN13, XZJ+13] et lieux [CALC13]. Il est également possible d'utiliser des motifs plus élaborés comme, par exemple, une forme "entité nommée + verbe d'action" afin de cibler des tweets plus spécifiques en exploitant complètement les possibilités de l'API REST [HCZ+13]. Finalement, Twitter propose également un accès direct à ses "Trending Topics", définis selon ses propres critères, qui permet de collecter les tweets traitant des sujets les plus populaires [GGZ+13].

Une autre conséquence de la difficulté à définir des mesures de qualité pour évaluer les thématiques détectées a été de conduire les chercheurs à valider leurs résultats avec une approche qualitative. En effet, même s'il n'est pas toujours possible de définir de manière exhaustive ce que la méthode aurait dû trouver, il reste malgré tout intéressant d'illustrer le type de connaissances qu'elle peut potentiellement extraire. Dans ces circonstances, la plupart des corpus élaborés répondent à des besoins de visibilité et traitent de sujets universels, qui toucheront une large audience. Les sujets d'actualité, comme les débats politiques [XWW+13] ou les catastrophes naturelles [SOM10], représentent des sources potentielles pour extraire des connaissances interprétables et originales à partir de tweets. Les confrontations sportives [MBB+11] ou les émissions télévisées [PK13] réunissent régulièrement une grande part de l'audience de Twitter et peuvent permettre d'établir des résultats représentatifs. Et l'analyse de tweets relatifs aux entreprises offre la possibilité d'extraire des informations à forte valeur ajoutée sur l'opinion des utilisateurs vis-à-vis d'une marque ou d'un produit [CALC13, PBG+14].

[11] dev.twitter.com/streaming/overview
[12] dev.twitter.com/rest/public
[13] dev.twitter.com/rest/public/search

2.4 L'outil de collecte ListenTwitter

Nos travaux s'appuient sur l'outil de collecte ListenTwitter ayant l'objet d'un dépôt de brevet par la R&D d'EDF [BS15], qui interroge quotidiennement l'API REST de Twitter afin de récupérer les messages postés en langue française et contenant le mot clé "edf". Cette requête est assez générale et, même si elle ne peut pas être exhaustive, permet de récupérer tous les messages qui mentionnent l'entreprise. Néanmoins, la limite de 140 caractères imposée par le service incite les utilisateurs à également employer le mot "edf" pour désigner "l'Equipe De France". ListenTwitter applique alors un filtre (classifieur bayésien naïf + marcheur aléatoire sur le graphe des utilisateurs) pour séparer les messages en deux classes "Entreprise" et "Sport" [SSP12]. Dans la suite, nos travaux supposent l'existence d'un ensemble de tweets cibles, et la recherche de la solution optimale pour collecter le plus précisément possible les tweets concernant une entité donnée n'entre pas dans le cadre de cette thèse.

Les données sont collectées depuis le 17/06/2012 et toutes les expérimentations présentées dans cette thèse ont été conduites sur un extrait du corpus d'une durée d'environ 300 jours allant du 17/06/2012 au 02/05/2013. Cela représente 73 000 tweets labellisés "Entreprise" pour une moyenne de 240 tweets par jour. Ces volumes sont plutôt faibles comparés à ce que l'on peut trouver dans la littérature (120 Millions pour le "large scale corpus" de McMinn et al. [MMJ13]) mais ils s'étendent toutefois sur une période particulièrement longue qui offre une opportunité intéressante d'analyser l'évolution de thématiques sur le long terme.

3 Pré-traitements

En tant qu'outil de microblogage, une des principales caractéristiques de Twitter est la limite de 140 caractères imposée aux tweets. Cette limitation favorise, en autre, l'usage d'abréviations et de segmentation informelle des textes. Les tweets bruts nécessitent donc une étape de pré-traitement avant de pouvoir être analysés. La principale opération de cette étape est la segmentation du texte en unités syntaxiques : mots, bigrammes, etc. Cette opération peut également être complétée par une opération d'enrichissement du texte initial en regroupant plusieurs tweets pour former un "document" unique plus volumineux ou en ayant recours à des ressources externes dont la taille n'est pas limitée.

3.1 Cadre général

3.1.1 Segmentation de texte

La segmentation de textes ("Tokenisation") consiste à diviser un texte brut en unités syntaxiques : les "tokens", qui servent ensuite à modéliser la structure sémantique des documents : les thématiques. Même si la plupart des travaux analysent les tweets en les segmentant au niveau des termes [SST+09, DNKS10, GGZ+13, GHN12], des alternatives ont proposé d'analyser des unités plus riches telles que les bigrammes et les multi-grammes [PK13]. Li et al. [LSD12] ont par exemple proposé d'extraire des "concepts significatifs" grâce au service N-Gram[14] de Microsoft et de valider leur caractère informatif en vérifiant leur présence dans des articles de Wikipédia[15]. De nombreux efforts ont été également faits pour extraire des thématiques interprétables en basant l'analyse sur les entités-nommées [REC+12] ainsi que sur les phrases nominales [CA13] extraites des tweets. Les hashtags, qui représentent une information sémantique apposée aux tweets par leur auteur, ont aussi été utilisés en tant qu'unités syntaxiques, faisant fi des autres données textuelles [AOGS13].

3.1.2 Agrégation de tweets et extension par des ressources externes

Les approches classiques d'extraction de thématiques basées sur des informations statistiques, telles que la co-occurrence des termes, semblent mises en défaut par la taille très restreinte des tweets qui donne aux données Twitter un caractère parcimonieux. Il a été montré qu'agréger les tweets par période [GSB+13] ou en fonction de leur auteur [WAB12] permet de construire des modèles de meilleure qualité [HD10, MSBX13].

Une autre solution consiste à enrichir les tweets avec des informations provenant de sources externes dont la taille n'est pas limitée [JLZ+11]. Le système TUCAN [GSB+13] propose par exemple de réduire la taille du vocabulaire en se basant sur l'ontologie du service Wordnet[16] pour généraliser des concepts. Les termes "pistolet" et "fusil" peuvent ainsi être identifiés au concept "arme" ce qui augmente les co-occurrences sans trop de perte d'information. Une approche complémentaire consiste à accéder aux contenus référencés par les URLs contenues dans les tweets et à en extraire les informations sémantiques pour les associer aux tweets [VJLN13].

Au-delà des informations accessibles de manière automatisée fournies par le web, l'intégration d'experts dans le processus d'extraction de thématiques permet également de cibler des contenus spécifiques, pertinents pour un corpus de tweets donnés, comme par exemple des articles de presse [HCZ+13] ou la retranscription écrite d'un discours [HJWK12]. Ces données additionnelles permettent d'extraire

[14] weblm.research.microsoft.com/
[15] www.wikipedia.org
[16] wordnet.princeton.edu/

des thématiques de manière unifiée entre le corpus de tweets et les sources externes [HJWK12, GLD12].

3.2 Chaîne de pré-traitements

Dans la suite, nous considérons un ensemble de tweets $\mathcal{D}$ publiés durant un intervalle de temps $\mathcal{T}$. De manière à modéliser et analyser l'évolution temporelle de son contenu sémantique, ce jeu de données non-structurées doit être pré-traité (voir deux exemples de tweets dans la figure 3.1 et figure 3.2). La chaîne de traitement est présentée sur la figure 3.3 : chaque tweet est découpé en segments séparés par un caractère blanc. Les caractères spéciaux (par exemple : ", /, #), les accents, les mentions (@nom_utilisateur) et les symboles "RT" (pour "retweet") sont retirés. Une stop-liste est utilisée pour retirer les termes non porteurs de sens et un outil de lemmatisation (TreeTagger [Sch94]) est appliqué afin de ramener les termes à une forme canonique. A l'issue de cette étape de pré-traitements, un tweet $\mathbf{d}$ est décrit par un ensemble de termes et une date de publication. L'ensemble des termes associé à $\mathcal{D}$ est appelé vocabulaire et est noté $\mathcal{V}$.

FIGURE 3.1 – Effets de la chaîne de pré-traitemenst sur un tweet (exemple 1).

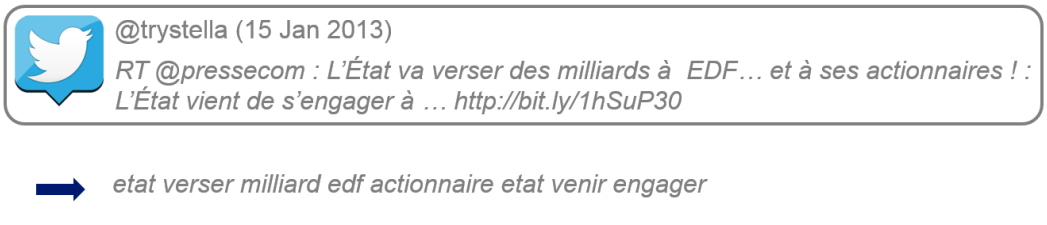


FIGURE 3.2 – Effets de la chaîne de pré-traitements sur un tweet (exemple 2).

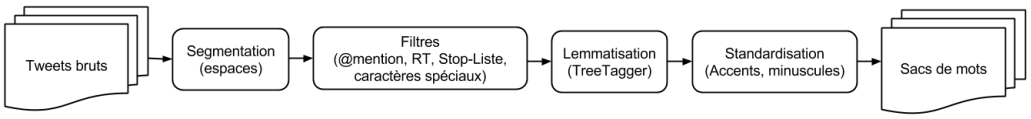


FIGURE 3.3 – Chaîne de pré-traitements des tweets bruts et construction du vocabulaire.

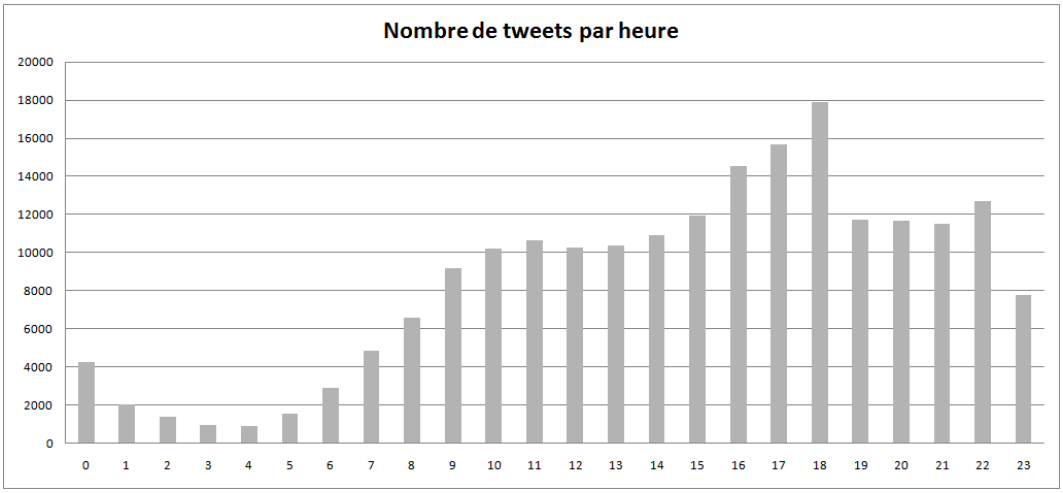


FIGURE 3.4 – Histogramme présentant le nombre cumulé de tweets publiés chaque heure de la journée. Une forte période d'inactivité apparaît la nuit entre 3h et 4h59.

Parallèlement, l'intervalle de temps $\mathcal{T}$ est découpé en périodes de 24h s'étalant de 4h à 3h59, notées $\mathcal{T}_1, \ldots, \mathcal{T}_n$. Ce découpage correspond à une double observation. Premièrement, les thématiques discutées sur Twitter évoluent rapidement et durent rarement plus d'une journée. Ainsi, le découpage en périodes de 24h permet de suivre l'échelle de temps de Twitter en modélisant des thématiques pertinentes pour la journée étudiée. Deuxièmement, comme illustré sur la figure 3.4, la période d'inactivité de Twitter, où les conversations se terminent, se situe au milieu de la nuit entre 3h et 4h59. En effet, le nombre cumulé de tweets publiés par heure montre la faible activité du réseau (francophone) au milieu de la nuit. Cette observation est confirmée par l'étude des 24 séries statistiques décrivant pour chaque heure et pour chaque journée le nombre de messages publiés. La représentation de ces distributions sous forme de diagrammes en boîte (voir figure3.5) montre que malgré une grande disparité dans les valeurs prises par ces distributions, avec notamment des journées de forte activité faisant augmenter la moyenne (la valeur moyenne est systématiquement au-dessus de la valeur médiane), le nombre de tweets publiés entre 3h et 4h59 est systématiquement faible, avec un nombre de tweets horaire inférieur à 5.

4 Clustering de tweets

Nos premières explorations de données s'appuient sur l'algorithme de clustering hiérarchique présenté dans [SSP12] pour regrouper les tweets similaires. Cette variante tronquée de la classification ascendante hiérarchique (CAH) tire parti de la composante temporelle des données pour accélérer le processus de fouille. Elle calcule des clusters de tweets proches d'une CAH classique (en fonction des paramètres choisis, 97 à 98 % des tweets sont regroupés dans les mêmes clusters) en

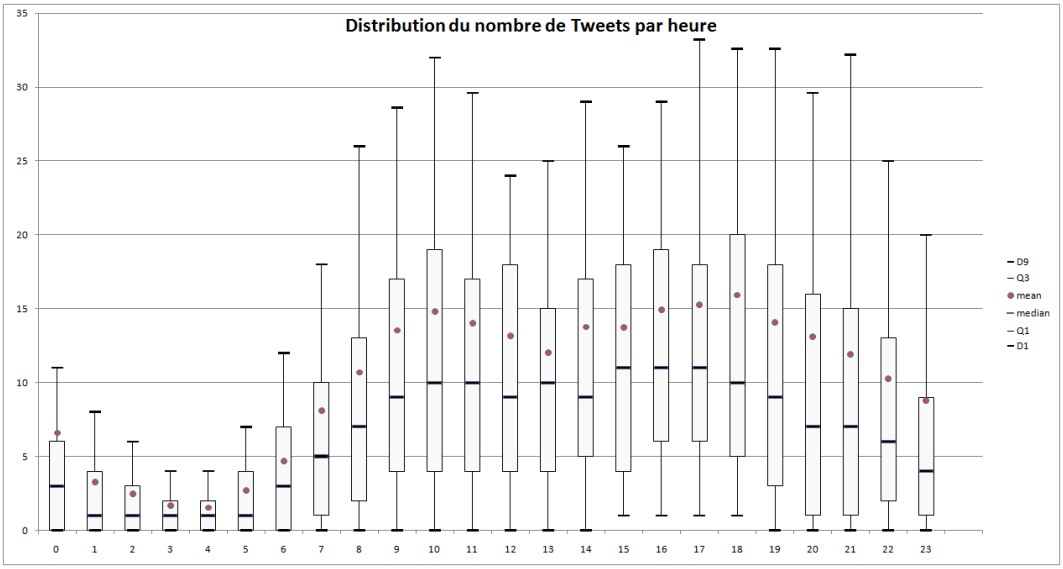


FIGURE 3.5 – Représentation des distributions du nombre de tweets par heure sous forme de diagrammes en boîte. Contrairement au reste de la journée, les distributions du milieu de la nuit montrent de faibles disparités.

un temps de calcul diminué d'un ordre de grandeur. Au lieu de calculer à chaque étape les distances entre tous les clusters pris deux à deux, cette variante ne calcule que les distances entre un cluster et ceux contenus dans une petite fenêtre de temps autour des jours de publication des tweets du cluster. Puisque, à chaque étape les dates de publication des nouveaux tweets sont ajoutées à celle du cluster, la fenêtre peut s'élargir d'étape en étape. En indiquant un seuil de similarité élevé ($> 0{,}9$), cette approche permet de regrouper les tweets similaires.

La grande similarité entre les clusters calculés par cette variante et ceux calculés par une CAH montre que la contrainte temporelle n'en est pas vraiment une. En effet, comme illustré sur la figure 1.1, les clusters sont concentrés sur une courte période de temps qui ne dure que rarement plus d'une journée.

Résultats. Le tableau 3.1 présente 10 des plus gros clusters décrits par leurs tailles et un tweet représentatif sélectionné parmi les tweets similaires du cluster. On y retrouve la thématique "Fin du monde" qui a agité les médias dans leur ensemble en fin d'année 2012, ainsi que des faits d'actualité (fermeture de la centrale de Fessenheim, location de véhicules électriques) et des messages humoristiques fréquents sur Twitter (#ChezLesAntillais). L'analyse de la répartition dans le temps des tweets similaires, représentée sur la figure 1.1, illustre le contexte de forte nouveauté des données Twitter. Parmi les 255 plus gros clusters représentés sur cette visualisation, seulement quatre durent sur plusieurs semaines (voir figure 3.6). Ces quatre clusters correspondent à des tweets relayant des traits d'humour ayant agité Twitter fin 2012 lorsque le calendrier Maya annonçait, soi-disant, la fin du monde programmée pour le 21 décembre de cette année-là. Plusieurs tweets humoristiques

ont été fortement retweetés durant les semaines précédant cette date et on remarque sur cette visualisation que l'engouement populaire a pris fin le 21 décembre 2012, à la date supposée de la fin du monde. Pour le reste, on observe que l'ensemble des thématiques est renouvelé quotidiennement et que l'analyse des tweets et de leurs retweets ne permet pas de suivre l'évolution de concepts au cours du temps. L'analyste est alors amené à suivre au jour le jour les sujets les plus populaires qui s'enchaînent sans lien les uns avec les autres.

Tweet représentatif du cluster	**# de tweets**
Le 21 décembre faudrait qu'EDF coupe l'électricité en France, juste pour faire flipper les gens #FinDuMonde#Tooozzz	1685
Si je travaillerai chez EDF, le 21/12/2012 je couperai le courant 10 minutes, juste pour faire flipper les gens mdrrrrrrrrr	1547
J'aimerais que le 21 décembre EDF éteint l'électricité de toutes les maisons de France pendant 10 minutes juste pour voir les gens flipper !	678
Moi je dis pour le 21 Décembre, EDF va tout éteindre pendant 10minutes pour faire flipper les gens.	525
EDF réclame 2 milliards d'euros d'indemnités pour la fermeture de Fessenheim www.lemonde.fr/planete/article/2012/09/16/edf-reclame-2-milliards-d-euros-d-indemnites-pour-la-fermeture-de-fessenheim_1761018_3244.html	402
Quand tu es petit t'as peur du noir. Mais crois moi quand tu verras tes factures EDF c'est de la lumière que t'auras peur	325
EDF, McDonald's, Airbus... ces entreprises qui créent encore des emplois en France www.lemonde.fr/emploi/article/2013/01/01/edf-mcdonald-s-airbus-ces-entreprises-qui-creent-encore-des-emplois-en-france_1811744_1698637.html	185
EDF se lance dans la location de voitures électriques www.leparisien.fr/automobile/edf-se-lance-dans-la-location-de-voitures-electriques-26-09-2012-2182328.php	161
#ChezLesAntillais quand t'oublie d'éteindre la lumière t'entend ta mère crier : " Limiè la ! Papaaaaaaaaw ka travail Edf ?"	156
250€ remboursés sur votre facture EDF www.leparisien.fr/economie/250-eur-rembourses-sur-votre-facture-edf-29-11-2012-2365495.php	148

TABLE 3.1 – Tableau présentant 10 tweets représentatifs de leurs clusters ainsi que la taille du cluster.

Discussion. L'analyse des groupes de tweets similaires permet donc de mettre en avant des thématiques du corpus correspondant à des instantanés, mais ne permet pas de suivre leurs évolutions au cours du temps car le facteur nouveauté est trop important. Les discussions s'enchaînent sans continuité et l'analyste est contraint d'observer une succession de tweets populaires sans rapport les uns avec les autres.

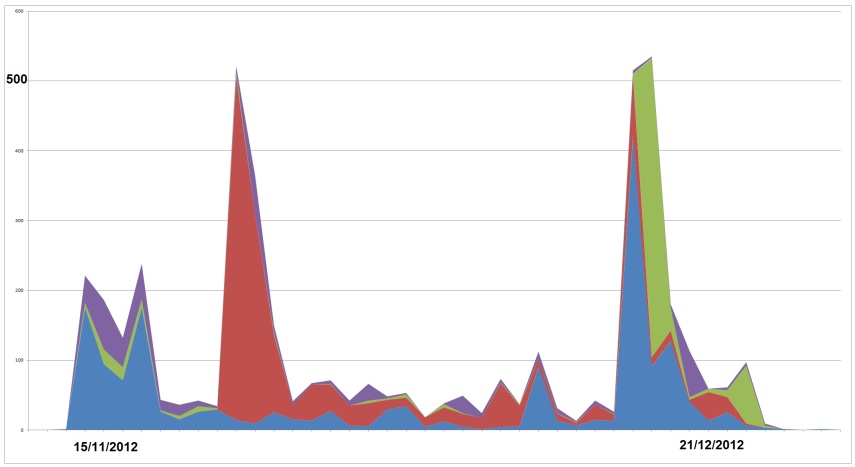


FIGURE 3.6 – Graphique sous forme d'aires empilées représentant la répartition dans le temps de 4 clusters répartis sur une période de 6 semaines. Le volume des aires correspond au nombre de tweets publiés par période.

Puisqu'il n'est pas possible de suivre l'évolution temporelle de groupes de tweets, nous proposons dans la suite de descendre à un niveau de granularité plus fin : celui des termes. En effet, l'étude de l'évolution temporelle de chaque terme indépendamment relâche les contraintes et pourrait permettre de suivre l'évolution dans le temps de groupes de termes.

5 Extraction de motifs fréquents

L'extraction de motifs fréquents est une deuxième approche classique en fouille de données catégorielles. Initialement développée dans le cadre de l'analyse des transactions dans un point de vente afin de trouver des règles d'association entre les achats de différents produits [AIS93], elle permet, appliquée à nos données, de déterminer des groupes de termes qui apparaissent souvent dans les mêmes tweets.

Définition 1. Étant donné un ensemble de tweets $\mathcal{D}$ écrits avec un vocabulaire $\mathcal{V}$, un *motif* est un sous-ensemble de $\mathcal{V}$. Un tweet $\mathbf{d}$ est *décrit* par un motif $\mathbf{P}$, et on note $\mathbf{P} \sqsubset d$, quand $\forall w \in \mathbf{P}, w \in d$. *L'ensemble des motifs* est noté $\mathcal{P}$ et est muni de la relation d'ordre d'inclusion $\subseteq$. Le *support* d'un motif $\mathbf{P}$ est l'ensemble des tweets décrits par $\mathbf{P}$: $support(\mathbf{P}) = \{d \in \mathcal{D} \mid \mathbf{P} \sqsubset d\}$. La *fréquence* de $\mathbf{P}$ est le cardinal de son support : $freq(\mathbf{P}) = |support(\mathbf{P})|$. Un motif $\mathbf{P}$ est dit *fréquent* relativement à un entier $f_{min} \in \mathbb{N}$ si sa fréquence est supérieure ou égale à f_{min} : $freq(\mathbf{P}) \geq f_{min}$.

Définition 2. Le problème de *recherche des motifs fréquents* associé à $\mathcal{D}$ et f_{min} consiste à déterminer l'ensemble des motif-fréquents $\mathcal{P}_{f_{min}} \subseteq \mathcal{P}$ ainsi que les fréquences associées.

Algorithme 1 : Pseudo-code pour l'algorithme Apriori

Apriori($\mathcal{D}$, f_{min}) :
$L_1 \leftarrow$ {motifs fréquents de longueur 1}
$k \leftarrow 2$
tant que $L_{k-1} \neq \emptyset$ **faire**
  $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \mathcal{V} \setminus \{a\}\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \not\subseteq L_{k-1}\}$
  pour chaque *tweet* $d \in \mathcal{D}$ **faire**
    $C_d \leftarrow \{c \mid c \in C_k \wedge c \sqsubset d\}$
    pour chaque *candidats* $c \in C_d$ **faire**
      $freq(c) \leftarrow freq(c) + 1$
  $L_k \leftarrow \{c \mid c \in C_k \wedge freq(c) \geq f_{min}\}$
  $k \leftarrow k + 1$
retourner $\bigcup_k L_k$

Il existe dans la littérature plusieurs algorithmes pour l'extraction de motifs fréquents (*Eclat* [Zak00], *FP-Growth* [HPY00]). Leurs particularités relèvent essentiellement de l'optimisation du temps de calcul et de l'utilisation de l'espace mémoire.

L'algorithme *APriori* (voir algorithme 1), qui est le plus populaire, repose sur un parcours par niveau de l'ensemble des motifs et sur la notion de motif candidat [AS+94]. Un *niveau* L_k de $\mathcal{P}$ correspond à l'ensemble des motifs de longueur k, c'est-à-dire composés de k éléments de $\mathcal{V}$; la recherche de motifs se fait dans l'ordre croissant des niveaux (parcours en largeur). Un *motif candidat* (de longueur k) est un motif $\mathbf{P}$ dont tous les motifs qui lui sont strictement inclus sont fréquents : $\forall \mathbf{P}' \in \mathcal{P}, \mathbf{P}' \subset \mathbf{P} \Rightarrow freq(\mathbf{P}') \geq f_{min}$. Un motif candidat est un motif pour lequel on ne peut prévoir qu'il n'est pas fréquent. Tout motif candidat doit donc être recherché dans les données en vue de déterminer sa fréquence. On note C_k l'ensemble des motifs candidats de longueur k.

La notion de motif candidat s'appuie sur l'observation suivante : si un motif est fréquent alors tous ses sous-motifs sont fréquents. Autrement dit, s'il existe un sous-motif qui n'est pas fréquent alors le motif ne peut pas être fréquent. Cette considération permet de diminuer significativement le nombre de motifs évalués.

Résultats 1. Le tableau 3.2 présente les motifs fréquents de taille supérieure ou égale 4 calculés sur notre corpus pour une valeur de seuil de 1%. Comme on peut le voir, la thématique "Fin du monde" est omniprésente dans ces résultats et en particulier les tweets proposant (avec humour) de couper le courant pour faire "flipper les gens" (voir tableau 3.1). Ce comportement, que l'on pourrait qualifier "d'arbre qui cache la forêt" vient du fait que les retweets déséquilibrent les données en multipliant le nombre de textes identiques. En effet, les associations entre termes sont fortes, les mêmes groupes de termes apparaissent de nombreuses fois ensemble sans

occurrence de termes complémentaires. Dans ce cadre, les motifs fréquents correspondent aux tweets les plus retweetés, et pour chacun, tout le tweet est fréquent. Ainsi, l'extraction des motifs fréquents revient à mesurer l'audience des tweets et ne permet pas d'extraire de connaissances supplémentaires.

Résultats 2. Afin de contourner cette limitation, nous proposons une étude complémentaire consistant à minimiser l'impact des retweets en ne prenant pas en compte la popularité d'un tweet dans l'analyse. Dans cette seconde approche, seuls les tweets initiaux sont utilisés dans le calcul des motifs fréquents. Comme l'illustre le tableau 3.3, les motifs ainsi calculés appartiennent au plus au niveau 2 de l'algorithme Apriori et correspondent à des notions très spécifiques de la relation avec un producteur d'électricité ("payer facture", "coupure courant"). Ce résultat s'explique par la faible intersection dans le vocabulaire employé dans différents tweets et illustre le contexte de forte nouveauté présent sur Twitter : les sujets de conversation s'enchaînent sans lien syntaxique. De façon à compléter cette étude, nous proposons un troisième ensemble de motifs extrait en diminuant le seuil à 0,5% afin de détecter des motifs de taille supérieure. Comme le montre le tableau 3.4, le support des motifs devient alors très faible et la proportion de documents représentés est alors loin d'être représentative de l'ensemble du corpus.

Discussion. D'après les analyses que nous avons conduites, l'extraction de motifs fréquents ne semble pas adaptée à l'étude de données déséquilibrées telles que les tweets. En effet, les retweets déséquilibrent les données en multipliant le nombre de documents identiques, ce qui rend impossible l'extraction de motifs informatifs. En prenant l'analogie avec le contexte du caddie de supermarché dans lequel l'extraction de motifs fréquents a été fortement employée, le système de retweets reviendrait à multiplier le nombre de caddies exactement identiques, à l'article prêt. Les consommateurs ne choisiraient alors plus librement leurs produits, mais choisiraient, à la place, des "paniers garnis", composés de différents ensembles de produits. L'analyse des passages en caisses enregistreuses ne nécessite plus alors l'extraction de motifs fréquents afin de déterminer, sur l'ensemble des produits du magasin, lesquels sont fréquemment achetés ensemble, mais consiste à déterminer quels paniers garnis sont les plus populaires. En somme, l'extraction des motifs fréquents appliquée aux tweets revient à étudier la popularité des tweets en termes de retweets. A l'inverse, lorsque les retweets sont retirés pour ne garder qu'un représentant par tweet, les recouvrements entre tweets se font rares, et les motifs calculés sont pauvres en information. Les supports nécessaires à l'extraction de motifs non triviaux sont très faibles et ne permettent pas de généraliser les concepts extraits. Ce comportement semble mettre en avant que, dans le cadre des données "edf", Twitter n'est pas un lieu d'échanges ou de communication entre utilisateurs, mais plus un

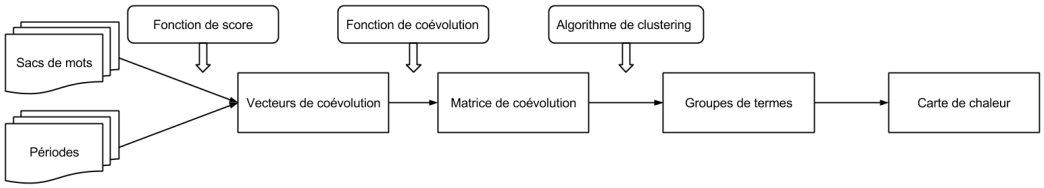


FIGURE 3.7 – Résumé de la chaîne de traitement de l'analyse de la co-évolution des termes dans les tweets et des différents paramètres.

outil de diffusion d'informations. Les utilisateurs reprennent et diffusent tels quels les tweets à leurs abonnés, sans modifications et sans entrer dans des conversations qui entraîneraient automatiquement une reprise du vocabulaire.

6 Analyse de la co-évolution des termes

Les études précédentes ont montré que l'analyse de la co-occurrence des termes dans un tweet était une contrainte trop stricte pour modéliser des thématiques sur Twitter. Nous avons donc ensuite tenté de modéliser le contenu sémantique d'un corpus de tweets en analysant la co-évolution des nombres d'occurrences des termes au cours du temps, sans exiger que ces termes apparaissent nécessairement au sein des mêmes tweets.

Étant donnés une fonction de score, une fonction de co-évolution et un algorithme de classification, notre méthodologie consiste à regrouper les termes présentant des comportements similaires pour visualiser leur évolution, et ainsi mettre en avant des motifs de co-évolution. La chaîne de traitement est représentée sur la figure 3.7 et peut être résumée ainsi : à partir de la représentation des tweets sous forme de sacs de mots et de leur période de publication, la fonction de score est utilisée pour construire, pour chaque terme, un vecteur appelé "vecteur d'évolution" ayant pour longueur le nombre de périodes et contenant le score associé à ce terme pour chaque période. Dans un second temps, la fonction de co-évolution est utilisée pour construire la matrice de co-évolution qui représente la similarité temporelle de chaque paire de termes. Enfin, l'algorithme de classification hiérarchique est appliqué à cette matrice pour regrouper les termes qui partagent un comportement commun. Finalement, un ordre compatible avec le dendrogramme fourni par cet algorithme est retenu pour classer les vecteurs d'évolution des termes et les visualiser sur une carte de chaleur.

6.1 Fonction de score

La fonction de score n'est pas imposée par notre méthodologie, son choix dépend de l'application et de la nature des données ; elle peut être, par exemple, une valeur

binaire correspondant au critère présence/absence ou une normalisation de type tf-idf.

Dans cette expérimentation, nous nous intéressons à l'évolution de la popularité des thématiques sur Twitter. Ainsi, en reprenant les notations de la section 3.3, l'intervalle de temps $\mathcal{T}$ est découpé en périodes $\mathcal{T}_1, \ldots, \mathcal{T}_n$ et le score de chaque terme $t \in \mathcal{V}$ pour chaque période est défini comme le nombre de tweets contenant t publiés pendant cette période.

La fonction de score s est définie par :

$$\begin{aligned} s: \quad \mathcal{V} \times \{1, \ldots, n\} &\to \mathbb{N} \\ t \times i &\mapsto N(t, i) \end{aligned} \tag{3.1}$$

où $N(t, i)$ est le nombre de tweets contenant le terme t publiés pendant la période $\mathcal{T}_i$.

6.2 Co-évolution des termes

Pour chaque terme $t \in \mathcal{V}$, nous définissons $\vec{ev}(t)$, le vecteur d'évolution de t tel que, $\forall i \in \{1, \ldots, n\}$ $\vec{ev}_i(t) = s(t, i)$. Pour chaque paire de termes (t_1, t_2), la co-évolution de t_1 et t_2 est définie comme la co-évolution de leurs vecteurs d'évolution $(\vec{ev}(t_1), \vec{ev}(t_2))$. Dans la suite, la co-évolution de deux vecteurs est calculée par un score de corrélation, mais en fonction des besoins de l'analyste, la fonction de co-évolution peut également être une mesure de similarité comme la similarité cosinus.

Dans cette thèse, la corrélation de chaque paire de termes (t_1, t_2) est calculée à l'aide du coefficient des rangs de *Kendall*. Contrairement au coefficient de *Pearson* et à celui de *Spearman*, le coefficient de *Kendall* présente l'avantage de détecter, non seulement, les relations linéaires entre termes, mais également, les relations non linéaires.

Définition 3. Soit $(i, j) \in \{1, \ldots, n\}^2$ une paire de coordonnées, on dit que (i, j) est une *paire de coordonnées concordantes* entre les vecteurs $\vec{ev}(t_1)$ et $\vec{ev}(t_2)$, si et seulement si, $\vec{ev}_i(t_1) < \vec{ev}_i(t_2)$ et $\vec{ev}_j(t_1) < \vec{ev}_j(t_2)$, ou $\vec{ev}_i(t_1) > \vec{ev}_i(t_2)$ et $\vec{ev}_j(t_1) > \vec{ev}_j(t_2)$.

Définition 4. Le *coefficient de Kendall* τ, des vecteurs $\vec{ev}(t_1)$ et $\vec{ev}(t_2)$ est défini par :

$$\tau(\vec{ev}(t_1), \vec{ev}(t_2)) = \frac{N_c - N_d}{\frac{1}{2}n(n-1)} \tag{3.2}$$

où N_c est le nombre de paires de coordonnées concordantes entre $\vec{ev}(t_1)$ et $\vec{ev}(t_2)$ et N_d est le nombre de coordonnées discordantes. Le dénominateur correspond au nombre total de combinaisons.

Le résultat du calcul de la co-évolution est une matrice carrée M_S, de taille $|\mathcal{V}|$, appelée matrice de co-évolution. Les lignes et les colonnes de M_S correspondent aux termes ordonnés de la même manière et $\forall i, j \in \{1, \ldots, |\mathcal{V}|\}$,

$$M_S(i,j) = \tau(\vec{ev}(t_i), \vec{ev}(t_j)) \tag{3.3}$$

A cette étape, l'ordre dans lequel les termes sont présentés dans M_S ne correspond à aucun comportement particulier. Selon la méthode utilisée pour construire $\mathcal{V}$, il peut être par exemple l'ordre alphabétique ou l'ordre dans lequel les termes apparaissent dans les documents. L'objectif de l'étape suivante est de réorganiser les termes de $\mathcal{V}$ en fonction de leurs co-évolutions de façon à faire apparaître des motifs.

6.3 Classification

Afin de construire des groupes de termes similaires, un algorithme de classification ascendante hiérarchique (CAH) est appliqué aux lignes de la matrice M_S représentant les termes (ou symétriquement sur les colonnes). Cet algorithme est également un paramètre de la méthodologie et le critère de regroupement peut notamment influencer les résultats. Dans cette thèse, à chaque étape de la classification, les deux classes les plus proches sont regroupées selon le critère du saut maximum qui produit des classes compactes et est moins coûteux en temps de calcul que le saut moyen. Pour chaque paire de classes C_1, C_2,

$$sim(C_1, C_2) = \min_{x \in C_1, y \in C_2} (sim(x,y)) \tag{3.4}$$

Le résultat de cette classification est ensuite utilisé pour réorganiser les termes et visualiser les termes similaires proches les uns des autres (voir figure 3.8).

6.4 Carte de chaleur

Déterminer l'ordre optimal à utiliser pour afficher le résultat d'une classification ascendante hiérarchique, c'est-à-dire celui minimisant la somme des distances entre instances adjacentes, est très coûteux en temps de calcul ([BJGJ01]). Or, la tâche proposée dans cette étude consiste uniquement à détecter des groupes de termes évoluant conjointement. Un ordre compatible avec le résultat de la classification suffit à mettre en avant les motifs recherchés. Cet ordre est utilisé pour construire une carte de chaleur représentant l'évolution des scores des termes au cours du temps. Les colonnes de cette carte de chaleur correspondent aux n périodes affichées dans l'ordre chronologique, les lignes correspondent aux $|\mathcal{V}|$ termes réordonnés par la CAH et le gradient de couleur est défini pour correspondre aux scores fournis par

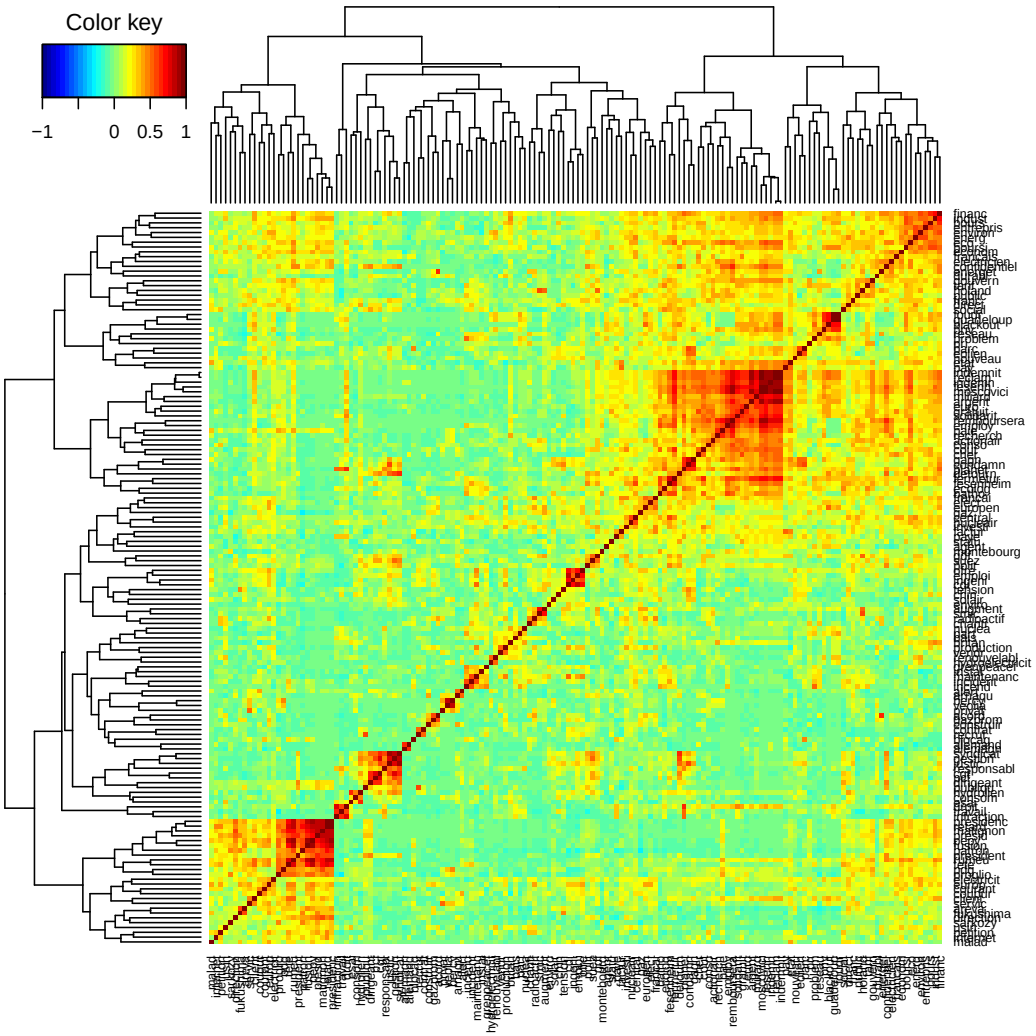


FIGURE 3.8 – Exemple de matrice de corrélations avec le dendrogramme associé. Des blocs apparaissent, mettant en avant des groupes de termes fortement corrélés.

la fonction de score pour chaque terme et chaque période.

Résultats. Plusieurs expérimentations ont été conduites en partenariat avec un expert du traitement du langage naturel et de la gestion de la relation client du groupe EDF. Nous lui avons demandé de détecter des groupes de termes à l'aide de la carte de chaleur construite par notre méthodologie. Nous présentons ici les résultats obtenus en étudiant les bigrammes (paires de mots consécutifs) en tant qu'unités syntaxiques. La figure 3.9 présente un extrait de la carte de chaleur visualisée par l'expert, construit pour un nombre restreint de bigrammes et centré sur leur période d'activité. Cet extrait permet de visualiser trois types de comportements distincts que l'on retrouve à plusieurs niveaux sur la carte complète :

Un premier groupe de termes correspond aux thématiques courtes décrites par [KLPM10], ces termes saillants n'apparaissent que quelques jours (autour du 17 Août 2012) et ne sont plus utilisés ensuite. Symptomatiques du contexte de forte

nouveauté inhérent à Twitter, ils correspondent à des messages rapportant une rumeur de remplacement à la tête de la direction d'EDF du président Henri Proglio par Guillaume Pepy, actuellement directeur de la SNCF. Cette rumeur a été relayée sur les médias traditionnels et n'est pas une nouveauté pour notre expert. Toutefois, savoir qu'elle a été discutée sur Twitter et pouvoir estimer dans quelles proportions, reste à ses yeux une information utile. Néanmoins, l'apport de notre méthodologie sur ce type de motifs est assez faible puisqu'elle apporte peu de valeur ajoutée par rapport à l'analyse des groupes de tweets similaires.

Le second groupe de termes présente une activité qui s'étend sur une période plus longue et qui est associée à un volume de messages plus important. Ils correspondent à la thématique de fin du monde déjà évoquée précédemment. Cette visualisation permet de distinguer deux types de termes parmi le vocabulaire employé pour décrire cette thématique. Certains bigrammes sont spécifiques à cette thématique, comme "flipper gens", alors que d'autres sont également employés dans d'autres contextes comme le montrent les cellules foncées réparties tout au long de la période. C'est par exemple le cas des bigrammes "couper courant" ou "éléctricité france".

Enfin, le troisième groupe correspond à des termes qui sont employés ensemble à plusieurs reprises durant la période considérée, créant un motif de répétition. Ces termes correspondent à un ensemble de tweets dans lesquels les utilisateurs demandent, sur le ton de l'humour, à ce que les tarifs appliqués par l'entreprise Free dans la téléphonie le soient également dans d'autres domaines, en particulier dans celui de l'électricité. Notre expert porte un intérêt à ces messages qui font partie des signaux faibles liés aux tarifs de l'énergie. Le fait que ce sujet soit répété à plusieurs reprises, contrairement à la blague sur la fin du monde, montre qu'il n'est pas anecdotique. Même si le ton est léger, les utilisateurs de Twitter sont attentifs aux prix pratiqués par EDF. Pour ce groupe, l'apport de notre méthodologie consiste à mettre en avant des signaux faibles qui n'avaient pas été détectés par les méthodes employées jusqu'à présent. Le volume de chaque occurrence de ces thématiques est trop faible pour être détecté par d'autres approches, néanmoins, mis bout à bout, le volume et la persistance de cette thématique témoignent d'un intérêt non négligeable des utilisateurs de Twitter.

Discussion. Contrairement aux méthodes de clustering et d'extraction de motifs fréquents, la visualisation des co-évolutions a permis de détecter à la fois différents types de termes au sein d'une thématique et des sujets de conversation récurrents alors même qu'ils sont relativement peu exposés sur Twitter, au regard du nombre de tweets. Cependant, si l'on met de côté ce type de co-évolution, la grande majorité des groupes de termes observés sur la carte de chaleur sont concentrés sur un unique intervalle de temps (qui souvent ne dure qu'une journée). Ceci ne permet

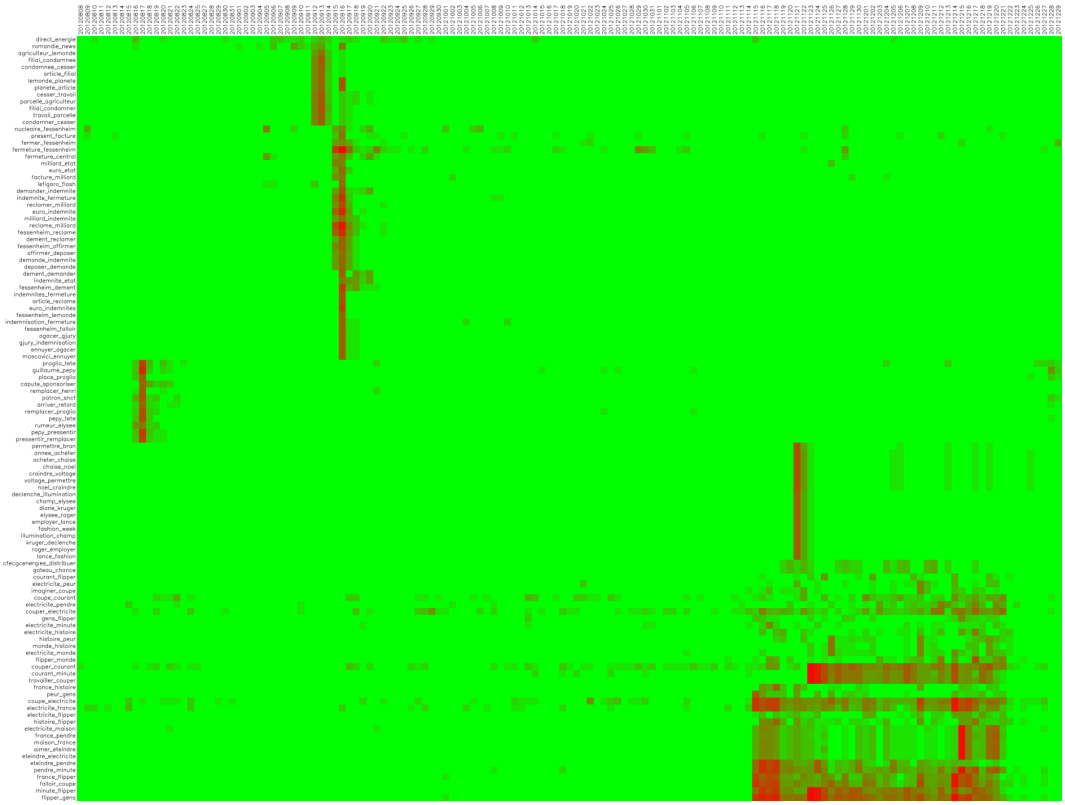


FIGURE 3.9 – Zoom sur la carte de chaleur illustrant les motifs de pics, la continuité de la thématique "fin du monde" et quelques motifs de ré-occurrence.

pas de suivre une véritable évolution du contenu sémantique du corpus de tweets. Encore une fois, la nouveauté est trop forte sur Twitter et la principale conclusion de cette approche est que le vocabulaire employé sur Twitter semble renouvelé quotidiennement et ne présente pas de continuité. Afin de proposer à l'analyste une méthode permettant de suivre l'évolution de concepts au cours du temps, nous proposons de nous placer à un niveau d'abstraction supérieur à celui des termes en modélisant des thématiques suffisamment générales pour être réemployées.

7 Conclusion

Dans ce chapitre, nous avons présenté la démarche de fouille d'exploration d'un corpus de tweets développée en collaboration avec un expert de la Direction Recherche et Développement d'EDF. En suivant le schéma classique d'une chaîne de traitement de données, nous avons commencé par décrire le processus de collecte qui interroge en continu les bases de données de Twitter, puis nous avons exposé les pré-traitements appliqués afin de transformer des tweets bruts en données exploitables. Nous avons ensuite proposé d'appliquer deux méthodes classiques de la fouille textuelle : le clustering de documents et l'extraction de motifs fréquents dans la co-occurence des termes. Les résultats obtenus avec ces deux approches, qui se placent à des niveaux d'analyse différents, ont confirmé l'impact de la forte nou-

veauté sur les approches classiques et ont révélé la difficulté d'identifier de la "continuité" dans les thématiques extraites. Pour tenter une analyse de l'évolution sur le long terme, nous avons proposé une approche originale reposant sur une visualisation sous la forme d'une carte de chaleur des groupes de termes aux comportements "similaires". Cette représentation a permis à l'analyste d'identifier des motifs de répétition non attendus. Cependant, le contexte de forte nouveauté limite aussi la portée de cette approche car les groupes construits se rapportent à un nombre faible de tweets. Puisque le système de retweets et le renouveau permanent permettent difficilement de suivre l'évolution temporelle des thématiques à l'échelle des termes et à l'échelle des documents, nous proposons, dans le chapitre suivant, de nous placer à un niveau d'abstraction intermédiaire en analysant les thèmes latents construits grâce au "topic model" LDA (Latent Dirichlet Allocation).

Motifs fréquents (avec RT)	**Support (# de tweets)**
courant gens flipper couper	1516
gens flipper couper minute	1514
courant flipper couper minute	1513
courant gens couper minute	1505
courant couper minute travailler	1503
courant gens flipper minute	1496
courant gens flipper couper minute	1495
electricite gens coupe falloir	1476
courant flipper couper travailler	1476
courant flipper minute travailler	1463
flipper couper minute travailler	1462
gens flipper couper travailler	1460
courant flipper couper minute travailler	1459
courant gens couper travailler	1459
courant gens flipper couper travailler	1457
gens flipper couper minute travailler	1445
courant gens flipper couper minute travailler	1443
electricite gens flipper coupe	1404
electricite flipper coupe falloir	1404
electricite france coupe falloir	1385
gens flipper coupe falloir	1384
electricite france gens falloir	1379
electricite gens flipper falloir	1379
france gens coupe falloir	1372
electricite france gens coupe	1371
electricite france gens coupe falloir	1367
electricite gens flipper coupe falloir	1367
electricite france gens flipper	1356
electricite france flipper falloir	1299
electricite france flipper coupe	1297
france flipper coupe falloir	1292
france gens flipper falloir	1290
electricite france flipper coupe falloir	1287
electricite france gens flipper falloir	1285
france gens flipper coupe	1280
france gens flipper coupe falloir	1278

TABLE 3.2 – Tableau présentant les motifs fréquents de taille supérieure ou égale à 4 calculés sur l'ensemble des tweets pour un seuil de 1% ainsi que leurs supports.

Motifs fréquents (sans RT)	**Support (# de tweets)**
facture payer	482
courant coupure	389
courant couper	298
electricite couper	276
electricite heure	251
electricite coupure	241
facture recevoir	196
electricite coupe	183
nucleaire central	178
electricite france	176
mail recevoir	170
electricite facture	163
heure coupure	158
proglio henri	154
courant heure	152
courant coupe	152
onpc bayrou	149

TABLE 3.3 – Tableau présentant les motifs fréquents de taille supérieure ou égale à 2 calculés sur les tweets sans la prise en compte des retweets pour un seuil de 1% ainsi que leurs supports.

Motifs fréquents (sans RT)	**Support (# de tweets)**
etre entrer contact viadeo	67
courant heure coupure	53
finance maroc otmane rhazi	44
heure technique ingenierie	44
facture mail recevoir	42
facture payer mail	41
facture onpc bayrou	41
electricite priver foyer	39
mail faux arnaque	37
privatiser dominique villepin	36
courant entrer passe	35
lumiere allumer laisse	34
proglio conseiller privatiser	34
electricite heure coupure	32
heure entrer passer	32
mail faux attention	32
mail attention arnaque	31
electricite facture payer	30
facture payer falloir	30
proglio dominique villepin	30
lumiere pere allumer	29

TABLE 3.4 – Tableau présentant les motifs fréquents de taille supérieure ou égale à 3 calculés sur les tweets sans la prise en compte des retweets pour un seuil de 0,5% ainsi que leurs supports.

4

Topic Modeling

If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.

—John von Neumann

Sommaire

1 Introduction

À l'origine, les modèles de thèmes ("topic models") ont été développés dans le cadre de la recherche d'information et ont montré de bonnes performances en termes de précision et rappel [YA09]. Ils ont depuis été fortement utilisés en extraction de connaissances. Le modèle de Latent Semantic Indexing (LSI) [DDL+90] est probablement le premier "topic model" à l'origine de la représentation des documents dans un espace latent. Il introduit les concepts de réduction de dimensions que l'on retrouve dans le modèle LDA : la matrice termes×documents est décomposée en valeurs singulières et les k plus grandes valeurs sont retenues pour définir les concepts latents. Comme l'illustre la figure 4.1, les documents sont représentés par ces concepts qui sont eux-mêmes représentés par des termes. Malgré ses bonnes performances, le manque de fondements statistiques de LSI a été critiqué et l'approche "probabilistic Latent Semantic Indexing" (pLSI) a été proposée [Hof99]. Le modèle pLSI étend le modèle LSI en offrant un cadre probabiliste à la modélisation des concepts latents. Ce modèle suppose que chaque mot d'un document a été généré selon un modèle de mélange composé de distributions multinomiales sur le vocabulaire (les concepts latents). Chaque document est représenté par une distribution sur ces concepts. Ce premier pas vers la représentation probabiliste de documents textuels a ensuite été complété dans le modèle Latent Dirichlet Allocation (LDA) [BNJ03] afin d'offrir un processus générateur au niveau des documents. En effet, pLSI n'offre pas de description justifiant la représentation des documents par un ensemble de concepts, et ne permet pas de généraliser le modèle à des documents autres que les données d'apprentissage.

Ces trois approches font deux hypothèses importantes sur les données, si la première est classique en analyse de données textuelles, la seconde relève d'une importance particulière dans l'étude d'un flux de texte, et tout particulièrement dans le contexte de forte nouveauté de Twitter. La première, souvent nommée hypothèse du "sac de mots", suppose que les mots sont interchangeables dans les documents. Autrement dit, l'ordre des mots n'est pas pris en compte lors de l'estimation des thèmes latents. La structure grammaticale des documents est ignorée et c'est à l'analyste de reconstituer mentalement les phrases à partir de listes de termes. Cette hypothèse est assez courante (elle est aussi faite par les algorithmes d'extraction de motifs fréquents et les algorithmes de clustering de textes représentant les documents sous forme de vecteurs de mots) et est compatible avec les pré-traitements fréquemment effectués sur les données textuelles : lemmatisation, stop liste, etc. En effet, puisque la structure grammaticale est ignorée, les mots outils (tels que les prépositions ou les verbes d'état) et les accords et conjugaisons ne sont plus nécessaires à la compréhension des documents. Seuls les mots porteurs de sens sont conservés afin de modéliser les documents. La seconde hypothèse suppose que les documents sont in-

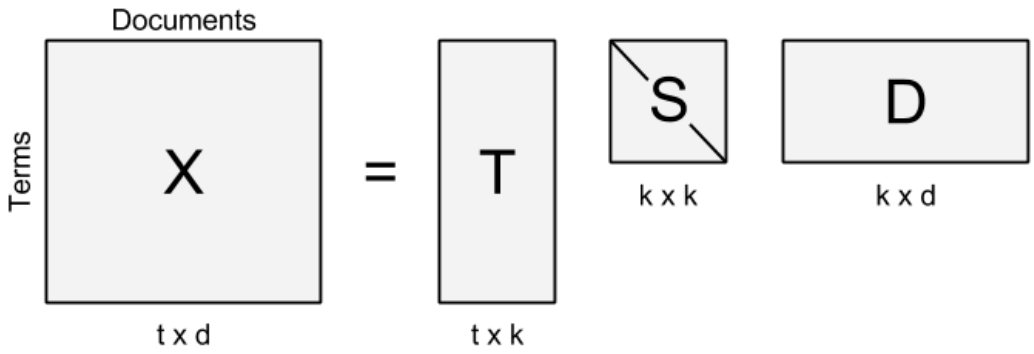


FIGURE 4.1 – Décomposition en valeurs singulières de la matrice termes×documents X. Les colonnes des matrices T et D sont unitaires et orthogonales, S est la matrice diagonale contenant les valeurs propres, t est le nombre de termes dans le vocabulaire, d est le nombre de documents et k est le nombre de concepts sélectionnés.

terchangeables au sein du corpus, c'est-à-dire que l'ordre des documents est ignoré et que l'information sur la date de publication n'est pas prise en compte. Ces approches sont donc dédiées à la modélisation de corpus statiques, dont le contenu n'est pas censé évoluer au cours du temps. Les thèmes construits sont donc représentatifs d'un corpus dans son ensemble et employer un modèle construit sur un ensemble de documents pour modéliser ceux d'un autre ensemble suppose qu'ils partagent les mêmes thématiques.

Dans ce chapitre, nous commençons par donner, dans la section 4.2, une présentation détaillée du modèle classique LDA décrivant le processus générateur permettant de généraliser l'approche pLSI à de nouveaux documents. Nous complétons la description par une présentation d'une méthode de la famille des chaînes de Markov-Monte-Carlo [GRS96] utilisée pour l'estimation des paramètres du modèle. Nous complétons la présentation théorique pas une expérimentation du modèle classique sur nos données. Dans la section 4.3, nous décrivons une méthode (Dynamic-LDA) proche de celle proposée par TopicFlow [MSH+13] que nous avons développé pour adapter les modèles de thèmes statiques à un corpus dynamique. Nous décrivons notamment une procédure pour comparer des thèmes construits par des modèles différents sur des vocabulaires différents. La description des thèmes se faisant sous la forme de distributions de probabilités représentées par des vecteurs, de nombreuses mesures peuvent être utiliser pour comparer les thèmes. Puisqu'aucune de ces mesures ne semblent, *a priori*, plus adaptée qu'une autre, nous proposons, dans la section 4.4, une comparaison expérimentale des distributions empiriques de sept mesures de divergence adaptées à la comparaison de distributions de probabilités. Cette comparaison nous a guidés dans le choix d'une mesure adaptée au suivi longitudinal des thèmes.

2 Latent Dirichlet Allocation

Afin de modéliser les thématiques abordées dans un corpus de tweets, nous proposons d'utiliser le modèle de Latent Dirichlet Allocation (LDA) dont les nombreuses variantes occupent une part de plus en plus importante des travaux sur la fouille de données textuelles [RZGSS04, RHNM09, WLD+14]. Son application dans un contexte industriel semble également déjà effective puisque même si les algorithmes de classement des moteurs de recherche sont une information hautement confidentielle, une analyse de corrélation entre les thèmes construits à partir des résultats de recherche fournis par Google et ceux construits à partir de la requête originelle[1] tend à montrer que LDA est depuis quelques années un ingrédient de la recette Google.

LDA est un modèle probabiliste génératif à variable latente qui suppose que la présence des termes dans les documents est imputable à une variable cachée : les thèmes (appelés "topics" en anglais). Les paramètres du modèle sont le nombre de thèmes K et les paramètres des distributions *a priori* $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$, appelés hyper-paramètres. Le paramètre $\boldsymbol{\alpha}$ utilisé dans la description des documents régule leur répartition sur les thèmes : des composantes faibles de $\boldsymbol{\alpha}$ auront tendance à forcer les documents à être décrits par peu de thèmes, alors que des valeurs élevées auront tendance à donner une description plus répartie des documents. De manière analogue, le paramètre $\boldsymbol{\beta}$ est utilisé dans la description des thèmes et régule leur répartition sur les termes du vocabulaire : des valeurs faibles de $\boldsymbol{\beta}$ auront tendance à produire des thèmes spécifiques, décrits par peu de termes, alors que des valeurs plus élevées auront tendance à créer des thèmes plus homogènes, décrits par un nombre plus importants de termes.

Par ailleurs, même si le modèle est décrit en utilisant le vocabulaire de la fouille de textes, (en raison de son origine), son application n'est pas restreinte à ce domaine. LDA s'applique à n'importe quelle variable catégorielle, et a été adapté aussi bien à l'analyse d'images [LMD10, FFP05], qu'à la bio-informatique [FGK+05, PSD00] ou aux systèmes de recommandation [BNJ03].

Enfin, puisque le modèle LDA distingue les différentes occurrences d'un même terme, il est nécessaire de distinguer les termes du vocabulaire $t \in \mathcal{V}$ et les occurrences des mots dans les documents $w \in \mathcal{W}$. En effet, un même terme peut apparaître plusieurs fois dans un document et il peut également apparaître dans plusieurs documents. Chaque occurrence est alors traitée séparément. Un mot $w \in W$ correspond alors à la donnée d'un terme $t \in \mathcal{V}$ et d'un document $d \in \mathcal{D}$ dans lequel il apparaît. La position exacte du mot dans le document n'intervient pas, c'est l'hypothèse du "sac de mots".

Le modèle LDA étant un modèle probabiliste génératif, il est nécessaire de définir son processus générateur, ainsi qu'une méthode permettant d'estimer les para-

[1] https ://moz.com/blog/lda-and-googles-rankings-well-correlated

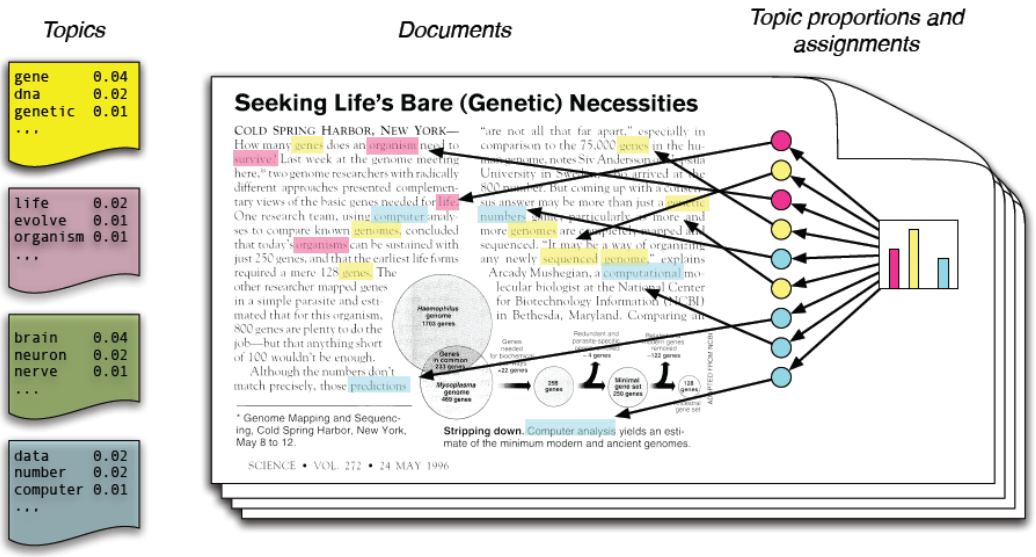


FIGURE 4.2 – Illustration du processus générateur du modèle LDA extrait de [Ble12]. Les thèmes du corpus sont représentés à gauche et la distribution du document en cours sur les thèmes est présentée à droite sous forme d'histogramme. Pour chaque mot du document (sans considération d'ordre), un thème est choisi selon la distribution du document, puis un terme est choisi selon la distribution du thème.

mètres du modèle étant donné un corpus de documents.

2.1 Processus générateur

Étant donné un nombre de thèmes K fixé *a priori*, le processus générateur est le suivant (voir figure 4.2) :

- Tirer K distributions sur les termes du vocabulaire $\mathcal{V}$ représentant les thèmes du corpus selon une loi de Dirichlet de paramètre $\boldsymbol{\beta}$:
  $\Phi = \{\boldsymbol{\phi_k}\}_{k=1}^{K} \sim Dir(\boldsymbol{\beta})$.
- Pour chaque document m :
  - Tirer un entier N_m décrivant la taille du document :
    $N_m \sim Poisson(\xi)$[2].
  - Tirer une distribution sur les thèmes représentant les thématiques du document selon une loi de Dirichlet de paramètre $\boldsymbol{\alpha}$: $\boldsymbol{\theta_m} \sim Dir(\boldsymbol{\alpha})$.
  - Pour chaque mot w_{mn} de ce document :
    * Tirer un thème selon une loi de Bernoulli multivariée de paramètres $\boldsymbol{\theta_m}$: $z_{m,n} \sim Mult(\boldsymbol{\theta_m})$.
    * Tirer un terme selon une loi de Bernoulli multivariée de paramètres $\boldsymbol{\phi}_{\boldsymbol{z_{m,n}}}$: $t_{mn} \sim Mult(\boldsymbol{\phi}_{\boldsymbol{z_{m,n}}})$.

Les thèmes et les termes sont donc générés selon une loi de Bernoulli multivariée (cas particulier pour $n = 1$ de la loi multinomiale) dont les paramètres sont eux mêmes tirés aléatoirement selon une loi de Dirichlet de paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ respectivement. Pour rappel, la distribution de Dirichlet est définie par $Dir(\boldsymbol{x} \mid \boldsymbol{\alpha}) = \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^{|\boldsymbol{x}|} x_i^{\alpha_i - 1}$ avec $\Delta(\boldsymbol{\alpha}) = \frac{\prod \Gamma(\alpha_i)}{\Gamma(\sum \alpha_i)}$ et Γ la généralisation aux nombres réels de la fonction factorielle et dans le cas $n = 1$, la loi multinomiale s'écrit simplement $Mult(\boldsymbol{x} \mid \boldsymbol{p}) = \boldsymbol{p}$.

Remarque 1. *Contrairement à ce que l'on trouve fréquemment dans la littérature anglophone et en particulier dans l'article original [BNJ03], les thèmes et les termes ne sont pas réellement générés selon une loi multinomiale mais selon le cas particulier, beaucoup plus simple, où l'expérimentation n'est reconduite qu'une seule fois. C'est pour cela que les données tirées sont un point, $z_{m,n}$ et t, et non pas un vecteur. Le point correspond en fait à la seule composante non-nulle du vecteur : celle qui a été tirée lors du seul tirage. Par analogie avec la relation entre la loi binomiale et la loi de Bernoulli, et afin de bien distinguer la loi multinomiale du cas particulier $n = 1$, cette loi est ici appelée loi de Bernoulli multivariée. Pour rappel, la loi de Bernoulli correspond au lancé d'une pièce biaisée, $\{succes, echec\}$, n'ayant que deux issues possibles dont les probabilités sont p et $1 - p$ respectivement. La loi de Bernoulli multivariée généralise cette loi en autorisant un nombre arbitraire d'issues possibles et s'illustre par exemple comme le lancé d'un dé biaisé à x faces, la contrainte étant alors que la somme des probabilités des faces vaille 1. L'analogie est donc que, comme la loi binomiale modélise le nombre de succès obtenus lors de la répétition de n tirages de Bernoulli, la loi multinomiale modélise le nombre fois où chaque face à été tirée lors de n lancés de dé indépendants.*

Remarque 2. *Malgré l'ambiguïté introduite par l'abus de langage identifiant la loi multinomiale et un de ses cas particulier, l'hypothèse selon laquelle les mots dans les documents sont générés selon une loi de Bernoulli multivariée est très commune. L'apport de LDA consiste en fait à introduire une distribution a priori sur ces probabilités permettant de généraliser le modèle à de nouveaux documents. En effet, les modèles précédents assignent une probabilité nulle aux mots nouveaux et donc aux documents nouveaux. L'introduction d'une distribution a priori permet de lisser ces probabilités. Le choix d'une loi de Dirichlet est alors naturel puisqu'elle est la loi conjuguée de la loi multinomiale (et donc de la loi de Bernoulli multivariée) ce qui permet d'estimer sans trop de difficultés "de calcul" les paramètres du modèle. Effet, il est alors possible de montrer que dans le cas de l'inférence Bayésienne, la loi a posteriori obtenue en combinant une vraisemblance multinomiale et une distribution a priori de Dirichlet suit alors, elle même, une loi de Dirichlet dont les paramètres sont connus.*

[2] Lors de l'estimation des paramètres, cette information est déduite des données.

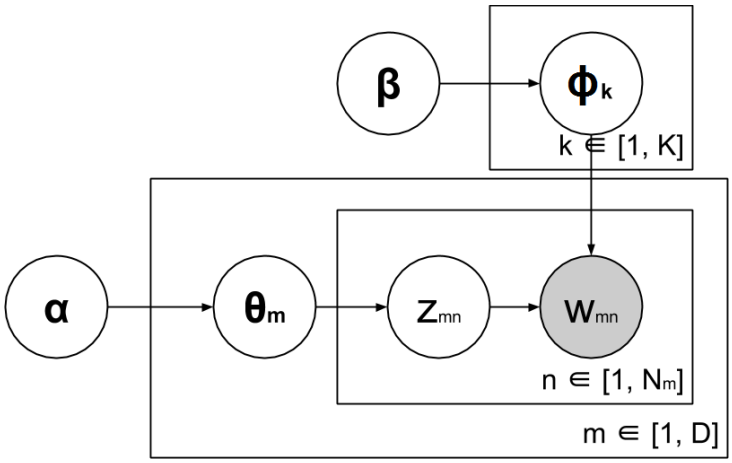


FIGURE 4.3 – Représentation graphique du modèle LDA. (Reproduit à partir de [Hei05])

La représentation graphique du modèle LDA est donnée sur la figure 4.3. Les cercles représentent les variables et les rectangles qui les entourent représentent les itérations sur l'ensemble indiqué dans le coin inférieur droit. Le cercle grisé représente la seule variable observée : les mots dans les documents. Les autres cercles représentent les variables cachées (les hyper-paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$, même s'ils sont donnés, ne font pas partie des variables observées). Ce schéma peut s'interpréter comme suit : une fois, pour l'ensemble du corpus, (c'est-à-dire indépendamment du document en cours), tirer K distributions représentant les thèmes $\{\phi_1, \ldots, \phi_K\}$. Ensuite, pour chaque document m tirer une répartition sur ces thèmes θ_m. Finalement, pour chaque mot de ce document tirer l'indice d'un thème z_{mn} ($z_{mn} \in [1, \ldots, K]$) selon les proportions θ_m propre à ce document et sélectionner un terme selon la distribution de ce thème : $\phi_{z_{mn}}$.

Dans un modèle LDA à K thèmes, la probabilité que le $n$ème mot w d'un document m (de longueur N_m) soit instancié par un terme t du vocabulaire est :

$$p(w_{m,n} = t) = \sum_{k=1}^{K} p(w_{m,n} = t \mid z_{m,n} = k) p(z_{m,n} = k) \tag{4.1}$$

La plupart des implémentations supposent que les distributions de Dirichlet sont symétriques (*i.e.* $\exists \alpha \in \mathbb{R}, \forall k \in \{1, \ldots, K\}, \alpha_k = \alpha$ et $\exists \beta \in \mathbb{R}, \forall t \in \{1, \ldots, \mathcal{V}\}, \beta_t = \beta$)), dans ce cas les hyper-paramètres sont simplement notés α et β. Cela signifie que l'on ne souhaite pas discriminer *a priori* les termes et les thèmes du modèle.

De façon à faire apparaître ces paramètres, l'équation 4.1 peut se réécrire :

$$p(w_{m,n} = t \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{k=1}^{K} p(w_{m,n} = t \mid z_{m,n} = k, \boldsymbol{\alpha}) p(z_{m,n} = k \mid \boldsymbol{\beta})$$

Par ailleurs, la distribution jointe du modèle est donnée par la formule :

$$p(\boldsymbol{w}, \boldsymbol{z}, \Theta, \Phi \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\Phi \mid \boldsymbol{\beta}) p(\Theta \mid \boldsymbol{\alpha}) p(\boldsymbol{z} \mid \Theta) p(\boldsymbol{w} \mid \Phi) \tag{4.2}$$

où $\boldsymbol{w} = \{w_{m,n}\}$ désigne les mots dans les documents, $\boldsymbol{z} = \{z_{m,n}\}$ leurs assignations aux thèmes et $\Theta = \{\boldsymbol{\theta_m}\}$ l'ensemble des distributions des thèmes pour chaque document m.

2.2 Estimation des paramètres

L'étape d'estimation consiste à déterminer les paramètres des distributions "*a posteriori*" des variables latentes étant donnés les mots observés dans les documents :

$$p(\Theta, \Phi, \boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{w}, \boldsymbol{z}, \Theta, \Phi \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}$$

Toutefois, cette probabilité n'est pas calculable en pratique, en particulier à cause du dénominateur qui met en jeu $K^{|\mathcal{W}|}$ opérations :

$$p(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{w_i \in \mathcal{W}} \sum_{k=1}^{K} p(z_i = k, w_i)$$

La solution est alors d'utiliser une méthode approchée, comme l'inférence variationnelle proposée dans l'article original [BNJ03] qui présente l'avantage d'offrir une approche rapide, mais qui est soumise au problème des maximums locaux, ou une approche de la famille des chaînes de Markov - Monte Carlo, comme le "Collapsed Gibbs Sampler" [GS04], que nous allons détailler ici, et dont la convergence vers une solution exacte est assurée, mais est coûteuse en temps de calcul. D'autres approches ont été proposées, comme la propagation de l'espérance [ML02], mais l'essentiel des efforts a été fait pour accélérer la convergence de l'échantillonnage de Gibbs [PNI+08, XS10, QWW+14, NASW09].

2.2.1 Échantillonnage de Gibbs (Gibbs sampling)

La méthode d'échantillonnage de Gibbs est de la famille des chaînes de Markov et est un cas particulier du Metropolis-Hastings [MRR+53, Has70]. Plusieurs tutoriels ont été proposés [SG07, RH10, Dar11, Gri02] et Heinrich en donne une présentation pédagogique reprenant les concepts clés [Hei05].

Chaînes de Markov - Monte Carlo. L'objectif des méthodes de la famille des chaînes de Markov - Monte Carlo est de tirer des échantillons selon une loi de probabilité P trop complexe en tirant selon une loi plus simple Q. Pour cela, on définit une chaîne de Markov dont la loi stationnaire est la loi de probabilité recherchée.

La difficulté est alors de définir une loi $Q(x_{i-1})$ qui permet de générer x_i à partir de x_{i-1} (Chaîne de Markov) et qui converge vers P.

Metropolis-Hastings. L'algorithme de Metropolis-Hastings [CG95], en particulier, offre une solution pour définir une telle loi $Q(x_{i-1})$ dans le cas où l'on sait tirer selon une loi F proportionnelle à P. (Par exemple comme c'est le cas ici, quand on ne sait pas calculer le coefficient de normalisation dans la règle de Bayes). Pour cela, on utilise une distribution intermédiaire $Q(x_{i-1})$ symétrique (par exemple une gaussienne centrée sur x_{i-1} comme fonction de transition de la chaîne de Markov, avec un coefficient d'affectation calculé selon un ratio entre deux valeurs de F). Comme F est proportionnelle à P, cela correspond à un ratio selon P. Cette technique assure que la loi stationnaire de cette chaîne de Markov sera bien P.

Plus formellement, le Metropolis-Hastings consiste à :

- 1. Initialisation :
  - - Choisir x_0 arbitrairement
  - - Choisir une distribution de probabilité symétrique $Q(x \mid y)$
- 2. Pour chaque itération t :
  - Tirer un candidat x selon $Q(x \mid x_t)$
  - Calculer le ratio d'acceptation $a = \frac{F(x)}{F(x_t)}$. (Comme $F \propto P$, $a = \frac{P(x)}{P(x_t)}$)
    - Si $a > 1$, x est plus probable que x_t et le candidat est retenu : $x_{t+1} = x$
    - Sinon, x n'est retenu qu'avec la probabilité a.
      - Si le candidat est retenu on a de la même manière : $x_{t+1} = x$
      - Sinon, on reste en x_t et on a : $x_{t+1} = x_t$

Intuitivement, cette procédure a tendance à rester plus longtemps (et donc à tirer plus d'échantillons) dans les zones les plus probables selon la loi P, même si elle permet toutefois de visiter des régions moins probables.

Remarque 3. *Il est important de noter que les premiers tirages ne sont pas forcément représentatifs de la loi P (notamment si x_0 est choisi dans une région peu probable) et ils sont généralement ignorés. On appelle cela la période de "burn-in".*

Remarque 4. *Un deuxième point important est que les tirages sont corrélés (chaîne de Markov). Pour limiter l'impact de cette corrélation, il faut rejeter la majorité des tirages et ne grader qu'un échantillon sur n, avec n> auto-corrélation. (Le calcul de l'auto-corrélation est un problème difficile, et en général, on choisit arbitrairement n=1000)*

Échantillonnage de Gibbs L'échantillonnage de Gibbs ("Gibbs Sampling") est un cas particulier de l'algorithme de Metropolis-Hastings, adapté à LDA, dont l'objectif est de traiter le cas de distributions multi-variées pour lesquelles il est plus simple de calculer $P(x_i \mid x_{-i})$ que $P(x_1, \ldots, x_n)$, où $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots$ $\ldots, x_n)$. Cette approche consiste à estimer tour à tour chaque composante séparément (en fixant toutes les autres), puis à mettre à jour la valeur de cette composante avant de passer à la suivante.

En reprenant le formalisme du Metropolis-Hastings, l'échantillonnage de Gibbs peut être décrit comme suit :

- 1. Initialisation : Choisir $\boldsymbol{x^{(0)}}$ aléatoirement
- 2. Pour chaque itération t (génération de la t-ième observation) :
  - Pour i allant de 1 à n (traitement de la i-ième composante) :
    - Tirer $x_i^{(t)}$ selon $P(x_i \mid \overbrace{x_1^{(t)}, \ldots, x_{i-1}^{(t)}}^{\text{déjà màj}}, \overbrace{x_{i+1}^{(t-1)}, \ldots, x_n^{(t-1)}}^{\text{pas encore màj}})$

Remarque 5. *Ici,* $P(x_i \mid x_{-i}) = \frac{P(x_1,\ldots,x_n)}{P(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)} \propto P(x_1, \ldots, x_n)$ *dans le sens où le dénominateur de dépend pas de* x_i*.*

Remarque 6. *Afin de respecter complètement la description de l'algorithme de Metropolis-Hastings, il serait nécessaire de tirer aléatoirement l'indice* i *à mettre à jour (ce serait la distribution* Q*). Toutefois, tant que les indices sont tous parcourus à chaque itération, cela ne pose pas de problème.*

Échantillonnage de Gibbs marginalisé Pour revenir au cas de l'estimation des paramètres du modèle LDA, les distributions à évaluer par échantillonnage de Gibbs, étant donnés les mots dans les documents $\boldsymbol{w} = \{x_{m,n}\}$, sont les affectations des thèmes aux mots $\boldsymbol{z} = \{z_{m,n}\}$, ainsi que les descriptions des thèmes Φ et leurs contributions aux différents documents Θ. Toutefois, puisque Φ et Θ peuvent être déduites à partir de $\boldsymbol{z}$, une approche plus simple a été proposée[3]. L'échantillonnage

[3]Nous précisons à la fin de la section comment évaluer les distributions de probabilités à partir des affectations des thèmes aux mots.

de Gibbs marginalisé ("Collapsed Gibbs Sampling") consiste à intégrer Φ et Θ dans la distribution jointe (voir équation 4.2) et à évaluer directement $p(\boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

La fonction de mise à jour de la i-ème composante, toutes les autres étant fixées, peut alors s'écrire : $p(z_i = k \mid \boldsymbol{z_{-i}}, \boldsymbol{w})$, où $\boldsymbol{z_{-i}}$ correspond au vecteur z privé de sa i-ème composante.

Or, par définition des probabilités conditionnelles, on a :

$$p(z_i = k \mid \boldsymbol{z_{-i}}, \boldsymbol{w}) = \frac{p(\boldsymbol{w}, \boldsymbol{z})}{p(\boldsymbol{w}, \boldsymbol{z_{-i}})} \tag{4.3}$$

L'évaluation de la fonction de mise à jour revient alors à évaluer les distributions jointes : $p(\boldsymbol{w}, \boldsymbol{z})$ et $p(\boldsymbol{w}, \boldsymbol{z_{-i}})$ qui prennent des formes similaires.

Fonction de mise à jour. Nous détaillons maintenant le calcul de la fonction de mise à jour de l'échantillonnage de Gibbs marginalisé dans le cas de LDA en commençant par l'évaluation de $p(\boldsymbol{w}, \boldsymbol{z})$.

Par définition, on a : $p(\boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{w} \mid \boldsymbol{z})p(\boldsymbol{z})$. De plus, chaque terme de ce produit peut être évalué séparément. En effet, en exprimant tous les paramètres, on a : $p(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$, mais comme $p(\boldsymbol{z})$ ne dépend que de $\boldsymbol{\alpha}$ et $p(\boldsymbol{w} \mid \boldsymbol{z})$ ne dépend que de $\boldsymbol{\beta}$, on peut écrire : $p(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\beta})p(\boldsymbol{z} \mid \boldsymbol{\alpha})$.

D'un côté, $p(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\beta})$ peut être évalué à partir de sa définition de loi multinomiale sur les mots observés étant donnés leurs thèmes associés :

$$p(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\beta}, \Phi) = \prod_{w_i \in \mathcal{W}} p(w = w_i \mid z = z_i)$$

Or, en regroupant toutes les occurrences du même terme qui sont associées au même thème, on obtient :

$$p(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\beta}, \Phi) = \prod_{k=1}^{K} \prod_{\{i, z_i = k\}} p(w = w_i \mid z = z_i)$$

Ce qui peut se réécrire :

$$p(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\beta}, \Phi) = \prod_{k=1}^{K} \prod_{t \in \mathcal{V}} p(w = t \mid z = k)^{n_k^{(t)}}$$

où $n_k^{(t)}$ correspond au nombre de fois où le thème k est associé au terme t.

En intégrant Φ, on obtient la probabilité souhaitée $p(\boldsymbol{w} \mid \boldsymbol{z})$:

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\beta}) &= \int p(\boldsymbol{w} \mid \boldsymbol{z}, \Phi) p(\Phi \mid \boldsymbol{\beta}) d\Phi \\ &= \int \prod_{k=1}^{K} \frac{1}{\Delta(\boldsymbol{\beta})} \prod_{t \in \mathcal{V}} p(w = t \mid z = k)^{n_k^{(t)} + \beta_t - 1} d\boldsymbol{\phi_k} \\ &= \prod_{k=1}^{K} \frac{\Delta(\boldsymbol{n_k} + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \end{aligned} \tag{4.4}$$

où $\boldsymbol{n_z} = \{n_z^{(t)}\}_{t=1}^{\mathcal{V}}$ et $\Delta(\boldsymbol{x}) = \frac{\prod \Gamma(x_i)}{\Gamma(\sum x_i)}$ avec Γ la généralisation aux nombres réels de la fonction factorielle.

De l'autre côté $p(\boldsymbol{z} \mid \boldsymbol{\alpha})$ peut être évalué à partir de sa définition de loi multinomiale sur les thèmes à partir des documents observés :

$$p(\boldsymbol{z} \mid \boldsymbol{\alpha}, \Theta) = \prod_{i=1}^{|W|} p(z_i \mid d_i)$$

où d_i correspond au document auquel appartient le mot i.

Cette égalité peut se réécrire :

$$p(\boldsymbol{z} \mid \boldsymbol{\alpha}, \Theta) = \prod_{m=1}^{D} \prod_{\{i, z_i = k\}} p(z_i = k \mid d_i = m)$$

Et finalement :

$$p(\boldsymbol{z} \mid \boldsymbol{\alpha}, \Theta) = \prod_{m=1}^{D} \prod_{k=1}^{K} p(z = k \mid d = m)^{n_m^{(k)}}$$

où $n_m^{(k)}$ correspond au nombre de fois où le document m est associé au thème k.

De la même manière que l'on évalue $p(\boldsymbol{w} \mid \boldsymbol{z}, \Phi)$, on peut ici intégrer Θ :

$$\begin{aligned} p(\boldsymbol{z} \mid \boldsymbol{\alpha}) &= \int p(\boldsymbol{z} \mid \Theta) p(\Theta \mid \boldsymbol{\alpha}) d\Theta \\ &= \int \prod_{m=1}^{D} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{z=1}^{\mathcal{K}} p(z = k \mid d = m)^{n_m^{(k)} + \alpha_k - 1} d\boldsymbol{\theta_m} \\ &= \prod_{m=1}^{D} \frac{\Delta(\boldsymbol{n_m} + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \end{aligned} \tag{4.5}$$

où $\boldsymbol{n_m} = \{n_m^{(k)}\}_{k=1}^{K}$.

Enfin, puisque :

$$p(\boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{w} \mid \boldsymbol{z})p(\boldsymbol{z})$$

On obtient le numérateur de la fonction de mise à jour (voir équation 4.3) :

$$p(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{z=1}^{K} \frac{\Delta(\boldsymbol{n_z} + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \times \prod_{m=1}^{D} \frac{\Delta(\boldsymbol{n_m} + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}$$

Par ailleurs, le dénominateur peut être réécrit :

$$p(\boldsymbol{w}, \boldsymbol{z_{-i}}) = p(\boldsymbol{w_{-i}} \mid \boldsymbol{z_{-i}})p(\boldsymbol{w_i})p(\boldsymbol{z_{-i}})$$

et que $p(\boldsymbol{w_i})$ est une constante calculée à partir des données, on a :

$$p(\boldsymbol{z_i} = k \mid \boldsymbol{z_{-i}}, \boldsymbol{w}) = \frac{p(\boldsymbol{w}, \boldsymbol{z})}{p(\boldsymbol{w}, \boldsymbol{z_{-i}})} \propto \frac{p(\boldsymbol{w} \mid \boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{w_{-i}} \mid \boldsymbol{z_{-i}})p(\boldsymbol{z_{-i}})}$$

Puisque ces distributions s'écrivent sous forme de produits, et que les termes sont les mêmes à l'exception de la composante i, tous les termes s'annulent sauf ceux du numérateur où apparaît i. On obtient alors :

$$p(\boldsymbol{z_i} = k \mid \boldsymbol{z_{-i}}, \boldsymbol{w}) \propto \frac{\Delta(\boldsymbol{n_z} + \boldsymbol{\beta})}{\Delta(\boldsymbol{n_{z,-i}} + \boldsymbol{\beta})} \times \frac{\Delta(\boldsymbol{n_m} + \boldsymbol{\alpha})}{\Delta(\boldsymbol{n_{m,-i}} + \boldsymbol{\alpha})}$$

Encore une fois, comme tous les termes sont les mêmes, à l'exception de la composante i, par définition de la fonction gamma, tous les produits s'annulent sauf pour i, ce qui permet de donner la formule de mise à jour :

$$p(\boldsymbol{z_i} = k \mid \boldsymbol{z_{-i}}, \boldsymbol{w}) \propto \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{\mathcal{V}} (n_{k,-i}^{(t)} + \beta_t)} \times \frac{n_{m,-i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_{m,-i}^{(k)} + \alpha_k) - 1}$$

Après la période de burn-in et une fois que la chaîne a convergé, l'échantillonnage de Gibbs permet alors d'estimer les affectations des thèmes aux mots des documents : $p(\boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

Paramètres des lois multinomiales Φ et Θ. A partir de ses affectations $\boldsymbol{z}$, (et étant donnés les mots dans les documents $\boldsymbol{w}$), il est alors possible de déduire les paramètres des distributions multinomiales des documents sur les thèmes et des thèmes sur les termes. En effet, en appliquant le théorème de Bayes, on a :

$$p(\boldsymbol{\theta_m} \mid \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\alpha}) = \frac{p(\boldsymbol{z} \mid \boldsymbol{\theta_m})p(\boldsymbol{\theta_m} \mid \boldsymbol{\alpha})}{p(\boldsymbol{z})}$$

Par conjugaison des lois de Dirichlet et Catégorielle, on a alors simplement :

$$p(\boldsymbol{\theta_m} \mid \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\alpha}) = Dir(\boldsymbol{\theta_m} \mid \boldsymbol{n_m} + \boldsymbol{\alpha})$$

Et de manière analogue :

$$\begin{aligned} p(\boldsymbol{\phi_k} \mid \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\beta}) &= \frac{p(\boldsymbol{w} \mid \boldsymbol{\phi_k}) p(\boldsymbol{\phi_k} \mid \boldsymbol{\beta})}{p(\boldsymbol{w})} \\ &= Dir(\boldsymbol{\phi_k} \mid \boldsymbol{n_z} + \boldsymbol{\beta}) \end{aligned}$$

En utilisant l'espérance de la distribution de Dirichlet qui est connue ($E[x_i] = \frac{x_i}{\sum_i x_i}$) on obtient alors chaque composante de chaque distribution :

$$\phi_k^t = p(w = t \mid z = k) = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{\mathcal{V}} (n_k^{(t)} + \beta_t)}$$

et :

$$\theta_m^k = p(z = k \mid d = m) = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_m^{(k)} + \alpha_k)}$$

Remarque 7. *Ces deux expression montrent une interprétation intuitive des hyper-paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$, qui est due au fait que les distributions de Dirichlet et de Bernoulli multivariée sont conjuguées. En effet, dans ces expressions, les termes de prieurs β_t et α_k jouent le même rôle que les observations $n_k^{(t)}$ et $n_m^{(k)}$; le décompte des termes ne fait pas la distinction entre les données réellement observées et la croyance a priori. Ainsi, l'expression de la distribution a posteriori est similaire à celle de la distribution a priori pour laquelle les paramètres auraient été mis à jour par les observations. Cette observation explique la dénomination "pseudo-observations" ("pseudo-count") souvent utilisée pour désigner les hyper-paramètres. Plus la valeur des hyper-paramètres est élevée, plus il faut d'observations pour modifier les connaissances apportées a priori, ce qui justifie l'impact des hyper-paramètres sur les distributions décrit dans l'introduction de ce chapitre.*

2.3 Expérimentations sur le corpus complet

Dans le but d'explorer les possibilités offertes par le modèle LDA, nous avons conduit une expérimentation consistant à modéliser l'ensemble du corpus à l'aide de thèmes. Cette utilisation statique du modèle LDA a pour but d'évaluer l'interprétabilité des thèmes construits, et de déterminer si le modèle LDA peut permettre à un analyste d'extraire des connaissances supplémentaires.

Pour cette expérimentation, nous avons utilisé la librairie Java Mallet[4] et en particulier sa classe ParallelTopicModel qui offre une grande flexibilité. Nous avons

[4]http ://mallet.cs.umass.edu/

fixé le nombre de thèmes à 25 afin de trouver un équilibre en richesse d'analyse et quantité de résultats, et suite à quelques essais, nous avons fixé les valeurs des priors symétriques α et β à 0,004 et 0,1 respectivement. (Nous détaillons le choix de $\boldsymbol{\alpha}$ dans la section 2.3.2). Pour l'estimation des paramètres par échantillonnage de Gibbs, nous avons fixé le nombre d'itérations à 2000, en ignorant les 200 premières valeurs (burn-in) et en sélectionnant ensuite un tirage sur 10 de façon à limiter l'impact de la corrélation. Une description des thèmes construits par ce modèle est présentée dans la section 2.3.1.

Dans un second temps, nous avons fait varier la valeur de α afin de mettre en avant son influence sur les distributions des documents. De plus, en comparant les vraisemblances des modèles selon différentes valeurs de α nous avons également mis en avant l'impact du contexte de forte nouveauté sur les thèmes construits par LDA. En effet, les modèles les plus vraisemblables sont ceux proposant une description très saillante des documents, ce qui signifie que les documents ne traitent que d'un thème, qui n'a que peu de liens avec les autres.

2.3.1 Analyse des résultats

Une description synthétique des thèmes construits est présentée dans le tableau 4.1 qui contient les 7 termes les plus représentatifs de chaque thème ainsi que leurs poids définis comme la somme cumulée des probabilités des documents ($\forall i \in \{1, \ldots, K\}, score(z_i) = \sum_{d \in \mathcal{D}} p(z_i \mid d)$). Cette mesure de l'importance relative des thèmes au sein d'un corpus exploite la propriété de clustering flou de LDA, qui ne se contente pas d'attribuer un document à son thème le plus représentatif, mais propose une distribution de probabilités sur l'espace des thèmes. Chaque document contribue, à hauteur de sa probabilité, aux poids des différents thèmes et comme la somme des probabilités vaut 1 ($\forall d \in \mathcal{D}, \sum_{i=1}^{K} p(z_i \mid d) = 1$), chaque document contribue au total à hauteur de 1. De fait, le poids des thèmes peut également être lu en termes de nombre de documents représentés et la somme des poids des thèmes correspond au nombre total de documents.

L'analyse des thèmes construits par LDA sur le corpus complet fait apparaître des concepts qui étaient passés, jusqu'à présent, inaperçus aux yeux de l'analyste. En effet, puisque LDA construit des thèmes plus complexes que la simple co-occurrence des termes dans les documents, et analyse les co-occurrences d'ordres supérieurs (on dit que t_1 et t_3 sont en co-occurrence d'ordre 2 s'il existe t_2 tel que t_1 apparaît souvent avec t_2 et que t_2 apparaît souvent avec t_3, etc...), ce modèle semble capable d'analyser un corpus de données parcimonieuses. Le thème 15 représente, par exemple, une thématique autour des mails frauduleux et du "phishing" (attaque informatique visant à obtenir des informations personnelles en se faisant passer pour un tiers de confiance) qui n'avait pas été détectée par les approches

précédentes. EDF, de par son nombre de clients important et la fréquence de ses contacts avec eux, est fréquemment utilisé comme "tiers de confiance" par les attaquants, ainsi l'analyse des messages ayant une probabilité élevée pour ce thème pourrait permettre d'améliorer la sécurité des utilisateurs de Twitter et de l'ensemble des clients en général. Le thème 22, quant à lui, représente un nombre important de documents autour de la thématique du compteur électrique et de la coupure de courant, qui n'avait pas semblé si conséquente jusque là. L'étude des documents en rapport avec ce thème montre de grandes disparités dans les cas décrits et illustre la capacité d'abstraction de LDA.

#	**25 Thèmes LDA pour le corpus complet**	**Score**
0	flamanville energie nucleaire enel central mettre solaire	2438
1	payer facture electricite lumiere bayrou onpc recevoir	3204
2	facture eolien payer parc nouveau mettre service	2565
3	electrique voiture lance lancer vehicule location fondation	2234
4	milliard euro fessenheim fermeture etat rembourser indemnite	3850
5	nucleaire areva vendre hydrolienne ouvrir mettre capital	1902
6	facture courant privatisation fermer bois manman site	1703
7	village acces like potable mali payer energie	1301
8	heure payer electricite energie ingenieur maintenance technicien	1840
9	nucleaire chine enquete projet central reacteur entrer	4333
10	lumiere mere eteindre facture peur compte crier	2145
11	central emploi france entreprise levallois creer electricite	3575
12	courant privatiser proglio coran electrocuter niquer mecque	1882
13	total anglais ecolo suer humilier saigner filial	2023
14	courant coupure electricite etre heure suivre couper	4968
15	mail faux phishing arnaque attention frauduleux recevoir	2632
16	minute flipper gens couper courant pendre travailler	4201
17	payer tarif salarier sncf electricite entreprise agent	4951
18	proglio pepy tete guillaume sncf remplacer pressentir	1612
19	proglio veolia henri conseil etat administration patron	2450
20	electricite coupe gens falloir flipper france monde	2845
21	proglio patron nouveau centrale directeur cahuzac nucleaire	1687
22	courant facture electricite compteur etre couper heure	8500
23	radioactif amende fuite condamner electricite euro nucleaire	1937
24	electricite priver foyer bayrou onpc lareunion dumile	1925

TABLE 4.1 – Tableau présentant les 25 thèmes construits sur le corpus complet ainsi que leurs scores (poids cumulés des documents).

Cette description peut toutefois être complétée en faisant apparaître les probabilités des termes au sein de chaque thème, par exemple à l'aide d'une visualisation sous forme de nuages de mots (voir figure 4.4). En mettant en correspondance, ici avec l'outil en ligne Wordle[5], les probabilités des termes au sein des thèmes avec la

[5]http ://www.wordle.net/

FIGURE 4.4 – Nuages de mots illustrant 6 topics construits avec le modèle LDA.

taille de la police de caractères, cette représentation permet, non seulement, d'interpréter facilement de multiples distributions de probabilités, mais également de distinguer plusieurs types de thèmes (comme, par exemple, ceux décrits pas quelques termes saillant et ceux décrits par plusieurs termes équiprobables).

Afin d'obtenir une vue globale du modèle construit, de ses thèmes et de leurs relations avec les documents et les termes qu'ils contiennent, nous proposons une visualisation sous forme de graphe inspirée de celle proposée par la communauté des humanités digitales de l'université de Stanford[6]. A l'aide de l'outil de visualisation et de manipulation Gephi [BHJ09], nous avons construit un graphe dans lequel trois types de nœuds sont représentés (voir figure 4.5) : les thèmes en rouge, les tweets en vert et les termes en bleu. La taille des nœuds décrivant les thèmes et les termes est constante alors que la taille des nœuds décrivant les documents est proportionnelle à leur nombre de retweets. Nous avons cherché à pondérer la taille des thèmes par leurs poids, mais les différences d'ordre de grandeur entre les poids des différents types de nœuds ne permettent pas de visualiser de variation significative. Par ailleurs, les tweets n'ayant jamais été retweetés ne sont pas visualisés afin d'alléger le graphe. Les thèmes correspondent à un élément central du graphe, auxquels les deux autres types de nœuds sont rattachés en fonction des distributions de probabilités calculées par le modèle. Une arête est créée entre un thème z et un document d si et seulement si $p(z \mid d) > 0.1$, c'est-à-dire si la description du document par ce thème est significative et le poids de l'arête est défini comme cette probabilité. Une arête est créée entre un thème et ses 15 termes les plus représentatifs avec un poids correspondant à la probabilité $p(t \mid z)$. L'algorithme de placement de type "force et ressort", Force Atlas 2[7], est appliqué afin de visualiser ce graphe.

Il ressort de cette visualisation que chaque thème semble attirer une grappe de documents pour lesquels il apparaît comme très spécifique. Cela tend à montrer que les tweets, courts, ne traitent que d'un sujet en particulier. Les thèmes semblent

[6] https ://dhs.stanford.edu/comprehending-the-digital-humanities/
[7] www.medialab.sciences-po.fr/publications/Jacomy_Heymann_Venturini-Force_Atlas2.pdf

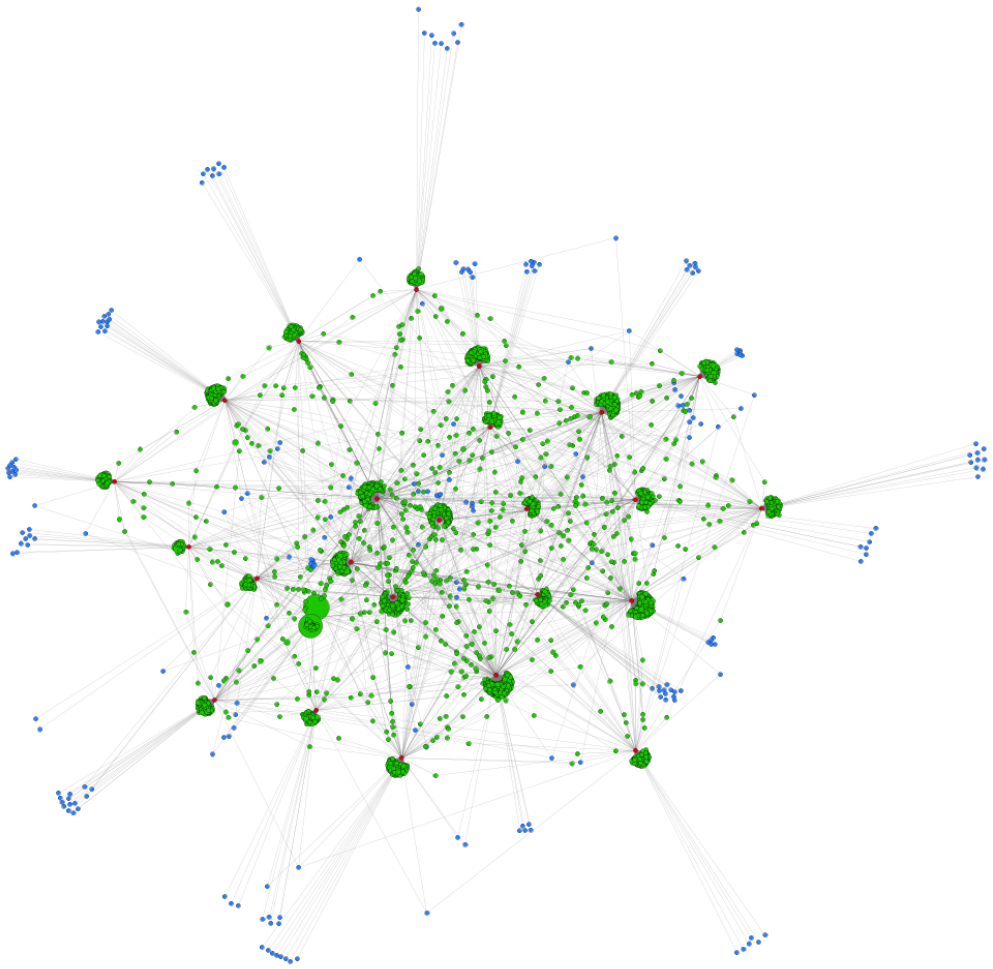


FIGURE 4.5 – Visualisation sous forme de graphe du modèle LDA construit sur le corpus complet. Les thèmes sont représentés en rouge, les termes en bleu et les documents en vert. Une arête relie chaque thème à ses 15 termes les plus représentatifs et une arête relie un document à un thème si sa probabilité est supérieure à 0,1.

également être décrits par un vocabulaire spécifique et de nombreux termes ne sont reliés qu'à un seul thème (voir figure 4.6). Toutefois, même si cette approche permet de donner une vue synthétique d'un modèle complexe, la distance sémantique introduite par le graphe n'est pas bien définie. La proximité entre thèmes ou entre documents résulte d'une simplification du problème à un espace à deux dimensions dans lequel les documents et les termes reliés à plusieurs thèmes ne sont pas mis en valeur. En effet, le positionnement de nœuds au cœur du graphe peut être dû à des relations complexes avec les nœuds avoisinants.

Puisque les documents semblent en majorité assignés à un thème en particulier, et que la visualisation sous forme de graphe ne permet pas de quantifier cette observation, nous proposons également de visualiser les probabilités des documents sur les thèmes $p(z_i \mid d)$. Le nombre de documents étant trop important pour visualiser chaque distribution séparément, nous proposons de discrétiser l'intervalle $[0; 1]$ et

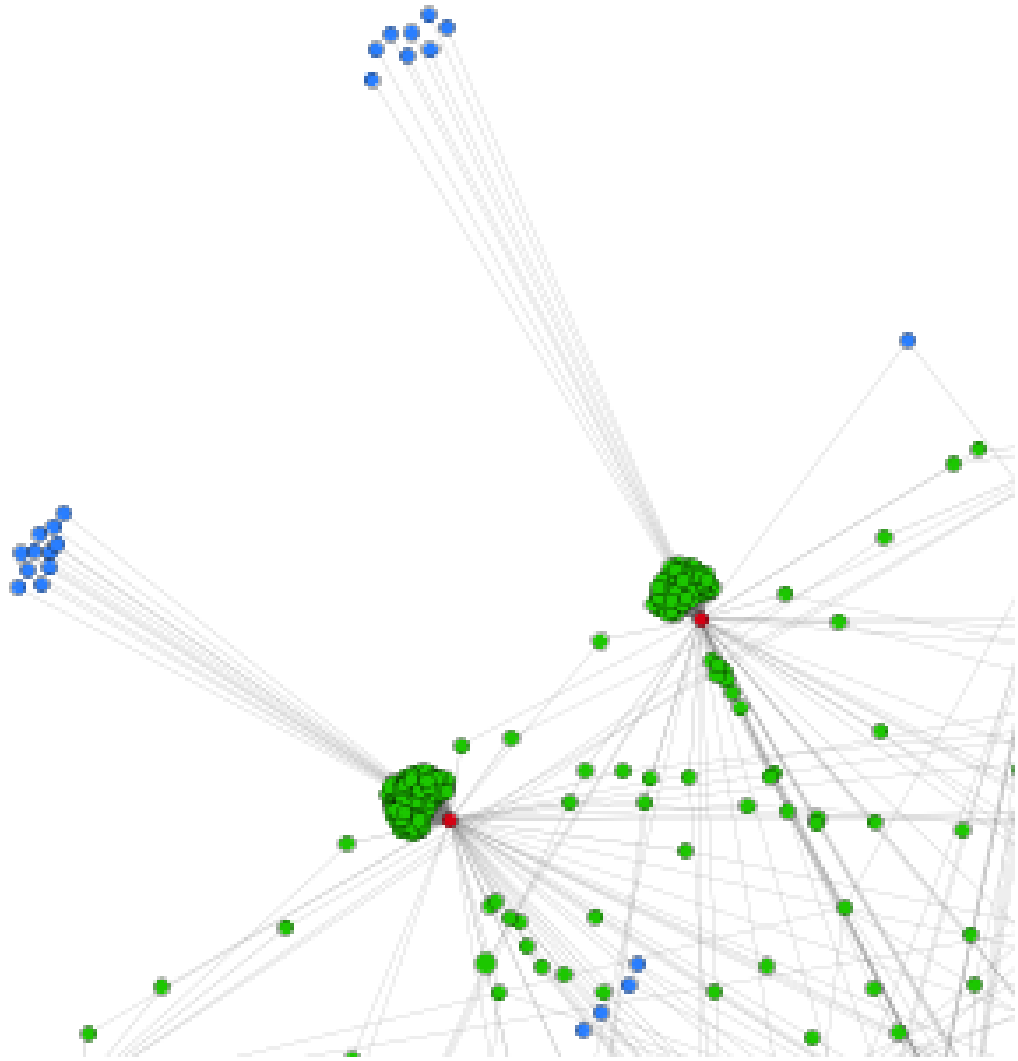


FIGURE 4.6 – Zoom sur deux noœuds "thèmes" attirant des documents et des termes spécifiques.

de visualiser pour chaque thème le nombre de documents dont la probabilité appartient à chaque sous-intervalle. La figure 4.7 présente un histogramme décrivant, pour chacun des 25 thèmes, la répartition des probabilités des documents dans chacun des intervalles $[0,1;0,2[$, $[0,2;0,3[$, …, $[0,9;1]$. Dans cette visualisation, le premier sous intervalle $[0;0,1[$ a été masqué afin d'améliorer la lisibilité. En effet, puisque, dans leur ensemble, les thèmes visent à fournir une description discriminante des documents, ils ne représentent qu'un "petit" nombre de documents. Les autres, largement majoritaires, ne sont pas associés à ce thème et se voient donc attribuer une probabilité faible pour ce thème. En conclusion, parmi tous les couples *document*×*thèmes* possibles, la grande majorité n'est pas en correspondance, et la probabilité $p(z \mid d)$ appartient à l'intervalle $[0;0,1[$. Une fois cet intervalle retiré, il apparaît que, dans leur ensemble, les documents sont associés à un thème unique, avec une probabilité $\geq 0,9$. (Et donc avec une probabilité < 0.1 pour les 24 autres thèmes).

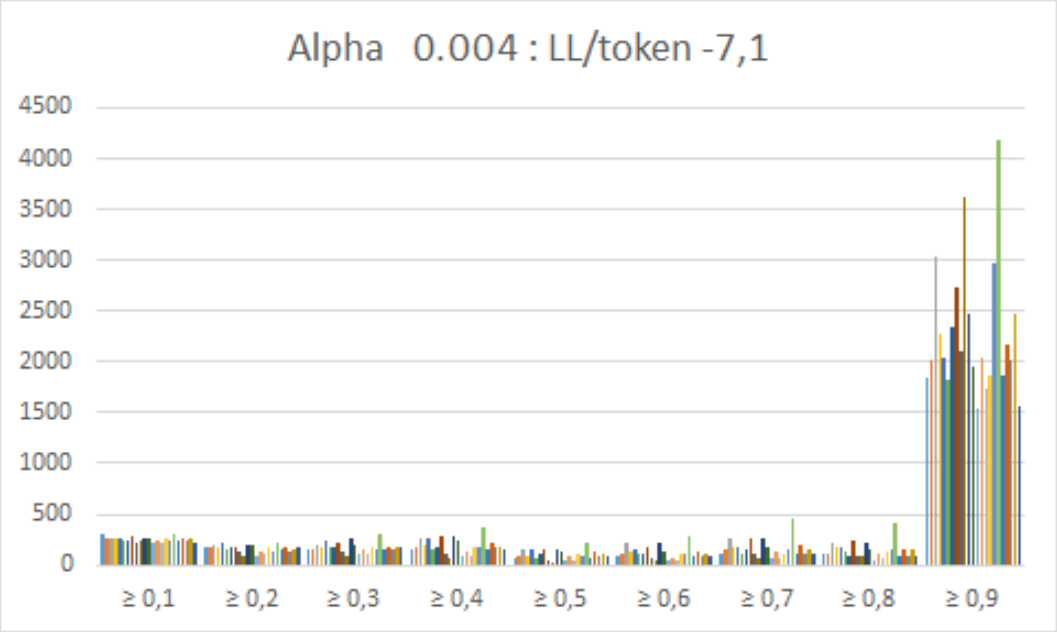


FIGURE 4.7 – Répartition des probabilités des tweets pour chaque thème. (Les documents dont la probabilité est inférieure à 0.1, largement majoritaires, sont ignorés afin d'améliorer la lisibilité.)

Toutefois, comme nous l'avons vu dans la section 2.2, la répartition des probabilités des tweets sur chaque thème est influencée par la valeur de l'hyper-paramètre $\boldsymbol{\alpha}$. Une valeur faible de α (ici, $\alpha = 0,004$) induit une description parcimonieuse des documents. Afin de justifier le choix de cette valeur pour l'analyse de notre corpus, nous proposons d'étudier, en parallèle, l'influence de α sur les distributions des documents et la qualité des modèles construits en termes de vraisemblance.

2.3.2 Influence de l'hyper-paramètre α sur les distributions des documents

Afin de comparer les distributions de probabilités des documents sur les thèmes, nous avons construit plusieurs instances du modèle LDA pour différentes valeurs de α. En reprenant la méthodologie utilisée pour la figure 4.7, nous présentons, pour chacun de ces modèle, la répartition des probabilités des documents dans chacun des intervalles $[0,1;0,2[$, $[0,2;0,3[$, …, $[0,9;1]$. Les résultats sont reportés sur la figure 4.8.

Cette analyse met en avant le glissement qui s'opère vers des valeurs de probabilités faibles lorsque $\boldsymbol{\alpha}$ augmente. Pour une valeur faible de α, les documents sont assignés à un thème majoritaire ($P(z \mid d) \geq 0.9$). Pour une valeur plus élevée, les "pseudo-observations" *a priori* sont délicates à contrebalancer et les documents sont répartis équitablement entre chaque thème ($P(z \mid d) \in [0.1; 0.2]$)

Puisque ce glissement influence grandement la représentation des documents et les thèmes construits, il est nécessaire d'évaluer la qualité des modèles pour différentes valeurs de α. Une mesure classique pour évaluer les modèles génératifs est la

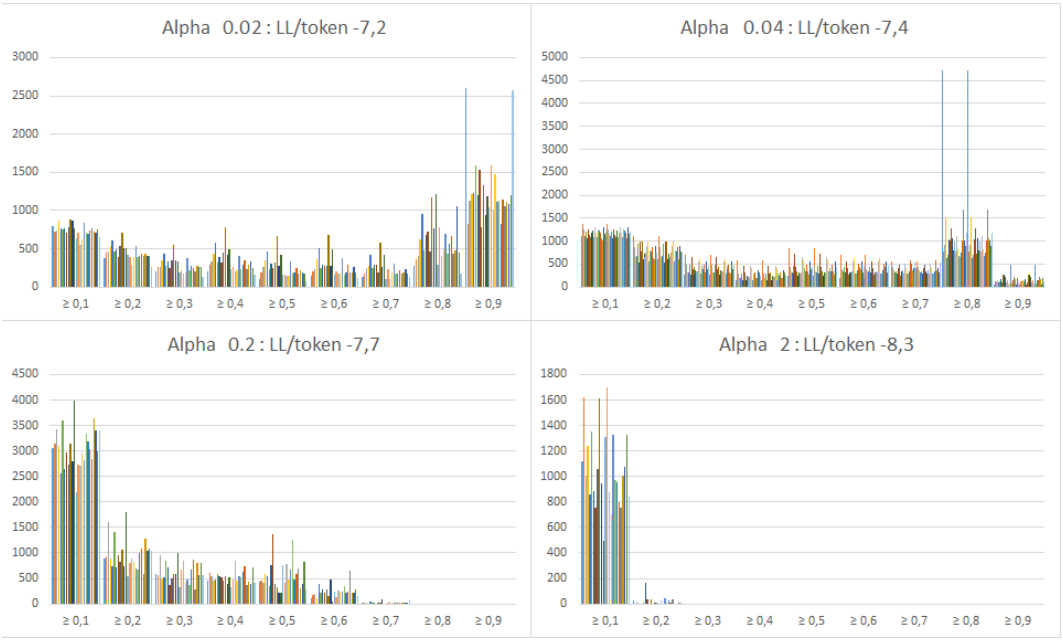


FIGURE 4.8 – Histogrammes illustrant l'influence de la valeur de α sur les distributions de probabilités des documents.

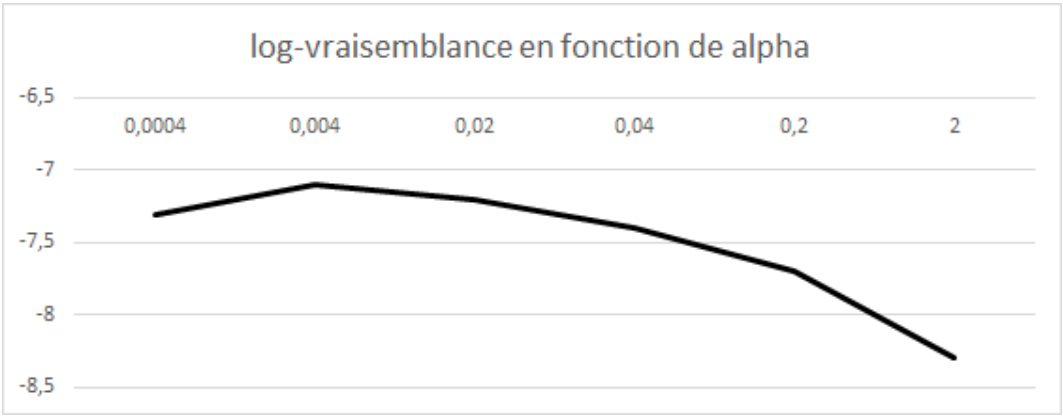


FIGURE 4.9 – Courbe représentant la log-vraisemblance des modèles pour différentes valeurs de α. Le maximum est atteint pour $\alpha = 0,004$.

vraisemblance d'un jeu de paramètres $\boldsymbol{\theta}$ étant donné un ensemble d'observations $\boldsymbol{x}$. Elle est définie par : $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\theta})$ et une vraisemblance élevée correspond à un meilleur modèle. Puisque la vraisemblance introduit souvent de grands nombres négatifs, une variante classique consiste à prendre le logarithme de cette valeur, que l'on appelle log-vraisemblance. De plus, comme cette mesure dépend de la longueur des documents, elle est fréquemment normalisée pour donner la log-vraisemblance par mot. Les résultats du calcul, pour chacun des modèles, sont donnés sur la figure 4.9. Ils permettent d'identifier que la vraisemblance atteint son maximum pour une valeur faible de α (0,004), ce qui correspond à une description parcimonieuse des documents.

2.3.3 Premières conclusions pour l'application de LDA.

Les premiers résultats obtenus avec LDA sur notre corpus sont encourageants et semblent montrer la pertinence de l'utilisation des modèles de thèmes pour la modélisation de thématiques dans les données Twitter. Toutefois, une des limitations du modèle LDA est qu'il suppose que les documents sont interchangeables au sein du corpus. Cette propriété limite l'utilisation de LDA à l'extraction de thématiques représentatives d'un corpus statique -par exemple pour produire une vue synthétique- puisque les dates de publication des documents ne sont pas prises en compte par le modèle, celui-ci n'est pas adapté à la modélisation d'un flux dynamique de documents. En effet, dans le cadre de Twitter, l'hypothèse qui voudrait que les thématiques abordées dans les nouveaux documents soient les mêmes que celles apprises sur des documents plus anciens n'est pas réaliste. Pour pallier ce problème, une solution consiste à diviser le corpus en sous-périodes et à entrainer un modèle par période. Ainsi, les documents ne sont plus supposés interchangeables au sein de l'ensemble du corpus, mais seulement au sein d'une sous-période. C'est dans ce cadre que nous proposons, dans la section 4.3, l'approche Dynamic-LDA qui permet de suivre l'évolution temporelle de thématiques dans un contexte de forte nouveauté.

3 Dynamic-LDA

Puisque les modèles de thèmes ne sont pas adaptés à la modélisation de flux de textes, notre méthodologie propose de découper le corpus en sous-périodes afin de construire un modèle par période. Dans un second temps, nous comparons les thématiques des modèles successifs afin de détecter la continuité temporelle. Les thématiques consécutives proches sémantiquement sont reliées entre elles dans une relation de type plusieurs-à-plusieurs (n : n) de manière à visualiser leur évolution au cours du temps. Contrairement à une recherche de correspondance un-pour-un entre les thèmes de modèles successifs, cette approche permet de détecter l'apparition et la disparition de thématiques, mais également des motifs de fusion ou de division de thématiques lorsque deux (ou plusieurs) sujets se regroupent ou à l'inverse lorsque que deux (ou plusieurs) sous-thématiques émergent à partir d'une thématique existante. De plus, comme cette approche ne force pas une correspondance un-pour-un, elle permet également d'étudier le cas où les modèles successifs n'ont pas le même nombre de thèmes.

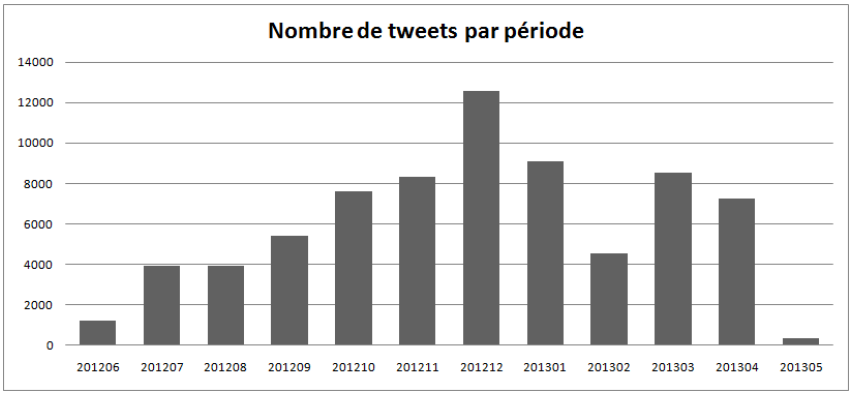


FIGURE 4.10 – Histogramme présentant le nombre de tweets publiés chaque mois.

3.1 Découpage Temporel

Afin d'obtenir une évolution fine des thèmes détectés, il est nécessaire de découper le corpus en périodes courtes. Toutefois, la phase d'estimation des paramètres des modèles d'apprentissage nécessite un nombre suffisant de données afin d'établir des corrélations significatives entre termes. Ainsi, un compromis doit être trouvé entre la finesse d'analyse et la représentativité des modèles. Étant donné le nombre moyen de tweets publiés pendant notre période d'étude, nous proposons de découper le corpus en périodes d'un mois (voir figure 4.10). Le premier et le dernier mois, qui ne contiennent que quelques jours de données, sont ignorés. Pour chaque période $\mathcal{T}_i$ (et donc chaque ensemble de documents $\mathcal{D}_i$) un modèle $\mathcal{M}_i = (\mathcal{D}_i, \mathcal{V}_i, K, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$ est entraîné et l'ensemble des distributions de probabilités $\boldsymbol{\phi}_i$ et $\boldsymbol{\theta}_i$ est estimé par l'échantillonnage de Gibbs présenté dans la section2.2.1.

3.2 Harmonisation des vocabulaires

Dans le modèle LDA, un thème k est décrit par une distribution de probabilités sur les termes du vocabulaire $\mathcal{V}_i$ et est représenté par un vecteur : $\boldsymbol{\phi_{ik}} = (\phi_{ik}(1), \ldots$ $\ldots, \phi_{ik}(|\mathcal{V}_i|))$. Or, notre approche propose de construire des modèles LDA sur des périodes différentes et puisque les tweets sont associés à la période correspondant à leur date de publication, chaque modèle est entraîné sur un ensemble de tweets différent. En conséquence, les vocabulaires des différents modèles sont potentiellement différents et leurs thèmes sont représentés par un vocabulaire différent. En effet, chaque nouvelle thématique introduit de nouveaux termes et lorsqu'une thématique n'est plus abordée, ses termes spécifiques ne sont plus employés. Ainsi, il n'est pas réaliste de supposer que le vocabulaire est stable au cours du temps. Il est alors nécessaire de mettre en place une étape d'harmonisation pour comparer des vecteurs définis sur des espaces différents.

Par ailleurs, la taille des vocabulaires est souvent très importante et les thèmes sont décrits par des vecteurs de grandes dimensions. Toutefois, toutes les compo-

santes de ces vecteurs n'ont pas la même importance. En effet, les termes auxquels sont assignées de faibles probabilités sont peu représentatifs du thème étudié et sont fréquemment ignorés dans la description du thème. Ainsi, les thèmes sont souvent présentés sous forme de listes pondérées d'un petit nombre de termes les plus représentatifs (15 dans l'article original [BNJ03]), c'est-à-dire les termes dont la probabilité est maximale présentés dans l'ordre décroissant. Dans le cadre du calcul de divergences entre thèmes, il est possible d'utiliser cette propriété afin de diminuer le temps de calcul tout en limitant la perte d'information.

En prenant en compte ces deux spécificités, nous proposons de ne retenir que les termes les plus représentatifs de chaque thème du modèle $\mathcal{M}_i$ afin de construire un sous-vocabulaire spécifique $\mathcal{V}_i^*$ avec : $\mathcal{V}_i^* \subset \mathcal{V}_i$.

Définition 5. Étant donné un modèle $\mathcal{M}_i$, le *vocabulaire r-représentatif* $\mathcal{V}_i^*$ de $\mathcal{M}_i$ est défini par : $\mathcal{V}_i^* = \{t \in \mathcal{V}_i \mid \exists k \in \{1, \dots K\}, \phi_{ik}(t) \in max_r(\boldsymbol{\phi_k})\}$, où $max_r(\boldsymbol{x})$ désigne les r-plus grands éléments de $\boldsymbol{x}$.

De plus, en vue de comparer les thèmes de deux modèles successifs, il est nécessaire de construire un vocabulaire commun aux thèmes des deux modèles.

Définition 6. Étant donnés deux modèles $\mathcal{M}_i$ et $\mathcal{M}_{i+1}$, le *vocabulaire commun* à $\mathcal{M}_i$ et $\mathcal{M}_{i+1}$ est défini par : $\mathcal{U}_{i,i+1} = \mathcal{V}_i^* \cup \mathcal{V}_{i+1}^*$.

Les thèmes de ces deux modèles sont alors décrits par des vecteurs de dimension $|\mathcal{U}_{i,i+1}|$, tels que $\forall t \in \mathcal{U}_{i,i+1}$:

$$\phi_{ik}^*(t) = \begin{cases} \phi_{ik}(t) & \text{si } t \in \mathcal{V}_i^* \\ 0 & \text{sinon} \end{cases} \qquad \phi_{i+1k}^*(t) = \begin{cases} \phi_{i+1k}(t) & \text{si } t \in \mathcal{V}_{i+1}^* \\ 0 & \text{sinon} \end{cases}$$

Finalement, de manière à continuer à représenter les thèmes sous forme de distributions de probabilités, un lissage de Laplace est appliqué aux vecteurs en ajoutant un "petit" nombre réel positif à chaque composante (ici α_i) et en divisant le résultat par la somme des composantes.

3.3 Détection des relations entre thèmes

Étant donnée une suite de modèles $(\mathcal{M}_i)_{i=1}^n$ définissant les thèmes de périodes successives, l'étape suivante consiste à modéliser l'évolution de ces thèmes au cours du temps. Est-ce que certains thèmes se maintiennent d'une période à l'autre ? Est-ce qu'ils évoluent ou sont remplacés ? Pour cela, nous proposons de calculer la divergence entre chaque paire de thèmes de deux modèles consécutifs et nous définissons la relation binaire entre thèmes : "évolue en".

Définition 7. Étant donnés deux ensembles de thèmes $\boldsymbol{\phi_i}$ et $\boldsymbol{\phi_{i+1}}$ de modèles successifs $\mathcal{M}_i$ et $\mathcal{M}_{i+1}$, une mesure de divergence δ entre distributions de probabilités et un seuil ρ, la relation *évolue en*, notée $\triangleright_{\delta\rho}$, est définie par $\phi_{ik} \triangleright_{\delta\rho} \phi_{i+1k}$ ssi $\delta(\phi_{ik}, \phi_{i+1k}) < \rho$.

Par construction, la relation *évolue en* est une relation de type plusieurs-à-plusieurs, un thème peut évoluer en plusieurs thèmes et plusieurs thèmes peuvent évoluer en un thème.

4 Choix d'une mesure de divergence entre thèmes

De nombreuses mesures de divergences ont été proposées dans la littérature (voir par exemple la synthèse de Lesot et al. [LRB09]) et le choix de la mesure utilisée est ici crucial car il influence grandement les résultats obtenus et l'interprétation visuelle des liens créés. Afin d'éclairer ce choix, nous proposons de comparer expérimentalement les distributions empiriques de sept mesures classiques adaptées à nos données : Cosinus, Euclidienne, Jaccard généralisée, Kullback-Leibler, Jensen-Shannon, Bhattacharyya et Hellinger. Pour cela, nous avons calculé, avec chacune des divergences, les similarités entre chaque paire de thèmes pris entre deux modèles consécutifs de notre corpus. Le corpus est divisé en 10 périodes d'un mois, et pour chacune, nous avons entraîné un modèle LDA avec 10 thèmes. Nous comparons donc 9 paires de modèles, et pour chacune nous calculons 10×10 similarités, pour un total de 900 mesures de divergence. Nous présentons dans la suite la répartition de ces 900 valeurs pour chacune des divergences, dont nous commençons par rappeler la définition.

4.1 Définitions et distributions empiriques

4.1.1 Divergence cosinus

Définition 8. La *divergence cosinus* de deux distributions de probabilités ϕ_1 et ϕ_2 est définie comme le complément à 1 de la similarité cosinus :

$$cosineDiv(\phi_1, \phi_2) = 1 - cosineSim(\phi_1, \phi_2)$$

où :

$$cosineSim(\phi_1, \phi_2) = \frac{\sum \phi_{1i}.\phi_{2i}}{\sqrt{\sum \phi_{1i}^2}.\sqrt{\sum \phi_{2i}^2}}$$

Puisque pour toute distribution de probabilités ϕ, $0 \leq \phi_i \leq 1$, on a : $0 \leq cosineDiv(\phi_1, \phi_2) \leq 1$. De plus, $cosineDiv(\phi_1, \phi_2) = 0$ si ϕ_1 et ϕ_2 sont similaires et $cosineDiv(\phi_1, \phi_2) = 1$ si elles divergent. La figure 4.11 présente la répartition

des 900 mesures de divergences calculées selon la divergence cosinus pour 10 modèles LDA consécutifs composés de 10 thèmes.

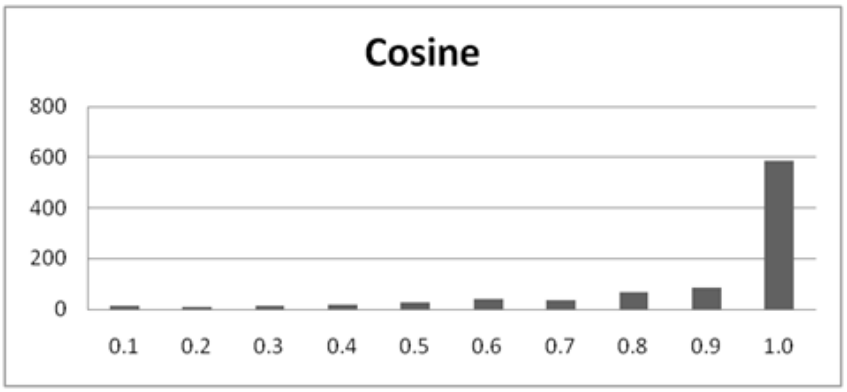


FIGURE 4.11 – Distribution empirique des valeurs prises par la distance cosinus.

4.1.2 Distance euclidienne

Définition 9. La *distance euclidienne* de deux distributions de probabilités ϕ_1 et ϕ_2 est définie par :

$$eucliDiv(\phi_1, \phi_2) = \sqrt{\sum(\phi_{1i} - \phi_{2i})^2}$$

Par construction, $eucliDiv(\phi_1, \phi_2) \geq 0$. De plus, $eucliDiv(\phi_1, \phi_2) = 0$ si ϕ_1 et ϕ_2 sont similaires et $eucliDiv(\phi_1, \phi_2) \to \infty$ si ϕ_1 et ϕ_2 divergent. La figure 4.12 présente la répartition des 900 mesures de divergences calculées selon la distance euclidienne pour 10 modèles LDA consécutifs composés de 10 thèmes.

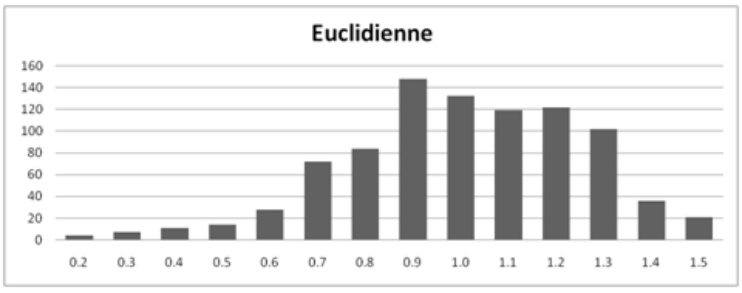


FIGURE 4.12 – Distribution empirique des valeurs prises par la distance euclidienne.

4.1.3 Distance de Jaccard généralisée

Définition 10. La *distance de Jaccard généralisée* de deux distributions de probabilités ϕ_1 et ϕ_2 est définie comme le complément à 1 de l'indice de Jaccard généralisé :

$$genJaccDiv(\phi_1, \phi_2) = 1 - genJaccInd(\phi_1, \phi_2)$$

où :

$$genJaccInd(\phi_1, \phi_2) = \frac{\sum \min(\phi_{1i}, \phi_{2i})}{\sum \max(\phi_{1i}, \phi_{2i})}$$

Par construction, $0 \leq genJaccInd(\phi_1, \phi_2) \leq 1$. De plus, $genJaccInd(\phi_1, \phi_2) = 0$ si ϕ_1 et ϕ_2 sont similaires et $genJaccInd(\phi_1, \phi_2) = 1$ si elles divergent. La figure 4.13 présente la répartition des 900 mesures de divergences calculées selon la distance de Jaccard généralisée pour 10 modèles LDA consécutifs composés de 10 thèmes.

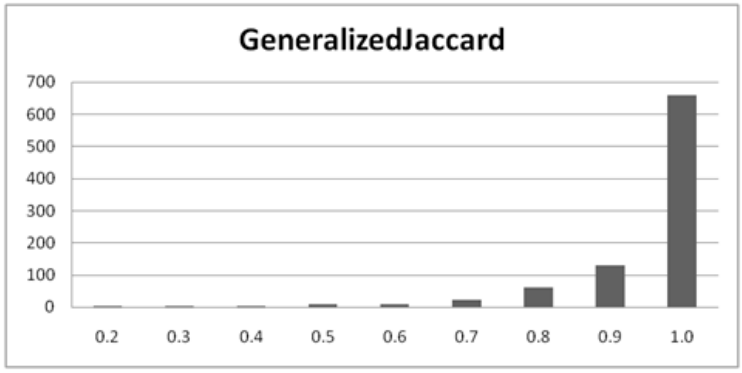


FIGURE 4.13 – Distribution empirique des valeurs prises par la distance de Jaccard.

4.1.4 Divergence de Kullback-Leibler

Définition 11. La *divergence de Kullback-Leibler* de deux distributions de probabilités ϕ_1 et ϕ_2 s'interprète comme la différence moyenne du nombre de bits nécessaires au codage d'échantillons de ϕ_1 selon que le codage est choisi optimal pour la distribution ϕ_1 ou ϕ_2. Elle est définie par :

$$KL(\phi_1, \phi_2) = \sum \phi_{1i}. \log(\frac{\phi_{1i}}{\phi_{2i}})$$

Afin de rendre compte de la nature de nos données, nous proposons d'utiliser ici une version symétrique classique de cette mesure :

$$\begin{aligned} KLDiv(\phi_1, \phi_2) &= \frac{1}{2} KL(\phi_1, \phi_2) + \frac{1}{2} KL(\phi_2, \phi_1) \\ &= \frac{1}{2} \sum \phi_{1i}. \log(\frac{\phi_{1i}}{\phi_{2i}}) + \frac{1}{2} \sum \phi_{2i}. \log(\frac{\phi_{2i}}{\phi_{1i}}) \end{aligned}$$

On a : $KLDiv(\phi_1, \phi_2) \geq 0$. De plus, $KLDiv(\phi_1, \phi_2) = 0$ si ϕ_1 et ϕ_2 sont similaires et $KLDiv(\phi_1, \phi_2) \to \infty$ si ϕ_1 et ϕ_2 divergent. La figure 4.14 présente la répartition des 900 mesures de divergences calculées selon la divergence de Kullback-Leibler pour 10 modèles LDA consécutifs composés de 10 thèmes.

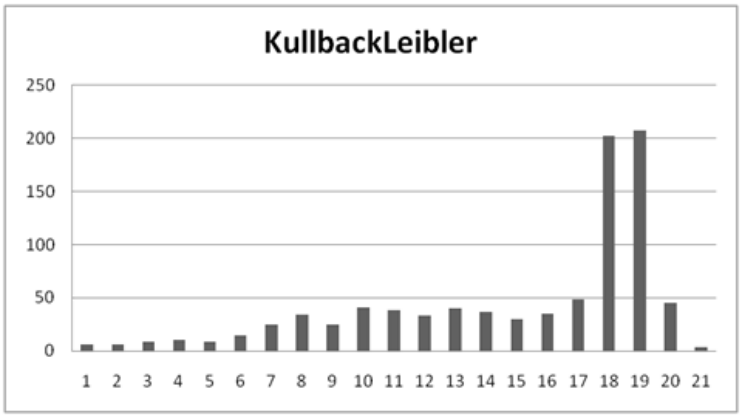


FIGURE 4.14 – Distribution empirique des valeurs prises par la divergence de Kullback-Leibler.

4.1.5 Divergence de Jensen-Shannon

Définition 12. La *divergence de Jensen-Shannon* est une autre version symétrique classique de Kullback-Leibler. Elle est définie par :

$$JS(\phi_1, \phi_2) = \frac{1}{2} KL(\phi_1, \phi) + \frac{1}{2} KL(\phi_2, \phi)$$

, où $\phi = \frac{1}{2}(\phi_1 + \phi_2)$. On a donc :

$$JSDiv(\phi_1, \phi_2) = \frac{1}{2} \sum \phi_{1i}. \log(\frac{\phi_{1i}}{\frac{1}{2}(\phi_{1i} + \phi_{2i})}) + \frac{1}{2} \sum \phi_{2i}. \log(\frac{\phi_{2i}}{\frac{1}{2}(\phi_{1i} + \phi_{2i})})$$

On a : $0 \leq JSDiv(\phi_1, \phi_2) \leq \ln 2$, avec $JSDiv(\phi_1, \phi_2) = 0$ si ϕ_1 et ϕ_2 sont similaires et $JSDiv(\phi_1, \phi_2) = \ln 2$ si ϕ_1 et ϕ_2 divergent. La figure 4.15 présente la répartition des 900 mesures de divergences calculées selon la divergence de Jensen-Shannon pour 10 modèles LDA consécutifs composés de 10 thèmes.

FIGURE 4.15 – Distribution empirique des valeurs prises par la divergence de Jensen-Shannon.

4.1.6 Divergence de Bhattacharyya

Définition 13. La *divergence de Bhattacharyya* s'appuie sur le coefficient de Bhattacharyya qui est une mesure statistique du recouvrement de deux ensembles d'échantillons. Elle est définie par :

$$BCDiv(\phi_1, \phi_2) = -\ln(\sum \sqrt{\phi_{1i}.\phi_{2i}})$$

On a : $BCDiv(\phi_1, \phi_2) \geq 0$. De plus, $BCDiv(\phi_1, \phi_2) = 0$ si ϕ_1 et ϕ_2 sont similaires et $BCDiv(\phi_1, \phi_2) \to \infty$ si ϕ_1 et ϕ_2 divergent. La figure 4.16 présente la répartition des 900 mesures de divergences calculées selon la divergence de Bhattacharyya pour 10 modèles LDA consécutifs composés de 10 thèmes.

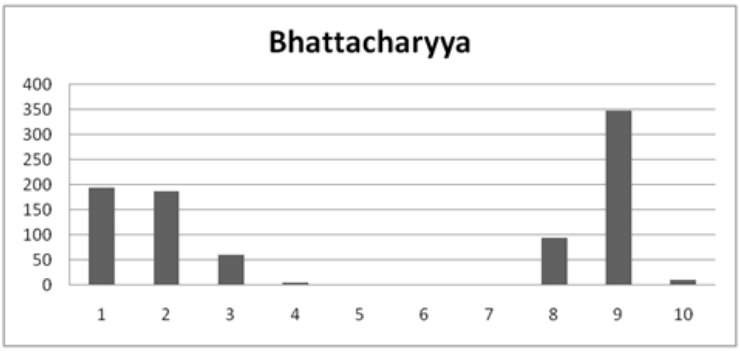


FIGURE 4.16 – Distribution empirique des valeurs prises par la divergence de Bhattacharyya.

4.1.7 Distance de Hellinger

Définition 14. Contrairement à la divergence de Bhattacharyya, la *distance de Hellinger* satisfait l'identité triangulaire. Elle est définie à partir du coefficient de Bhattacharyya par :

$$HDiv(\phi_1, \phi_2) = \sqrt{1 - \sum \sqrt{\phi_{1i}.\phi_{2i}}}$$

Par construction, $0 \leq HDiv(\phi_1, \phi_2) \leq 1$. De plus, $HDiv(\phi_1, \phi_2) = 0$ si ϕ_1 et ϕ_2 sont similaires et $HDiv(\phi_1, \phi_2) = 1$ si elles divergent. La figure 4.17 présente la répartition des 900 mesures de divergences calculées selon la distance de Hellinger pour 10 modèles LDA consécutifs composés de 10 thèmes.

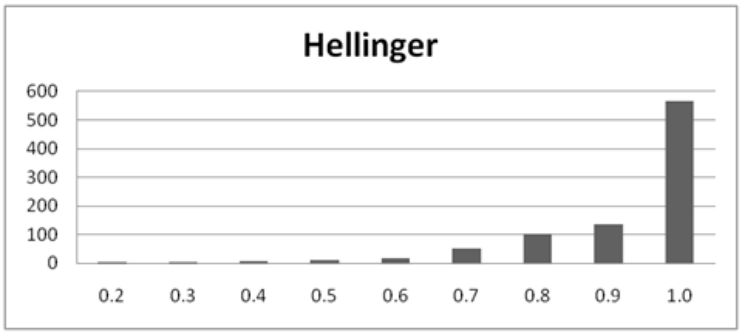


FIGURE 4.17 – Distribution empirique des valeurs prises par la distance de Hellinger.

Discussion. L'analyse des valeurs prises par ces sept mesures de divergence montre de grandes disparités dans les distributions observées. Un premier groupe émerge toutefois, formé de plusieurs divergences présentant une allure générale en forme de fonction exponentielle et où les thèmes calculés apparaissent différents les uns des autres (Cosine, Jaccard, Jensen-Shannon et Hellinger, voir figure 4.11, 4.13, 4.15 et 4.17). Néanmoins, les distances Euclidienne et de Bhattacharyya présentent des comportements complètement opposés : la première semble concentrer beaucoup de ses valeurs dans le ventre de la distribution impliquant une faible variance (voir figure 4.12), alors que la seconde tend à diviser les thèmes en deux groupes (voir figure 4.16). Avec la divergence de Bhattacharyya, les thèmes sont soit similaires, soit différents, mais il n'y a pas d'entre deux. En outre, la divergence de Kullback-Leibler, fréquemment utilisée dans la littérature pour comparer des distributions de probabilités, semble répartir ses valeurs un peu plus uniformément que les divergences du premier groupe, même si la tendance générale reste une grande dissimilarité entre thèmes (voir figure 4.14). Ce comportement illustre à la fois la capacité de LDA à construire des thèmes discriminants et le contexte de forte nouveauté de Twitter qui implique que les thèmes construits sur des périodes différentes ne partagent pas le même vocabulaire.

4.2 Courbes cumulées

En normalisant ces distributions entre $[0; 1]$ et en visualisant les sommes cumulées (voir figure 4.18), il est possible de comparer sur un même graphique la proportion de liens créés par chaque divergence en fonction du seuil choisi par l'analyste. Cette représentation permet d'anticiper sur le chapitre suivant qui présente un outil interactif permettant d'observer les chaînes de thèmes crées pour différentes valeurs de seuil. L'axe des abscisses peut, par exemple, être considéré comme un curseur que l'analyste peut déplacer afin de faire varier le seuil de similarité, et l'axe des ordonnés correspond alors à la proportion de liens affichés sur la visualisation.

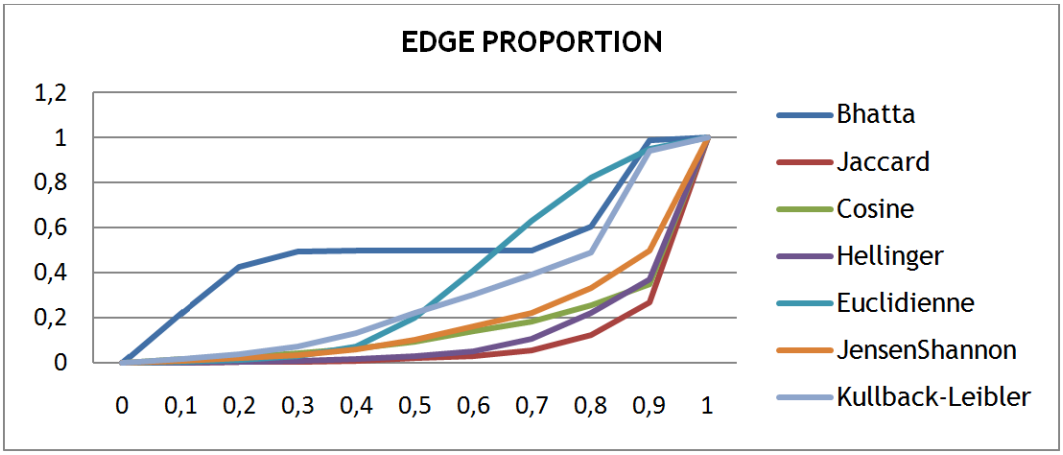


FIGURE 4.18 – Évolution du nombre de liens créés en fonction du seuil de similarité pour les sept divergences étudiées.

Discussion. Il ressort de cette visualisation que les divergences du premier groupe (Cosine, Jaccard, Jensen-Shannon et Hellinger) concentrent tous leurs liens sur la partie droite de l'axe des abscisses, ce qui n'offre pas beaucoup de marge de manœuvre à l'analyste. En effet, celui-ci ne verra pas beaucoup de changements dans les liens affichés sur la visualisation sur toute la partie gauche de l'axe. La distance Euclidienne, quant-à-elle, présente une courbe en forme de sigmoïde indiquant une grande variabilité au centre de l'axe. L'analyste sera alors à même de modifier rapidement la visualisation autour de ces valeurs centrales. La divergence de Bhattacharyya, au contraire, affiche un plateau au milieu de la distribution qui offre un intervalle dans lequel n'importe quelle valeur de seuil mènera au même ensemble de liens affichés à l'écran. Ce comportement semble offrir une bonne opportunité de fournir à l'analyste une valeur de seuil par défaut, pour laquelle la divergence a tranché : les liens affichés correspondent aux thèmes similaires. Il pourrait alors être intéressant d'analyse les chaînes construites dans cette configuration. Pour finir, la courbe de la divergence de Kullback-Leibler, même si elle présente une allure similaire aux divergences du premier groupe, semble être celle qui s'approche le plus de la diagonale, ce qui se traduirait par une évolution du nombre de liens la plus homogène. L'analyste pourrait alors explorer les variations induites par un changement de seuil avec d'avantage de précisions.

Ces différentes mesures proposées dans la littérature semblent avoir des comportements très différents sur nos données, ce qui peut s'avérer problématique dans le cadre d'un outil de fouille interactif. En effet, le fait que les relations entre thèmes différent totalement d'un mesure à l'autre, serait signe d'une forte sensibilité à ce paramètre. Dans le but d'évaluer la robustesse de notre approche, nous proposons de calculer sur l'ensemble des mesures deux coefficients de corrélation très répandus dans la littérature : le coefficient τ de Kendall et le coefficient ρ de Spearman. Une étude de la corrélation entre les rangs des valeurs prises par ces mesures de

divergence nous permettra de mesurer à quel point les chaînes de thèmes présentées à l'analyste correspondent à une évolution fiable de concepts.

4.3 Corrélations entre divergences

Nous avons comparé chaque mesure de divergence deux à deux en analysant les valeurs attribuées à chaque paire de thèmes pris entre les deux premiers modèles de notre période d'analyse (10 thèmes par modèle, soit 10*10 = 100 valeurs pour chaque divergence). Étant données deux mesures de divergence D_x et D_y, nous avons construit deux ensembles d'observations $(x_{ij})_{10\times10}$ et $(y_{ij})_{10\times10}$ tels que : $\forall i,j \in \{1,\ldots,10\}, x_{ij} = D_x(i,j)$ et $y_{ij} = D_y(i,j)$. Chaque ensemble est donc composé de 100 individus.

4.3.1 Coefficient de Kendall

Le coefficient τ de Kendall est une mesure de corrélation de rang qui, étant donné un ensemble d'observations composé de paires $P_n = \{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, compare le nombre de paires concordantes et le nombre de paires discordantes. Il estime à quel point deux mesures induisent le même ordre indépendant des valeurs absolues. Deux paires (x_i,y_i) et (x_j,y_j) sont dites concordantes si $x_i < x_j$ et $y_i < y_j$ ou si $x_i > x_j$ et $y_i > y_j$ à l'inverse, elles sont dites discordantes si $x_i < x_j$ et $y_i > y_j$ ou si $x_i > x_j$ et $y_i < y_j$. Le coefficient de Kendall τ est défini par :

$$\tau = \frac{(\textit{Nombre de paires concordantes}) - (\textit{Nombre de paires discordantes})}{\frac{1}{2}n(n-1)}$$

Par construction, on a : $-1 \leq \tau \leq 1$ avec $\tau = 1$ si les observations sont positivement corrélées, $\tau = 0$ si elles sont indépendantes et $\tau = -1$ si elles sont négativement corrélées.

Dans notre cas, le calcul s'effectue en comparant deux à deux les paires prises dans $P_{100} = \{(x_{00},y_{00}),(x_{01},y_{01}),\ldots,(x_{nn},y_{nn})\}$, soit $\frac{100(100-1)}{2} = 4950$ comparaisons.

4.3.2 Coefficient de Spearman

Le coefficient ρ de Spearman est une mesure de corrélation de rang entre deux variables qui quantifie à quel point une des deux variables est une fonction monotone de l'autre. Il mesure la similarité entre les ordres induits. Étant donnés deux ensembles d'observations $(x_i)_n$ et $(y_i)_n$, le calcul du coefficient de Spearman nécessite en premier lieu de construire les variables intermédiaires $(X_i)_n$ et $(Y_i)_n$ correspondant respectivement au rang de l'élément x_i dans la suite $(x_i)_n$ et au rang de

KENDALL TAU	Bhattacharyya	Cosine	Euclidean	Jaccard	Hellinger	Jensen-Shannon	Kullback-Leibler
Bhattacharyya		0.92	0.5	0.87	1.0	0.97	0.92
Cosine	0.92		0.5	0.9	0.9	0.9	0.9
Euclidean	0.5	0.5		0.5	0.5	0.5	0.5
Jaccard	0.87	0.9	0.5		0.9	0.9	0.8
Hellinger	1.0	0.9	0.5	0.9		0.96	0.9
Jensen-Shannon	0.97	0.9	0.5	0.9	0.96		0.9
Kullback-Leibler	0.92	0.9	0.5	0.8	0.9	0.9	

FIGURE 4.19 – Corrélation de rang (coefficient tau de Kendall) entre les différentes divergences.

l'élément y_i dans la suite $(y_i)_n$. Le coefficient de Spearman ρ est alors défini par :

$$\rho = 1 - \frac{6\sum(X_i - Y_i)^2}{n(n^2-1)}$$

Par construction, on a : $-1 \leq \rho \leq 1$ avec $\rho = 1$ si les observations sont positivement corrélées, $\rho = 0$ si elles sont indépendantes et $\rho = -1$ si elles sont négativement corrélées.

Remarque 8. *Si les observations sont rangées dans le même ordre, on a :* $\forall i \in X_i = Y_i$ *et* $\sum(X_i - Y_i)^2 = 0$. *A l'inverse, si les observations sont rangées exactement dans l'ordre contraire, on a :* $\sum(X_i - Y_i)^2 = \frac{n(n^2-1)}{3}$. *Afin de normaliser cette somme entre* $[-1, 1]$ *où* $\rho = 1$ *indique une corrélation positive en* $\rho = -1$ *indique une corrélation négative, la transformation affine,* $f : X \to -\frac{6}{n(n^2-1)}X + 1$ *est appliquée. Cette application définit alors un chemin de la forme* $f(x) = ax + b$ *tel que* $f(0) = 1$ *et* $f(\frac{n(n^2-1)}{2}) = -1$.

Dans notre cas, le calcul du coefficient de Spearman revient à comparer le rang de chaque paire de thèmes (i, j) dans les suites $(x_{ij})_{nn}$ et $(y_{ij})_{nn}$.

Discussion. Malgré des répartitions apparentes différentes, les valeurs prises par ces sept mesures de divergence semblent très corrélées. Les liens sont alors créés dans le même ordre et il est possible d'obtenir des visualisations équivalentes en faisant varier le seuil de similarité. Cette étude quantitative présente pour nous un intérêt particulier puisqu'elle permet de valider le faible impact du choix de la mesure de divergence sur les résultats produits par notre méthodologie. En effet, il est alors possible de choisir la mesure de divergence dont la répartition des liens correspond le plus aux besoins de visualisation de l'analyste sans affecter la qualité des

SPEARMAN RHO	Bhattacharyya	Cosine	Euclidean	Jaccard	Hellinger	Jensen-Shannon	Kullback-Leibler
Bhattacharyya		0.99	0.7	0.98	1.0	0.99	0.99
Cosine	0.99		0.7	0.98	0.99	0.99	0.97
Euclidean	0.7	0.7		0.68	0.68	0.68	0.68
Jaccard	0.98	0.98	0.68		0.98	0.98	0.95
Hellinger	1.0	0.99	0.68	0.98		0.99	0.99
Jensen-Shannon	0.99	0.99	0.68	0.98	0.99		0.98
Kullback-Leibler	0.99	0.97	0.68	0.95	0.99	0.98	

FIGURE 4.20 – Corrélation de rang (coefficient rho de Spearman) entre les différentes divergences.

résultats.

Cette observation nous permet de sélectionner un nombre de mesures restreint en vue de conduire une évaluation complémentaire sur la base des visualisations induites. En effet, la forte corrélation des mesures étudiées nous permet d'en sélectionner trois en fonction de la répartition des valeurs de similarité calculées sur nos données :

1. La divergence de Bhattacharyya qui présente un caractère discriminant pouvant offrir une valeur de seuil par défaut pour une version non-paramétrique de notre méthodologie.

2. La distance Euclidienne qui exhibe une forte variabilité dans la répartition des liens et dont la corrélation plus faible avec les autres mesures tend à montrer que les chaînes créées seront différentes.

3. La divergence de Kullback-Leibler qui est très fortement corrélée aux mesures du premier groupe et dont la répartition homogène des valeurs en fait un bon représentant.

Remarque 9. *Malgré leurs distributions empiriques très différentes, les divergences de Bhattacharyya et de Hellinger affichent une corrélation parfaite. (Il est possible de le déduire analytiquement à partir de leurs définitions : l'une est fonction croissante de l'autre). Autrement dit, même si les valeurs absolues prises par ces deux mesures sont très éloignées, les distances relatives entre thèmes sont identiques.*

5 Conclusion

Dans ce chapitre, nous avons commencé par rappeler la définition du modèle de Latent Dirichlet Allocation [BNJ03] qui a contribué à populariser les "topic models"

dans la communauté de la fouille de texte. Dans un premier temps, nous avons précisé le processus générateur de ce modèle qui permet de décrire un corpus statique de documents (sans date de publication) au travers d'un ensemble de thèmes latents, définis sous forme de distributions de probabilités. Nous avons, ensuite, donné une présentation détaillée d'une approche, de la famille des chaînes de Markov pour estimer les paramètres du modèle à partir d'un corpus particulier. Puis, nous avons décrit une instance du modèle LDA calculée sur les données Twitter de notre corpus et nous avons analysé finement l'impact des paramètres du modèle sur les thèmes construits. Dans un second temps, nous avons développé une approche pour adapter ces modèles statiques aux flux de textes dynamiques. En nous inspirant des travaux de Malik et al. [MSH+13], nous avons proposé de partitionner le corpus par périodes temporelles de façon à construire un modèle par période ; ce qui permet aux thèmes construits d'être représentatifs d'une courte période temporelle. L'enjeu consiste ensuite à détecter les similarités entre les thèmes pour explorer leur évolution temporelle. Ainsi, nous avons proposé une comparaison expérimentale de sept mesures de divergences classiques sur un jeu de données réelles qui nous a permis d'identifier les mesures les plus appropriées. La restitution à l'analyste de l'étude longitudinale des proximités thématiques est décrite dans le chapitre suivant : elle s'appuie sur une représentation visuelle inspirée des diagrammes de Sankey.

5

Visualisation interactive

Mathematics, rightly viewed, possesses not only truth, but supreme beauty ? a beauty cold and austere.

—Bertrand Russel

Sommaire

1 Introduction

Si les efforts de la fouille de données restent encore ciblés sur le développement d'algorithmes de traitement performants capables d'affronter une volumétrie qui ne cesse d'exploser, une attention accrue est portée au rôle de l'utilisateur/décideur dans le processus de fouille (« human in the loop »). Elle a conduit à l'essor de la fouille visuelle de données ou Visual Analytics définie par Thomas et Cook dans leur papier séminal en 2005 comme « la science du raisonnement analytique facilité par des interfaces visuelles interactives » [TC05]. Les éléments structurants du processus de fouille visuelle ont été introduits dans un modèle générique présenté par Keim et al [KMS+08], à la suite d'une proposition antérieure de Wijk (2005) [Van05]. Le processus (voir figure 5.1) y est présenté comme une boucle de rétroaction dans laquelle l'utilisateur joue le rôle d'une heuristique qui guide les algorithmes d'exploration et d'analyse. Suite à une première analyse des données, l'utilisateur rentre dans une boucle de rétro-action permettant la formulation et le test de nouvelles hypothèses. La visualisation interactive fournit de nouvelles images des données qui permettent d'acquérir des connaissances et d'imaginer des hypothèses, hypothèses qui peuvent ensuite être testées grâce à la visualisation interactive. Cette approche centrée sur l'utilisateur qui combine des algorithmes de fouille avec des visualisations interactives a aujourd'hui fait ses preuves pour de nombreuses applications comme l'atteste la vitalité des conférences dédiées à la visualisation de l'information (e.g. IEEE Information Visualization, IEEE Visual Analytics Science and Technology, International Conference on Information Visualisation). Les réseaux sociaux ont été, et restent encore, un terrain fécond pour son développement. Mais au-delà des graphes « à la Facebook », les autres médias sociaux ont plus récemment conduit au développement de différents systèmes de visualisation interactive de l'information [SK13].

Dans ce chapitre, nous nous focalisons sur la visualisation de données textuelles issues de Twitter. Sans prétendre à l'exhaustivité, nous commençons, dans la section 5.2, par présenter quelques représentations de données textuelles qui sont basées sur des modèles de visualisation différents. Puis, nous décrivons, dans la section 5.3, l'approche que nous avons développée qui repose sur un diagramme de Sankey. Nous précisons les différentes composantes de la représentation et présentons ensuite des retours d'interprétation sur les données réelles du groupe EDF, dans la section 5.4. Finalement, puisque le choix du nombre de thèmes des modèles construits apparaît comme étant un paramètre crucial de notre méthodologie, nous proposons, dans la section 5.5, d'intégrer une variante non-paramétrique du modèle LDA afin d'estimer le nombre de thèmes à partir des données.

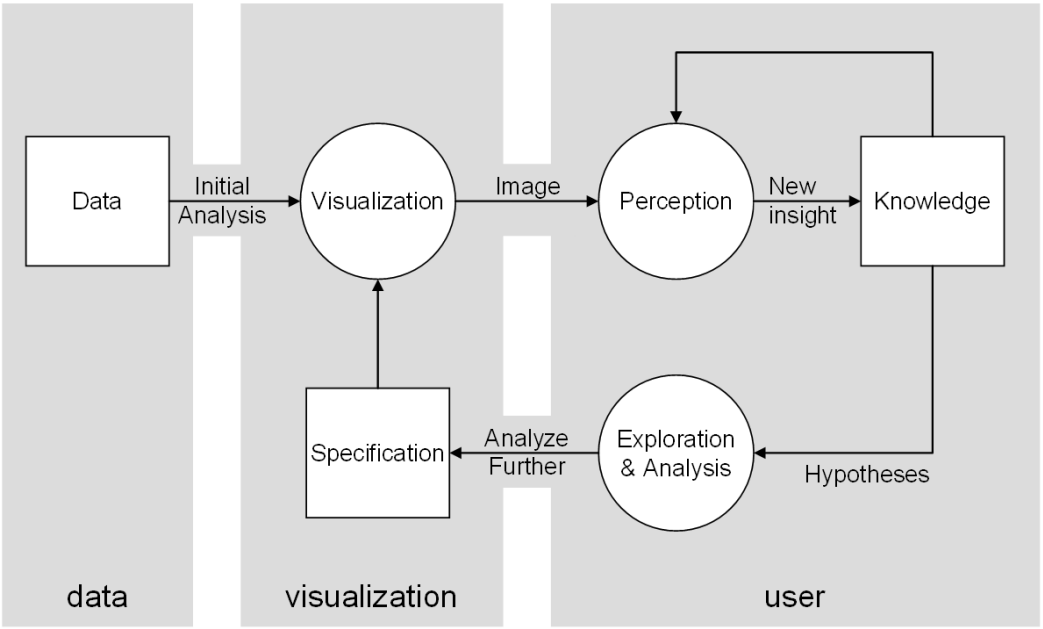


FIGURE 5.1 – Illustration du pipeline de fouille visuelle tiré de [KMS+08].

2 Visualisation de données issues de Twitter.

Stimulées par le nombre de données disponibles, de nombreuses méthodes de visualisation ont été développées pour proposer des vues synthétiques de corpus textuels volumineux. La quantité de messages publiés sur Twitter, même en se restreignant à une catégorie particulière, empêche toute lecture individuelle exhaustive. Il est donc nécessaire, non seulement de développer des outils d'analyse pour synthétiser l'information, mais également des méthodes de visualisation pour explorer les résultats. La récente synthèse proposée par Šilić et Bašić [vB10] regroupe plusieurs visualisions adaptées aux flux de textes. Nous nous contentons ici de présenter quelques propositions, qui nous paraissent parmi les plus pertinentes, pour présenter des données issues de Twitter.

Nuages de mots Popularisés par les premiers sites du Web 2.0 au début des années 2000, les nuages de mots ("tag clouds") [HK07] mettent en avant l'importance relative des termes à l'aide de variables visuelles comme la taille ou la couleur des mots. Bien que leur pertinence ait parfois été discutée, par exemple à cause du biais introduit par la différence de longueurs des mots[1], les nuages de mots restent très utilisés notamment pour les thèmes du modèle LDA [RDL10] ou les hashtags présents dans les tweets [SST+09]. Plusieurs nuages étudiés en parallèle peuvent également permettre de comparer des ensembles de mots différents, comme le vocabulaire de l'échantillon de tweets fourni par Twitter via l'API Streaming et celui du flux total obtenu grâce à un accès privilégié [MPLC13] ou encore les vocabulaires

[1] www.niemanlab.org/2011/10/word-clouds-considered-harmful/

associés à différentes émoticônes ("smileys") [PBFC13]. Mais, ces approches se restreignent à la représentation de corpus statiques en synthétisant l'information qu'ils contiennent. En plaçant une succession de nuages dans un graphe orienté (voir figure5.2), Lee et al. [LLM13] illustrent l'évolution du vocabulaire d'une thématique et adaptent les nuages de mots à la visualisation de la dimension temporelle.

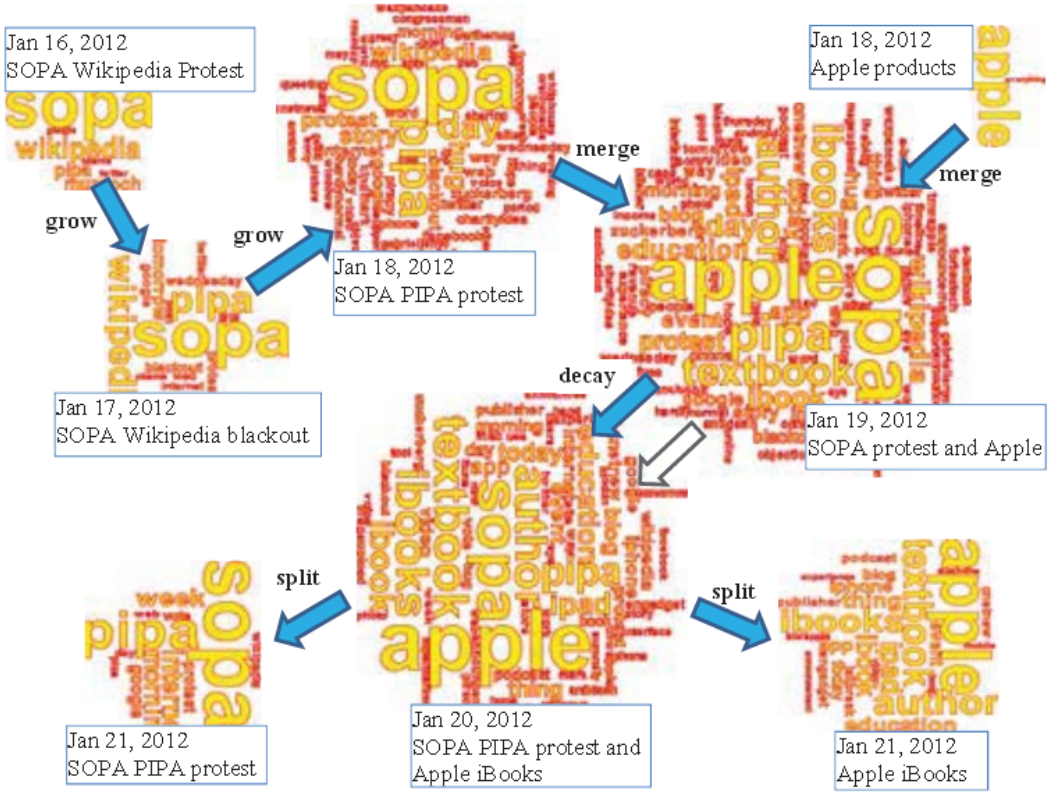


FIGURE 5.2 – Succession de nuages de mots illustrant l'évolution temporelle du vocabulaire. (Extrait de [LLM13])

Clustering L'approche de Gansner et al. [GHN12], inspirée des graphes dynamiques, propose une visualisation reprenant la métaphore des cartes de pays afin de visualiser l'évolution de groupes de tweets (voir figure 5.3). Chaque groupe est représenté sous forme d'un "pays" sur une carte et partage avec ses voisins des frontières qui évoluent au cours du temps. La visualisation est mise à jour régulièrement, et les nouveaux tweets sont intégrés en recalculant les groupes et leurs proximités. Si de nombreux efforts ont été faits pour conserver les positions des pays et leurs frontières, la vitesse des données Twitter et l'hétérogénéité des thématiques abordées peuvent rendre difficile la préservation de la carte mentale de l'utilisateur.

Séries temporelles Une approche intermédiaire entre la juxtaposition de représentations statiques et l'affichage dynamique peut être l'emploi de visualisations provenant de l'analyse des séries temporelles. La représentation sous forme d'aires empilées ("stacked area chart") permet notamment de mettre en correspondance les

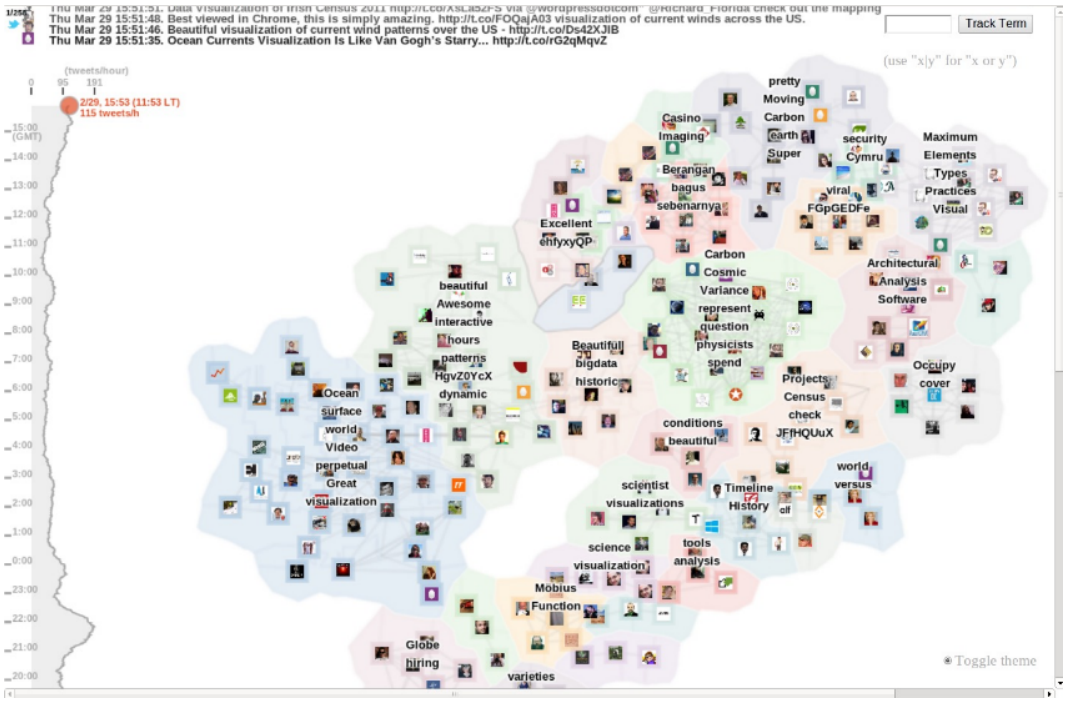


FIGURE 5.3 – Métaphore de carte utilisée par TwitterScope pour représenter l'évolution des résultats d'un algorithme de clustering dynamique. (Extrait de [GHN12])

valeurs d'une série temporelle avec la taille de l'aire qui la représente. En superposant plusieurs aires sur un même graphique, ce type de représentation permet de visualiser l'évolution de plusieurs séries temporelles ainsi que leurs contributions au motif total. Lu et Yang [LY12] l'utilisent, par exemple, pour illustrer comment les thématiques s'enchaînent dans le temps, en prenant chacune la place de la précédente. La contribution au flux total est donc à chaque étape le fait d'une unique thématique. Dou et al. [DWS+12] combinent, quant-à-eux, plusieurs graphes de flux ("stream-graphes" : variante centrée des aires empilées) pour visualiser un événement et ses sous-événements (voir figure 5.4). Le flux total est représenté en image de fond par transparence et chaque sous-événement est représenté par dessus grâce à un flux dont la taille correspond au volume d'information représenté. Une des limitations de la représentation sous forme d'aires empilées est que, même si elle permet de visualiser les contributions relatives de chaque composante au flux total, elle n'est pas adaptée à la visualisation de motifs de fusion et de division. En effet, les aires empilées représentent l'évolution de séries temporelles en termes de volume et il n'est pas possible de séparer ou de regrouper les différentes contributions.

Diagrammes de Sankey Les diagrammes de Sankey répondent à ce besoin en proposant une visualisation sous forme de diagramme de flux dans laquelle ceux-ci peuvent converger ou se séparer. Popularisés dans le monde industriel par leur utilisation dans l'illustration des transferts d'énergie d'une machine à vapeur [San98],

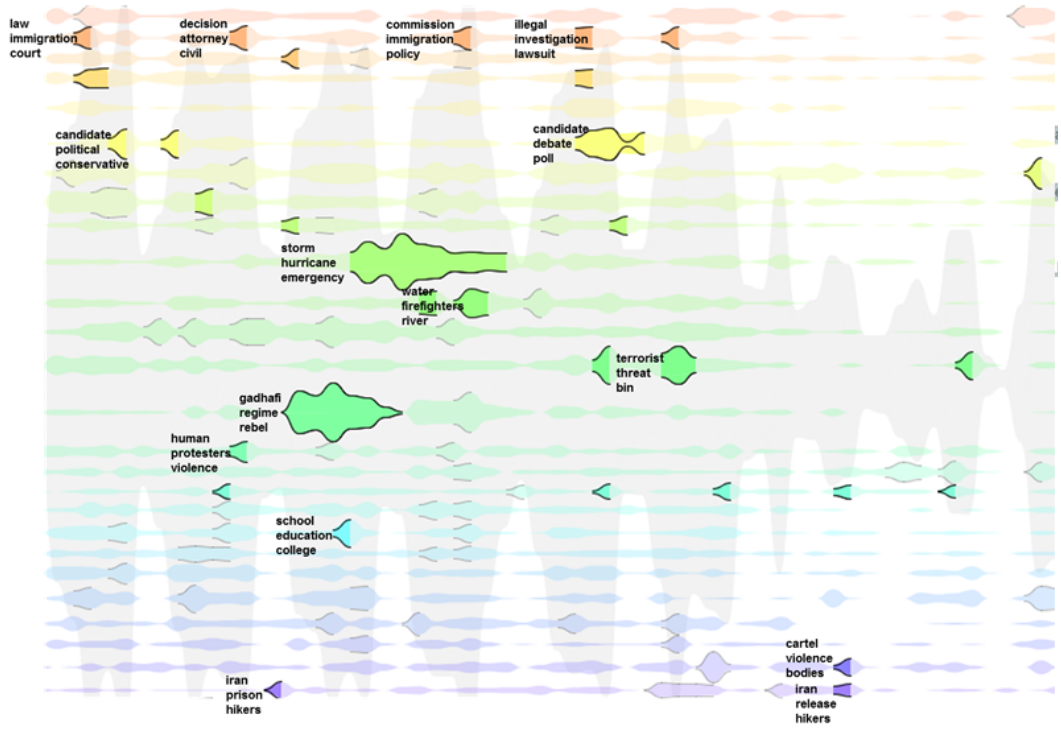


FIGURE 5.4 – Combinaison de plusieurs stram-graphes permettant de visualiser un événement et ses sous-événements. (Extrait de [DWS+12])

les diagrammes de Sankey avaient déjà été utilisés précédemment dans d'autres contextes, notamment pour visualiser l'évolution du nombre de soldats de Napoléon durant la campagne de Russie de 1812 (voir figure 5.5). Dans le premier cas, les flux représentent la quantité d'énergie échangée entre les différents composants du système, alors que dans le second, ils représentent les effectifs de l'armée napoléonienne à chaque étape clé de la campagne. Adaptés à la visualisation de flux de textes, ils ont été utilisés par Xu et al. [XWW+13] afin de visualiser l'évolution des sujets discutés dans les médias et les moments décisifs dans leur compétition. Ils ont également été utilisé dans l'outil TopicFlow [MSH+13], dont notre visualisation s'inspire, afin de représenter l'évolution de thèmes construits grâce au modèle LDA (voir figure 5.6). Toutefois, dans cette implémentation, les flux semblent partiellement discrétisés et les transferts entre flux n'apparaissent pas naturellement. Nous proposons dans la suite, une implémentation des diagrammes de Sankey adaptée à la visualisation de thèmes construits avec le modèle LDA qui conserve l'aspect de flux des visualisations classiques.

3 Diagramme de Sankey interactif

Afin de représenter l'évolution sémantique des thématiques en mettant en avant leurs motifs de fusion et de division, nous proposons une visualisation sous forme

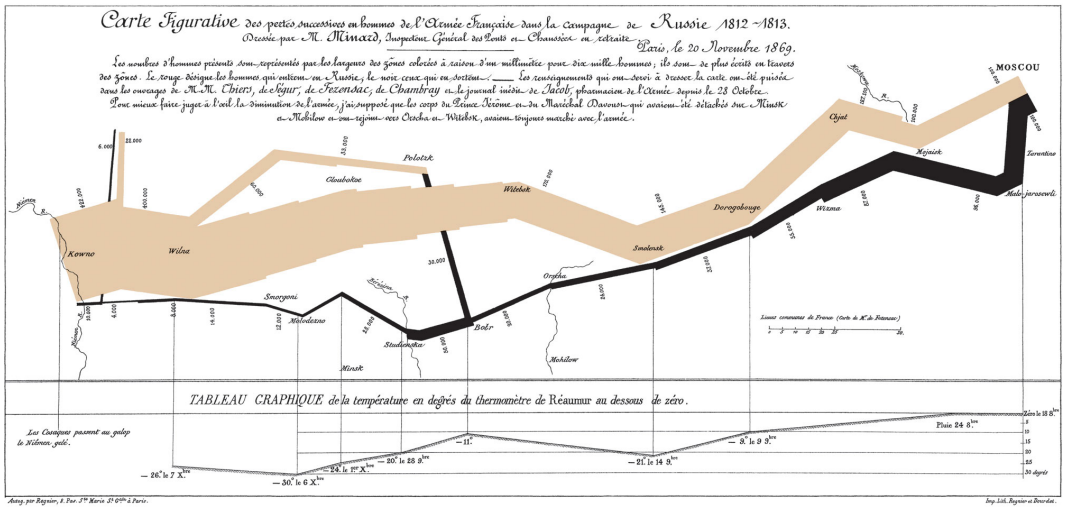


FIGURE 5.5 – Diagramme de Sankey combiné avec une carte géographique représentant l'évolution du nombre de soldats de l'armée napoléonienne au cours de la campagne de Russie. (Charles Minard)

de diagramme de Sankey utilisant la librairie Javascript "Data-Driven Document"[2] (voir figure 5.7). Dans cette visualisation, l'échelle temporelle est représentée en abscisse et les thèmes construits par chaque modèle successif sont représentés en colonne. Chaque thème z est associé à un nœud (rectangle vertical) dont la taille w_z est proportionnelle aux probabilités cumulées des tweets de la période.

$$w_z \propto \sum_{d \in \mathcal{D}_i} p(z \mid d)$$

Contrairement à l'implémentation de TopicFlow, dans laquelle chaque document ne contribue, à hauteur de 1, qu'à son thème le plus représentatif sans prendre en compte sa probabilité, ici, chaque tweet contribue au poids de plusieurs thèmes en fonction de la probabilité qui leur est associée. Néanmoins, puisque la somme de ses contributions vaut 1 (distribution de probabilités), comme dans TopicFlow, la somme des poids des thèmes d'une période correspond bien au nombre total de tweets. Au lieu de donner tout son poids à un unique thème, chaque document distribue ses contributions à l'ensemble des thèmes en fonction de leur probabilité.

Les liens entre thématiques permettent de construire des chaînes sémantiques et sont représentés par des arêtes entre les nœuds (voir figure 5.8).

La similarité entre thèmes dépend de la mesure de divergence utilisée, et une comparaison des visualisations induites par différentes mesures est donnée dans la section 5.4.

Étant donnés une mesure de divergence δ et un seuil ρ, une arête est construite entre deux nœuds de deux périodes consécutives si la divergence entre les thèmes correspondants, z_{source} et z_{cible}, est inférieure au seuil ρ sélectionné ; c'est-à-dire, si

[2] d3js.org/

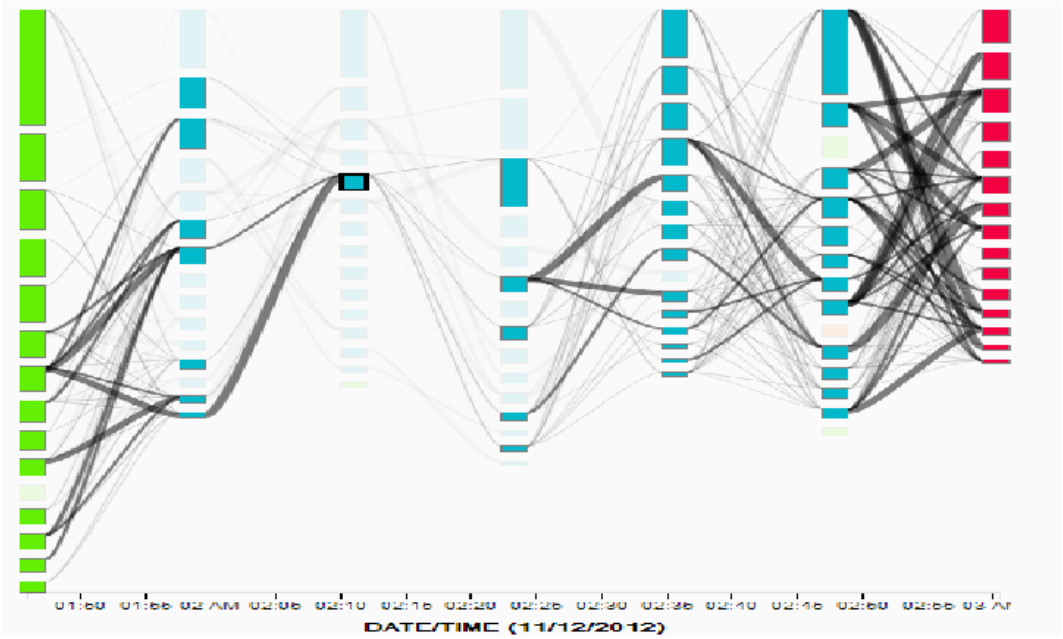


FIGURE 5.6 – Diagramme de Sankey représentant une succession de thèmes construits avec le modèle LDA. (Extrait de [MSH+13])

le thème z_{source} "évolue en" z_{cible} : $z_{source} \triangleright_{\delta\rho} z_{cible}$. Un lien entre thèmes est représenté par une arête a dont l'épaisseur w_a est fonction des poids des deux thèmes reliés et de leur similarité. La fonction utilisée ici est donnée par :

$$w_a = (w_{z_{source}} + w_{z_{cible}}) \times (1 - \delta(z_{source}, z_{cible}))$$

Cette fonction permet d'illustrer la part de la thématique source qui se retrouve dans la thématique cible.

Un flux gris est utilisé arrière-plan pour représenter le nombre total de documents de chaque période. Il permet de mettre en regard le poids des thèmes et le volume de documents. Les thèmes isolés sont regroupés sur la partie haute de la visualisation afin d'identifier les thèmes spécifiques à une période et les thèmes faisant partie d'une chaîne thématique sont superposés au flux gris afin de mettre en avant la contribution des chaînes au flux total. Cette répartition permet de visualiser la part du flux expliquée par les chaînes.

Par ailleurs, les thèmes faisant partie d'une chaîne sont organisés verticalement, au sein du flux, de telle sorte que le nombre de croisements d'arêtes soit minimisé, et la couleur de chaque chaîne est choisie sur la distribution de couleurs du CIELAB[3] afin de maximiser les distances perçues entre thèmes non reliés.

Par défaut, chaque thème est décrit par les deux termes dont la probabilité est maximale. Toutefois, afin d'accéder à une description plus précise, un nuage de mots est affiché au dessus de la visualisation lorsque l'utilisateur clique sur un nœud (voir figure 5.9). Ce nuage contient les 50 termes les plus représentatifs du thème

[3] scanline.ca/hue/cielab.html

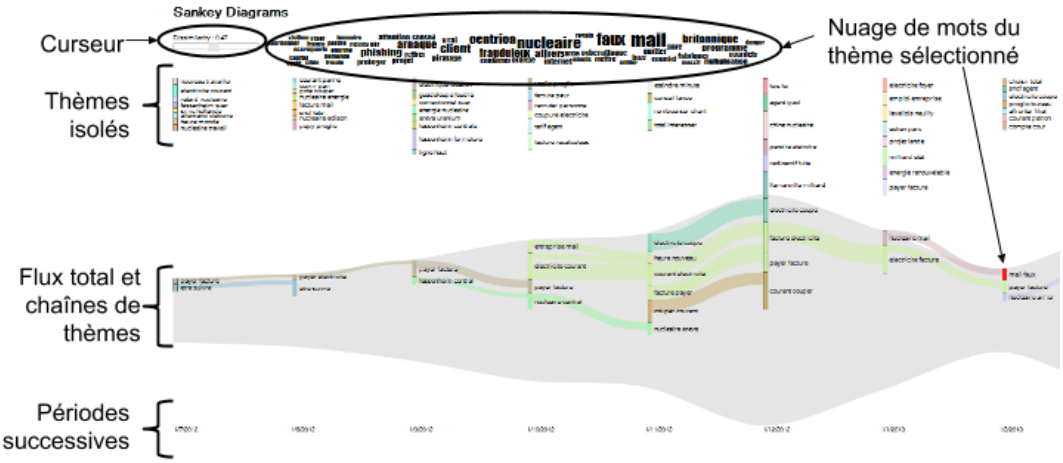


FIGURE 5.7 – Vue générale de l'interface graphique décrivant les différents éléments de visualisation.

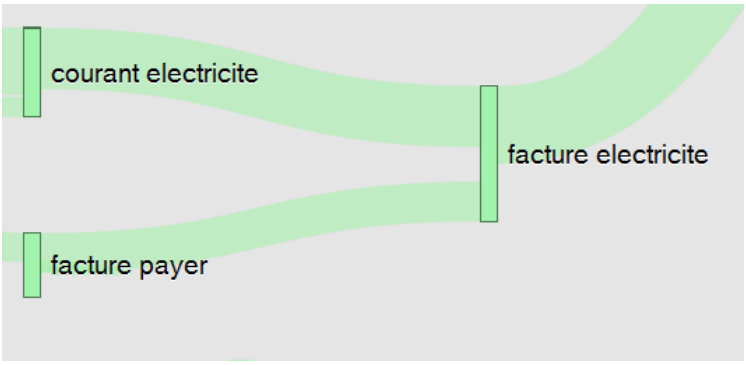


FIGURE 5.8 – Zoom sur un motif de fusion. Deux thèmes convergent vers un thème commun.

sélectionné et la taille de la police de caractères est proportionnelle à la probabilité du terme. La visualisation, sous forme de nuages de mots, de l'importance relative des termes au sein d'un thème permet à l'analyste d'identifier les thèmes décrits par peu de termes très saillants ou, à l'inverse, ceux décrits par un nombre plus important de termes dont les probabilités sont plus homogènes.

L'interface permet à l'utilisateur de déplacer verticalement des nœuds afin de mettre de côté un thème qu'il a déjà étudié, de visualiser des zones où il resterait des croisements d'arêtes (qui sont parfois inévitables) ou encore d'ajouter ces nœuds au flux total afin d'en visualiser la contribution.

Un mécanisme de zoom permet de mettre en avant des zones particulières, ou, à l'inverse, d'observer une vue d'ensemble.

Pour finir, un curseur permet de sélectionner un seuil de similarité régissant les liens à afficher sur la visualisation. Cette fonctionnalité permet de visualiser des tendances fortes lorsque le seuil est faible et exige de forte similarités ou à l'inverse d'explorer des évolutions moins évidentes lorsque la contrainte sur la similarité est relâchée.

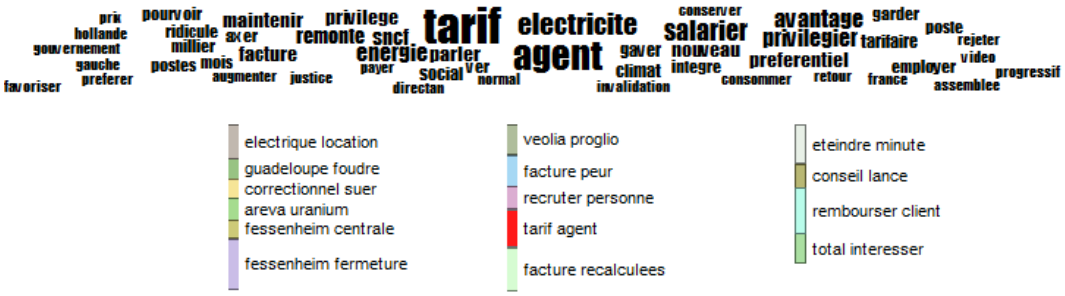
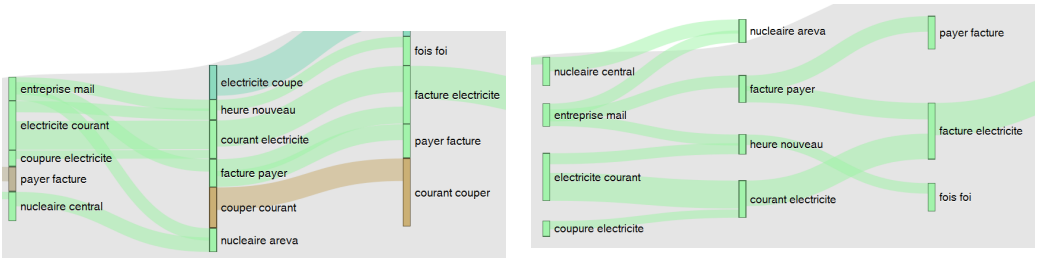


FIGURE 5.9 – Zoom sur un nuage de mot affiché lorsqu'un thème est sélectionné (en rouge), ici un des thèmes isolés.



(a) Positionnement des nœuds par défaut.

(b) Nœuds réorganisés afin d'identifier les croisements.

FIGURE 5.10 – Utilisation du glisser/déplacer pour réorganiser les nœuds et explorer une zone où les croisements sont inévitables.

Lorsque l'analyste modifie le seuil à l'aide du curseur, une animation fluide fait monter ou descendre les thèmes entre le flux et le haut de la visualisation afin de valoriser les zones modifiées. Si l'analyste relâche la contrainte, de nouveaux thèmes peuvent rejoindre les chaînes, ils en prennent la couleur et se placent de sorte à ne pas introduire de nouveaux croisements (si possible). A l'inverse, si l'analyste augmente les contraintes pour visualiser des liens plus forts, les thèmes remontent sur la partie supérieure et une couleur libre leur est assignée. Ces interactions permettent à l'analyste de choisir le bon niveau de granularité pour suivre des évolutions pertinentes. Notons qu'une stratégie d'interaction possible pourrait être d'initialiser la visualisation avec une contrainte forte, conduisant à la détection d'un nombre restreint de liens, puis d'augmenter progressivement le seuil afin d'afficher davantage de liens. Cela pourrait contribuer à faciliter la détection des motifs surprenants en raffinant la recherche de façon "continue". Cette stratégie d'exploration reprend le célèbre mantra de Shneiderman [Shn96] : "Overview first, zoom and filter, then details-on-demand".

4 Comparaison des visualisations induites par les trois mesures de liens entre thèmes sélectionnées

Suite à l'analyse des distributions empiriques de sept mesures de divergence entre les thèmes et à l'étude de leurs corrélations, conduites dans la section 4.4, nous avons sélectionné trois mesures présentant des comportements différents : la divergence de Bhattacharyya, dont la distribution présente un plateau illustrant son caractère discriminant ; la distance euclidienne qui affiche une grande variabilité au centre de la distribution et dont la corrélation avec les autres mesures étudiées est moindre ; et la divergence de Kullback-Leibler qui offre la répartition la plus homogène des liens. Puisque rien n'indique *a priori* quelle mesure est à privilégier, nous proposons, dans une logique de fouille exploratoire, d'étudier leur impact sur la visualisation de chaînes de thématiques dans un diagramme de Sankey. Nous présentons, pour chacune de ces divergences plusieurs diagrammes obtenus pour différents seuils de similarité et exhibant des comportements différents.

4.1 Divergence de Bhattacharyya

Le plateau observé sur la courbe cumulée de la distribution de la divergence de Bhattacharyya nous permet d'obtenir la même visualisation en choisissant arbitrairement une valeur de seuil entre 0,3 et 0,8. En effet, aucun lien n'est ajouté entre ces valeurs. Ce caractère discriminant, illustrant un comportement binaire de la mesure, entre paires de thèmes fortement similaires et paires de termes fortement différentes, semblait pouvoir nous permettre de proposer une visualisation par défaut à l'analyste. Malheureusement, il apparaît sur la figure 5.11 que les liens affichés pour une valeur de seuil comprise entre 0,3 et 0,8 sont trop nombreux et toutes les chaînes fusionnent en une unique chaîne difficile à analyser.

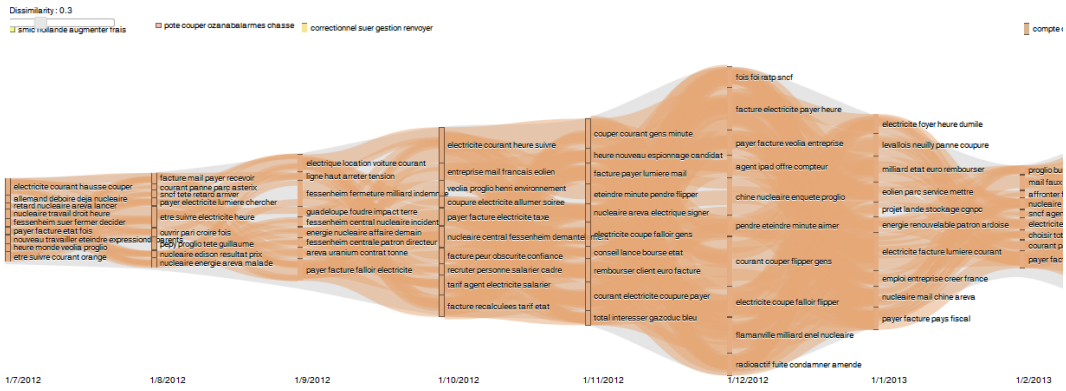


FIGURE 5.11 – Diagramme de Sankey illustrant la chaîne unique de thèmes LDA calculées selon la divergence de Bhattacharyya pour une valeur de seuil de 0,3. Selon cette valeur par défaut qui se situe sur le "plateau" de la divergence, les thèmes sont presque en totalité considérés comme similaires et reliés sur la visualisation.

Il est alors nécessaire de baisser le seuil afin d'augmenter les contraintes et de faire apparaître des chaînes séparées (voir figure 5.12).

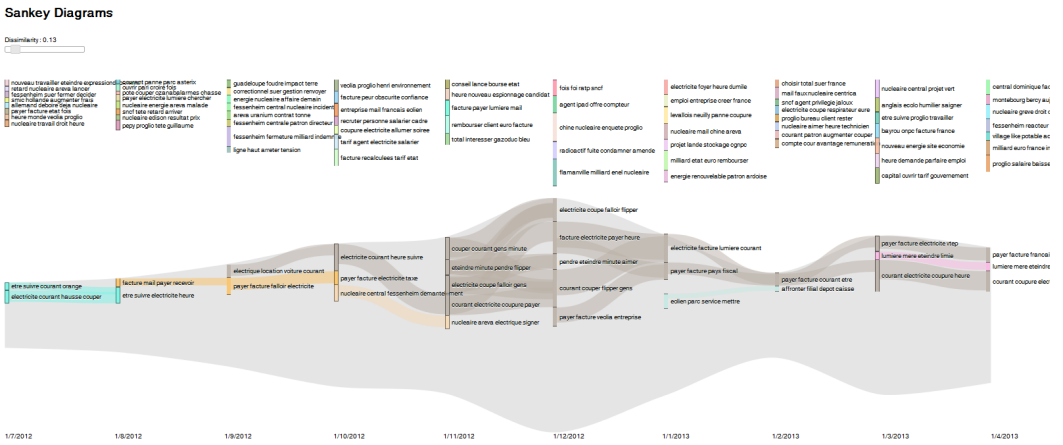


FIGURE 5.12 – Diagramme de Sankey illustrant les chaîne de thèmes LDA obtenues selon la divergence de Bhattacharyya pour une valeur de seuil de 0,13. Il est possible de diminuer le nombre de liens affichés en baissant la valeur de seuil, toutefois toute la variabilité est contenue dans l'intervalle [0 ;0,3].

Cette configuration n'est pas la plus adaptée pour l'analyste puisque la variabilité des liens est alors concentrée entre 0 et 0,3. Par ailleurs, puisque la divergence de Bhattacharyya est fortement corrélée à la divergence de Kullback-Leibler, il est préférable d'utiliser cette dernière.

4.2 Distance Euclidienne

La forme en sigmoïde de la courbe cumulée de la distribution de la distance euclidienne indique que très peu de liens seront créés sur la visualisation pour des valeurs de seuil faibles. Et en effet, comme l'illustre la figure 5.13, il est nécessaire de placer le seuil de similarité à 0,47 afin de voir apparaître les premières chaînes de thèmes.

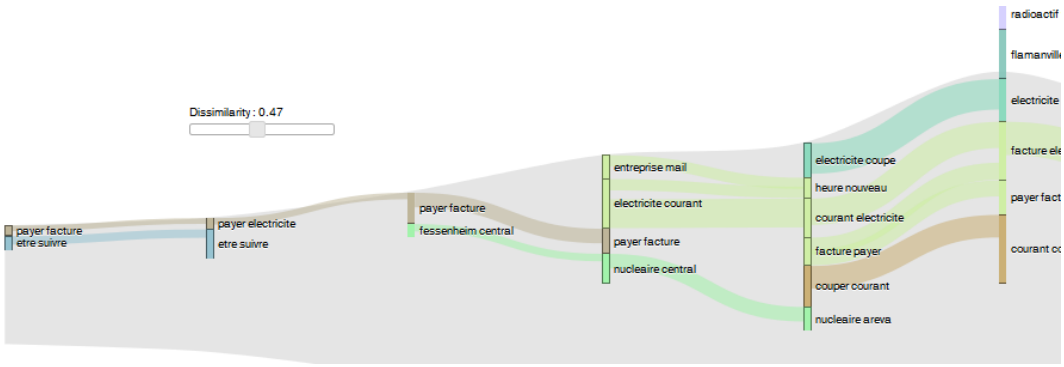


FIGURE 5.13 – Diagramme de Sankey illustrant les chaînes de thèmes LDA calculées selon la distance euclidienne pour une valeur de seuil de 0,47. Plusieurs thèmes sont liés d'une période à l'autre et des motifs de fusion et de séparation apparaissent, signe d'une évolution temporelle des thématiques.

L'analyse des chaînes créées dans cette configuration permet d'identifier des comportements intéressants : par exemple, comme l'illustre la figure 5.14, trois thèmes de trois modèles successifs mettent en avant un glissement de concept ayant lieu dans les données. Le thème "centrale de Fessenheim", à gauche sur la figure, décrit une thématique autour d'une centrale spécifique et se généralise au concept de "centrale nucléaire" dans le thème du milieu pour enfin évoluer, dans le thème de droite, vers "Areva", l'un des acteurs incontournable de la filière du nucléaire.

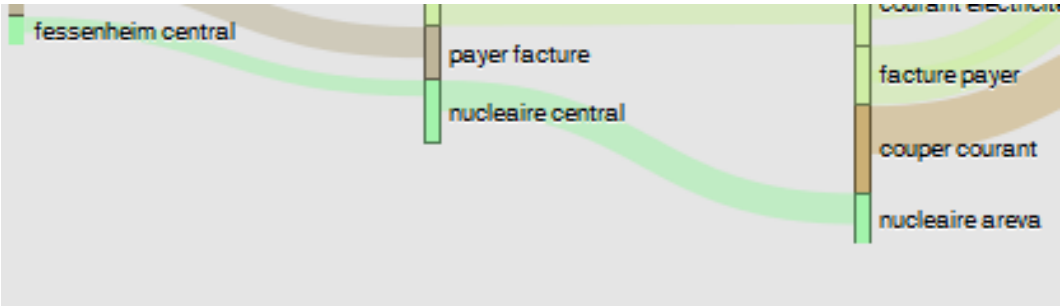


FIGURE 5.14 – Zoom sur la chaîne "centrale nucléaire" du diagramme construit selon la distance euclidienne pour une valeur de seuil de 0,47.

Cette souplesse, intéressante dans l'analyse de l'évolution des thématiques, est permise par la représentation des thèmes sous forme de distributions de probabilités. En effet, le calcul de similarité selon ses probabilités permet de rapprocher des thèmes dont le vocabulaire n'est pas identique. Toutefois, cette représentation présente des limites et entre parfois en contradiction avec l'interprétation humaine. La figure 5.15 illustre, par exemple, deux chaînes de thèmes qui ne sont pas reliées sur la visualisation alors que l'œil humain identifie une continuité sémantique entre les thèmes.

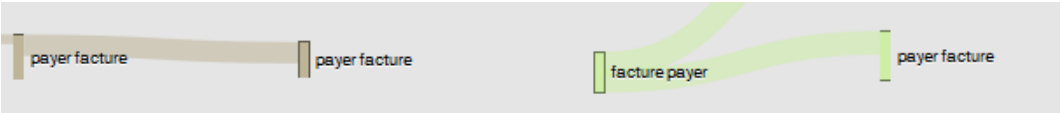


FIGURE 5.15 – Zoom sur la chaîne "payer facture" du diagramme construit selon la distance euclidienne pour une valeur de seuil de 0,47.

Cette différence de jugement vient du fait que les deux premiers thèmes sont, en réalité, portés uniquement par le terme "payer", le terme "facture" apparaissant bien les deux fois en deuxième position, mais avec une probabilité beaucoup plus faible. A l'inverse, dans les troisième et quatrième thèmes, les deux termes, "payer" et "facture", apparaissent avec des probabilités équivalentes. Selon le calcul de la distance euclidienne, qui prend en compte les écarts entre chaque probabilité, les deux premiers thèmes sont distincts du troisième, puisqu'ils ne contiennent pas (ou peu) le terme "facture", alors que le troisième et le quatrième sont proches (malgré l'inversion du terme dominant) car les probabilités des deux termes sont proches.

Bien évidemment, les deux thèmes centraux peuvent être reliés en modifiant le seuil pour relâcher les contraintes, mais dans ce cas, d'autres thèmes, qui semblent pourtant moins pertinents, se retrouvent également reliés.

Finalement, comme l'illustre la figure 5.16, lorsque la valeur du seuil est encore légèrement augmentée, le nombre de liens créés diminue la qualité de la visualisation. Pour une valeur de seuil de 0,57, les croisements de liens introduits ne permettent plus de suivre l'évolution des thématiques. Pour cette mesure, toute la variabilité, dans la répartition des liens, est alors contenue dans un intervalle d'étendue 0,1. Ce comportement, qui semblait, *a priori*, pouvoir permettre à un expert d'étudier avec finesse l'impact du choix de la valeur de seuil sur les chaînes de thématiques, s'avère, *a posteriori*, trop restrictif.

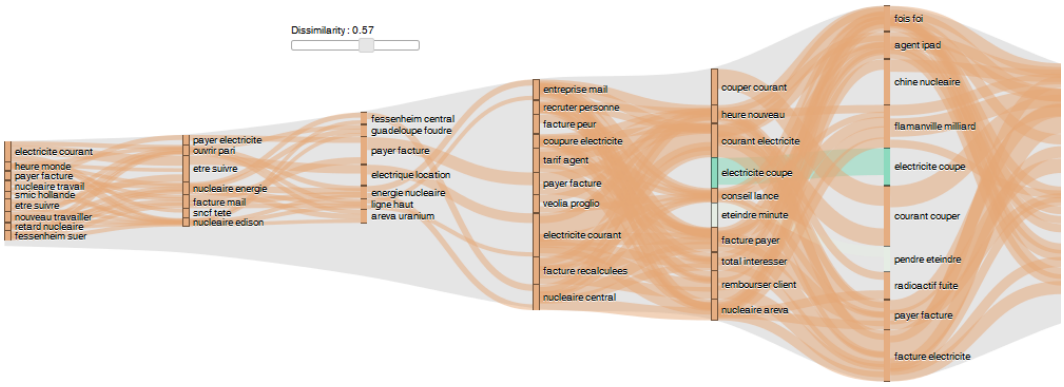


FIGURE 5.16 – Diagramme de Sankey illustrant les chaînes de thèmes LDA calculées selon la distance euclidienne pour une valeur de seuil de 0,57. Le nombre de liens créés est déjà trop important et les croisements de liens ne permettent plus de suivre l'évolution des thématiques.

4.3 Divergence de Kullback-Leibler

La dernière mesure de divergence étudiée ici est celle de Kullback-Leibler. Nous l'avons retenue, suite à l'analyse des distributions empiriques et des corrélations entre mesures, pour sa répartition plus uniforme des liens, qui fait d'elle un bon représentant de l'ensemble des mesures. Comme le montre les figures 5.17 et 5.18, l'étendue de son intervalle exploitable est, en effet, supérieure à celle des autres mesures, puisqu'il s'étend, au moins, entre 0,45 et 0,66.

FIGURE 5.17 – Diagramme de Sankey illustrant les chaînes de thèmes LDA calculées selon la divergence de Kullback-Leibler pour une valeur de seuil de 0,45. Des chaînes de thématiques commencent à apparaître.

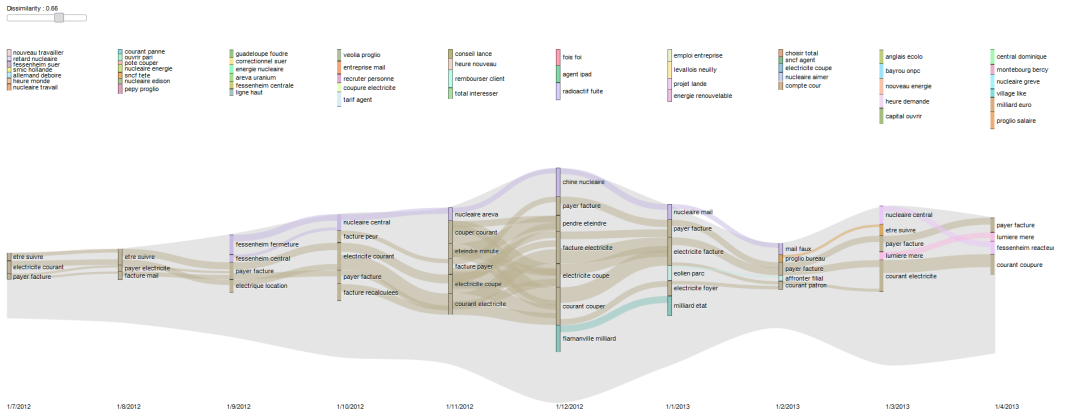


FIGURE 5.18 – Diagramme de Sankey illustrant les chaînes de thèmes LDA calculées selon la divergence de Kullback-Leibler pour une valeur de seuil de 0,66. Même si la chaîne principale contient beaucoup de croisements de liens, de grandes chaînes indépendantes continuent à émerger.

Pour une valeur de seuil de 0,45, plusieurs chaînes commencent à se former, et une chaîne composée de thèmes appartenant à cinq modèles consécutifs est déjà visible. Pour une valeur de seuil de 0,66, la chaîne principale, en marron, commence à agglomérer une grande partie des thèmes, mais les croisements de liens ne sont pas encore trop importants et le suivi de thématiques est encore possible. Pour cette valeur de seuil, d'autres chaînes indépendantes sont encore créées et permettent de suivre des concepts différents de ceux évoqués dans la chaîne principale. Comme le montre la figure 5.19, la chaîne "nucléaire", déjà mis en avant par le diagramme construit avec la distance euclidienne, s'étend ici sur six périodes consécutives et glisse vers des concepts de "nucléaire chinois" et de "mails frauduleux" liés à l'énergie nucléaire.

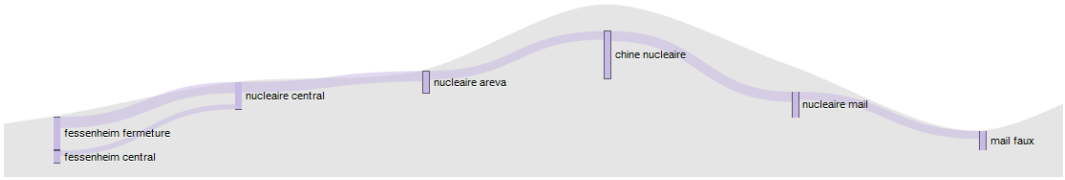


FIGURE 5.19 – Zoom sur la chaîne "nucléaire" du diagramme construit selon la divergence de Kullback-Leibler pour une valeur de seuil de 0,66.

Même si plusieurs chaînes ont fusionné pour former la chaîne principale, la forme générale de celle-ci ressemble, globalement, à celle de sous-chaînes thématiques évoluant en parallèle (voir figure 5.20). Néanmoins, les deux périodes centrales, 1/11/2012 et 1/12/2012, font exception, puisque pour elles, le nombre de croisements de liens est très important (voir figure 5.21). Comme le montre la figure 5.22, cette observation peut déjà être faite à partir du diagramme construit pour une valeur de seuil de 0,45. Ces nombreux croisements s'expliquent par le fait que, pour certaines périodes, les thèmes construits par un même modèle LDA ne sont pas suffisamment distincts les uns des autres. Pour ces modèles, le nombre de thèmes fixé n'apparaît pas comme étant optimal, et certaines thématiques sont fusionnées en un seul thème ou à l'inverse une thématiques est divisée en plusieurs thèmes.

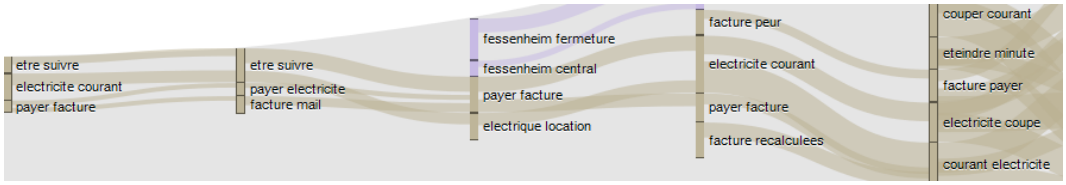


FIGURE 5.20 – Zoom sur la chaîne principale du diagramme construit selon la divergence de Kullback-Leibler pour une valeur de seuil de 0,66.

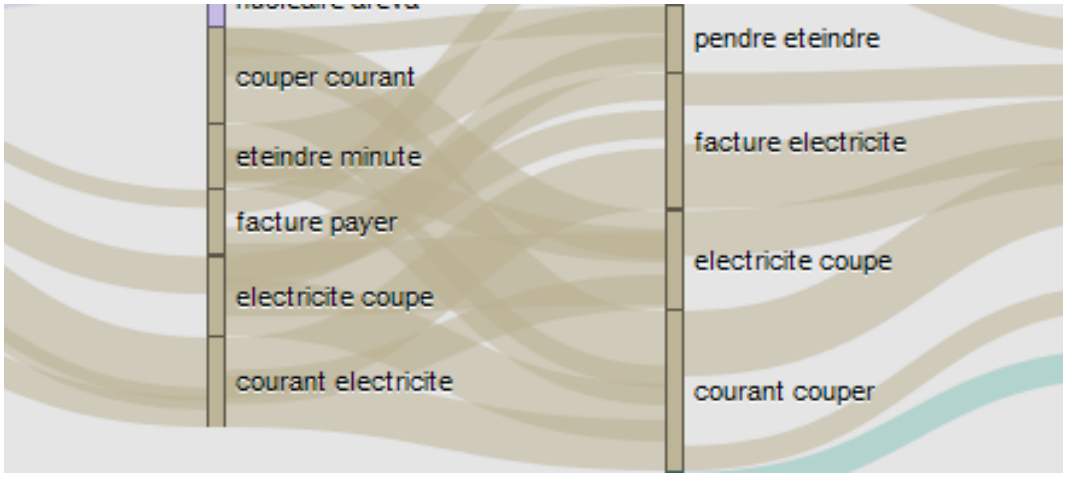


FIGURE 5.21 – Zoom sur les périodes 1/11/2012 et 1/12/2012 du diagramme de Sankey construit selon la divergence de Kullback-Leibler pour une valeur de seuil de 0,66.

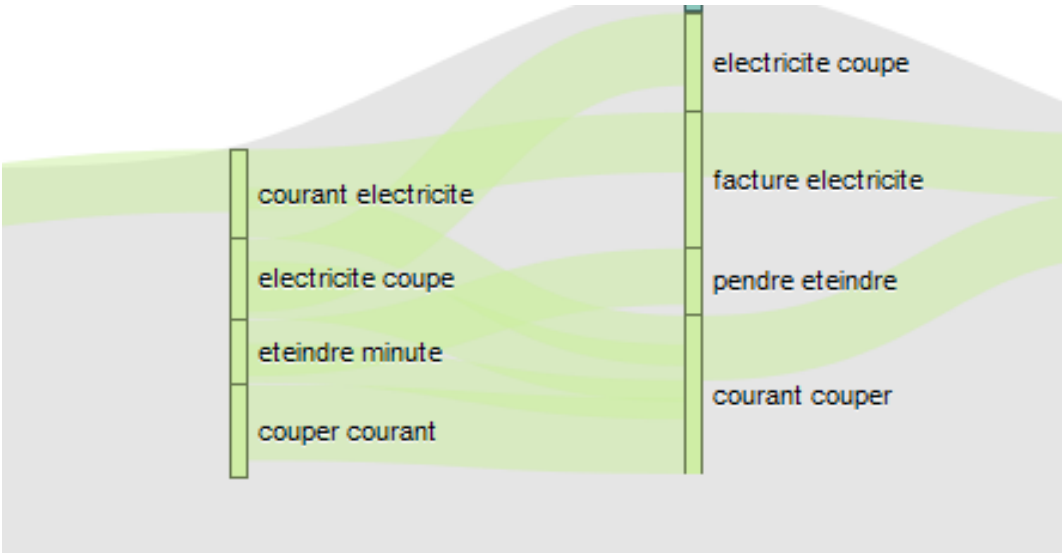


FIGURE 5.22 – Zoom sur les périodes 1/11/2012 et 1/12/2012 du diagramme de Sankey construit selon la divergence de Kullback-Leibler pour une valeur de seuil de 0,45.

Discussion. Ces différentes observations ont montré que les thèmes construits avec le modèle LDA ont un niveau de granularité permettant de détecter une continuité temporelle dans les thématiques abordées sur Twitter. Les nombreuses chaînes construites par l'approche Dynamic-LDA permettent de suivre la dynamique de ces thématiques et mettent en avant plusieurs types de comportements : suivi de thématiques et de leurs volumes, thématiques perdant de l'intérêt, détection de nouvelles thématiques, apparition de sous-thématiques et rapprochement entre thématiques.

Toutefois, et c'est un problème classique en apprentissage supervisé, le fait que le nombre de thèmes soit fixé a priori introduit un biais dans l'analyse. En effet, le modèle est parfois amené à diviser un thème volumineux en sous-thèmes, du fait de la contrainte imposée, alors que les données ne le justifient pas. A l'inverse, si les données contiennent plus de thèmes que ce que préconise le paramètre, le modèle peut être forcé de fusionner deux thèmes minoritaires. Enfin, les chaînes de thématiques et notamment les motifs de fusion et division sont impactés par cette limitation car ces motifs seraient amenés à disparaître si les thèmes étaient fusionnés.

Afin de déterminer l'impact du choix du nombre de topics sur la construction de chaînes de thématiques, nous proposons de modifier notre approche en remplaçant le modèle LDA par un modèle de thèmes non-paramétrique qui estime automatiquement le nombre de sujets à partir des données. En ajustant le nombre de thèmes aux données, la pertinence des chaînes et des motifs de fusion/division devrait être améliorée.

5 Intégration d'une approche non-paramétrique

Puisque le nombre de thèmes construits pour chaque période apparaît comme étant un facteur crucial de notre méthodologie, nous proposons d'intégrer un modèle de thème non-paramétrique : le modèle Hierachical Dirichlet Process (HDP) [TJBB06], qui estime automatiquement le nombre de thèmes en fonction des données. Nous rappelons, dans un premier temps, le processus générateur de ce modèle et une méthode pour estimer ses paramètres à partir d'un corpus de documents, puis nous décrivons, dans un second temps, les principes de son intégration dans notre méthodologie.

5.1 Hierachical Dirichlet Process

Le modèle Hierachical Dirichlet Process (HDP) s'applique aux problèmes où les données appartiennent à des groupes, ici les mots dans les documents, où chaque groupe est décrit par une distribution sur une variable catégorielle de cardinalité inconnue *a priori*, ici les thèmes, et où la variable catégorielle est partagée par les groupes (les composantes sont les mêmes, mais en proportions différentes). Son processus générateur décrit, comme LDA, un mécanisme pour expliquer l'apparition des mots dans les documents. Toutefois, il enrichit le modèle LDA en intégrant l'estimation du nombre de distributions de probabilités nécessaires à la description d'un corpus. La cardinalité de la variable latente (c'est-à-dire le nombre de thèmes) n'est pas bornée *a priori* (contrairement au modèle LDA) et grandit en moyenne comme le logarithme du nombre de données [TJBB06].

Plus formellement, HDP est un processus de Dirichlet hiérarchique à 2 niveaux [EW95] qui suppose l'existence d'une infinité de thèmes dont seulement un nombre fini est utilisé à chaque instant. Les documents sont représentés par une collection de processus de Dirichlet G_k (premier niveau) qui partagent une distribution de base commune G_0 qui est elle-même issue d'un processus de Dirichlet (deuxième niveau). Les thèmes correspondent aux atomes des processus de Dirichlet.

Puisque les processus de Dirichlet sont la généralisation au cas infini de la distribution de Dirichlet, le modèle HDP peut être vu comme la généralisation du modèle LDA permettant un nombre non borné de thèmes.

5.1.1 Processus générateur

Dans un premier temps, nous donnons une définition formelle des processus de Dirichlet afin d'introduire le processus générateur du modèle HDP. Dans un second temps, nous décrivons les métaphores du "Chinese Restaurant Process", qui permet de donner une définition constructive d'un processus de Dirichlet, et de la "Chinese Restaurant Franchise" (CRF) [TJBB06] qui généralise ce concept au cas infini et

permet de décrire le processus générateur de HDP.

Définition 15. Soit (Θ, B) un espace mesurable, H une mesure de probabilité sur cet espace et $\boldsymbol{\alpha} \in \mathbb{R}^+$, un *processus de Dirichlet* est la distribution d'une mesure de probabilité G sur (Θ, B) telle que, pour toute partition finie $(A_1, \ldots, A_n)$ de Θ, le vecteur $(G(A_1), \ldots, G(A_n))$ est distribué selon une distribution de Dirichlet (finie) : $(G(A_1), \ldots, G(A_n)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_n))$. On écrit : $G \sim DP(\alpha, H)$ et on appelle H la *mesure de base* et $\boldsymbol{\alpha}$ le *paramètre de concentration*.

Une telle mesure G existe et est discrète avec la probabilité 1 et peut se réécrire $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$, où les distributions ϕ_k sont appelés *les atomes*, δ_ϕ est la mesure de probabilité concentrée en ϕ et les poids $\pi_k \in \mathbb{R}$ somment à 1 [Fer73].

En utilisant ces notations, et étant donnés une mesure de base H (ici une distribution de Dirichlet sur le vocabulaire : $H = Dir(\boldsymbol{\beta})$) et deux paramètres de concentration α et γ, le processus générateur de HDP peut alors s'écrire :

- Tirer $G_0 \mid \gamma, H \sim DP(\gamma, H)$. On a alors : $G_0 = \sum_{k=1}^{\infty} \pi_{0k} \delta_{\phi_k}$ où les atomes ϕ_k représentent les thèmes et $\sum_{k=1}^{\infty} \pi_{0k} = 1$.
- Pour chaque document m :
  - Tirer une mesure $G_m \mid \alpha, G_0 \sim DP(\alpha, G_0)$. On a alors : $G_m = \sum_{k=1}^{\infty} \pi_{mk} \delta_{\phi_k}$ où G_m partage les mêmes atomes que G_0 avec des poids π_{mk} spécifiques à chaque document.
  - Pour chaque mot w_n de ce document :
    - Tirer un thème associé à ce mot $z_{m,n} \sim G_m$
    - Tirer un terme $t \sim Mult(\boldsymbol{\phi}_{\boldsymbol{z_{m,n}}})$

5.1.2 Estimation des paramètres

Ce formalisme présente toutefois la limitation de ne pas donner de définition constructive d'un processus de Dirichlet (même en supposant qu'un tel processus existe, il n'est pas possible de tirer d'échantillons selon cette loi) et ne permet pas, en l'état, de développer un algorithme d'estimation des paramètres. Plusieurs présentations équivalentes ont alors été proposées afin de générer des échantillons selon cette loi : la "stick-breaking construction" [Set91], le "Chinese Restaurant Process" (CRP) [BM73] et la limite infinie de modèles de mélange finis [TJBB06], et différentes méthodes d'inférence ont été proposées [KWT07, GGJ06, TGG07]. Nous détaillons ci-dessous la métaphore du CRP dans laquelle les échantillon sont représentés par les clients entrant dans un restaurant.

Ce restaurant est constitué d'une infinité (dénombrable) de tables, elles-mêmes de capacité infinie. Le premier client (1) à entrer dans le magasin s'assoie à une table libre, puis chaque nouveau client $(n+1)$ choisit de façon équiprobable de s'asseoir parmi les $n+1$ places suivantes : soit directement à gauche d'un des n clients déjà assis, soit à une nouvelle table.

Formellement, tirer une distribution $G \sim DP(\alpha, H)$ puis tirer $X_1, X_2, \ldots$ indépendamment selon G revient à tirer directement $X_1, X_2, \ldots$ selon H (mais de façon non indépendante) selon le procédé suivant [Fer73], appelé CRP :

- Tirer X_1 selon H
- Pour $n > 1$:
  - Tirer X_n selon H avec la probabilité $\frac{\alpha}{\alpha+n-1}$
  - Assigner X_n à une des valeurs précédentes X_j $(j < n)$ avec la probabilité $\frac{n_j}{\alpha+n-1}$ où n_j est le nombre d'observations ayant déjà été assignées à X_j

Remarque 10. *Cette métaphore fait déjà apparaître le caractère non borné du nombre de thèmes, qui sera directement lié au nombre de tables. Plus de nouveaux clients entreront dans le restaurant, plus il y aura de chances que de nouvelles tables soient utilisées. Par ailleurs, elle met également en avant l'effet de clustering du processus de Dirichlet pour lequel les valeurs ayant déjà été utilisées ont plus de chances d'être réutilisées.*

En utilisant cette représentation, le modèle du processus hiérarchique de Dirichlet peut être décrit par la métaphore de la franchise de restaurants chinois (CRF) dans laquelle les mots d'un document sont assignés à des "tables" selon un CRP propre à ce document et où les "tables" sont assignées à des "plats" selon un CRP commun à tout le corpus. Dans cette métaphore, les tables servent à faire le lien entre des thèmes communs au corpus (les plats) et les mots dans les documents. Les mots attachés au même plat (via leurs tables) sont tirés selon le même thème.

En formalisant cette métaphore, le processus générateur du Hierarchical Dirichlet Process (HDP) peut alors se réécrire :

- Pour chaque document m :
  - Pour chaque mot w_n de ce document :
    - Si $n = 1$, assigner w_n à la table 1
    - Sinon :
      - Assigner w_n à la table j avec la probabilité $\frac{n_j^{(t)}}{\sum_{j'} n_{j'}^{(t)} + \alpha}$ et associer à w_n le thème $z_{m,n} = k_j$ de la table j
      - Assigner w_n à une nouvelle table j_{new} avec la probabilité $\frac{\alpha}{\sum_{j'} n_{j'}^{(t)} + \gamma}$ et associer à w_n un thème existant k avec la probabilité $\frac{n_k}{\sum_{k'} n_{k'} + \alpha}$ ou un nouveau thème avec la probabilité $\frac{\gamma}{\sum_{k'} n_{k'} + \gamma}$

5.2 Dynamic-HDP

Nous décrivons maintenant comment le modèle HDP peut être aisément intégré à notre outil de visualisation afin d'explorer des chaînes de thématiques construites par plusieurs modèles successifs.

5.2.1 Principes

L'intégration d'un autre modèle ne présente pas de difficulté majeure à partir du moment où celui-ci décrit les documents par des distributions de probabilités sur les thèmes qui sont eux-mêmes décrits par des distributions de probabilités sur les termes du vocabulaire. Le principe général reste le même, la période de temps $\mathcal{T}$ est découpée en sous-périodes $\mathcal{T}_1, \ldots, \mathcal{T}_n$ et pour chaque période $\mathcal{T}_i$, un modèle HDP est appris. La seule distinction est que, dans ce contexte, le nombre de thèmes de chaque modèle dépend de la période considérée : $\mathcal{M}_i = (\mathcal{D}_i, \mathcal{V}_i, K_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$. La phase d'harmonisation des vocabulaires est la même et le calcul de similarité entre thème se fait par mesure de divergence entre leurs distributions de probabilités.

Pour cette implémentation, nous avons utilisé la librairie Java HDP développée par Armin et al.[4] qui intègre l'algorithme d'échantillonnage de Gibbs basé sur le CRP présenté dans [TJ10]. Les valeurs par défaut des paramètres ont été conservées, les hyper-paramètres γ et α ont été fixés à 1,5 et 1, respectivement et le nombre d'itérations a été fixé à 2000. Le nombre de thèmes estimés par chaque modèle HDP

[4] github.com/arnim/HDP

est représenté sur la figure 5.23. Comme on peut le voir en mettant cette information en parallèle du nombre de documents publiés chaque mois (voir figure 4.10), il semble que, contrairement à ce que prévoit le modèle, le nombre de thèmes n'augmente pas nécessairement avec le nombre de documents. Le comportement inverse se produit même. Ce phénomène peut s'expliquer par le système de retweets de Twitter qui tend à concentrer l'information sur un seul sujet. En effet, les tweets supplémentaires traduisent uniquement une thématique particulièrement populaire et n'entraînent pas l'ajout de nouveaux thèmes.

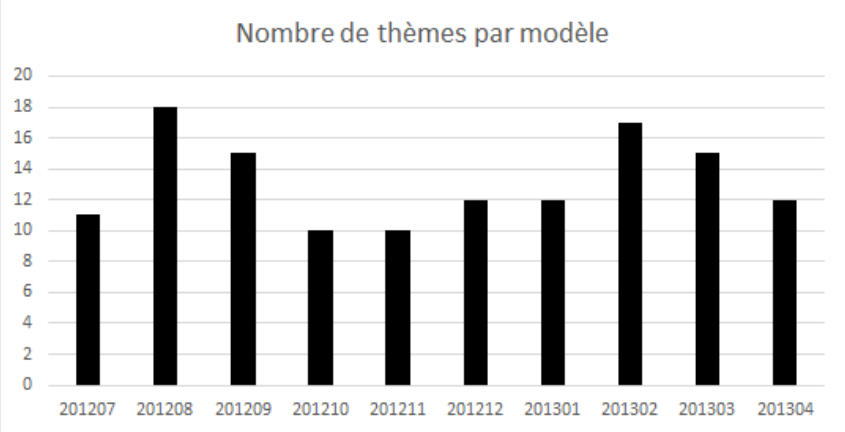


FIGURE 5.23 – Histogramme présentant le nombre de thèmes calculés par chaque modèle HDP sur les 10 périodes d'un mois.

Résultats. Nous présentons sur la figure 5.24 une capture d'écran de l'outil de fouille exploratoire présenté à l'analyste. Il apparaît que certains thèmes, qui avaient été artificiellement séparés par le modèle LDA à cause du paramètre fixé K, se retrouvent ici regroupés pour former des thèmes aux volumes plus importants. Comme le montre la figure 5.25, les chaînes construites avec ces thèmes présentent moins d'interconnexions, ce qui tend à montrer que les thèmes appris sont mieux séparés les uns des autres. De fait, il semble que l'utilisation du modèle HDP permet d'estimer des thèmes plus discriminants.

Néanmoins, une nouvelle limitation semble mise en avant par cette visualisation : en effet, il apparaît que chaque modèle introduit plusieurs thèmes dont les volumes, en termes de documents représentés, sont très faibles. Ces thèmes, qui ne participent que très peu à l'explication du flux total, diminuent la qualité de la visualisation en augmentant artificiellement le nombre de nœuds affichés.

En première approximation, une solution possible pour clarifier la visualisation consiste à exploiter les interactions permises par notre outil de fouille exploratoire pour déplacer manuellement les noœuds les plus volumineux. Ainsi, comme le montre la figure 5.26, il est possible de sélectionner les thèmes les plus importants et d'expliquer la quasi-totalité de l'information contenue dans le flux en juxtaposant

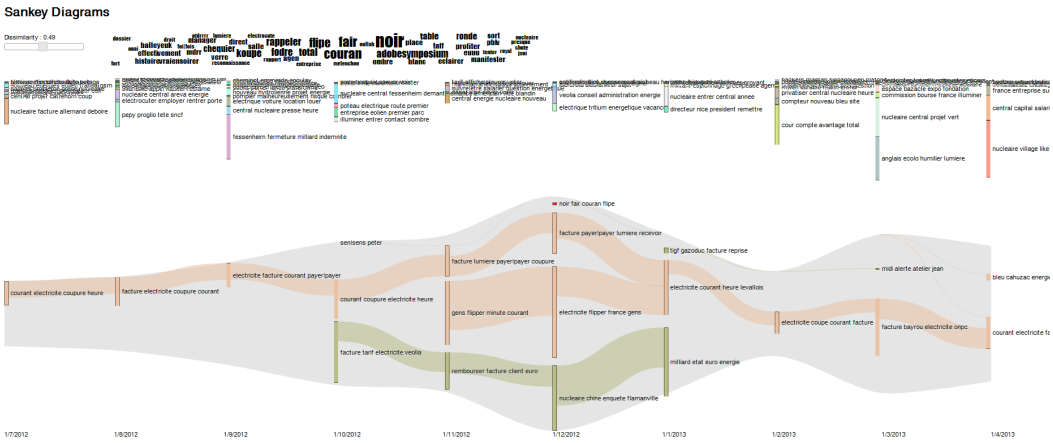


FIGURE 5.24 – Diagramme de Sankey illustrant les chaînes de thèmes HDP calculées selon la divergence de Kullback-Leibler pour une valeur de seuil de 0,47.

les chaînes et les thèmes isolés. Cette représentation permet d'identifier un nombre raisonnable de thèmes à explorer individuellement.

Cette solution n'est cependant pas idéale car elle limite fortement l'intérêt d'utiliser un modèle non-paramétrique. Puisque le nombre de thèmes construits dépend indirectement des paramètres γ et α, une analyse de leurs influences pourrait permettre de corriger ce problème dès la construction des modèles.

Discussion. Comme les thèmes construits grâce au modèle LDA, les thèmes construits grâce à son extension non-paramétrique HDP permettent de mettre en avant des chaînes de thématiques dans le contexte de forte nouveauté des données Twitter. En estimant automatiquement le nombre de thèmes pour chaque période, HDP semble plus à même de produire des thèmes spécifiques, représentatifs de chaque période.

La question de la validation d'un processus de fouille de données est une question délicate qui est encore largement débattue à la fois dans les communautés de visualisation et d'interface homme machine (IHM) et dans la communauté de fouille de données. Dans le cadre d'un processus de fouille, les tâches sont des tâches cognitivement complexes non préalablement bien définies et le cadre classique de l'évaluation en IHM est difficile à appliquer. Nous nous sommes restreints ici à des retours d'expérience d'un expert. Ces retours nous ont permis d'affiner à différents reprises notre représentation initiale, qui est considérée aujourd'hui comme une preuve de concept. Le prototype a été présenté récemment à la direction du domaine DACC et son devenir y est en réflexion, à la fois pour son intégration à la chaîne de traitement "Text Mining" et pour son évaluation sur d'autres jeux de données.

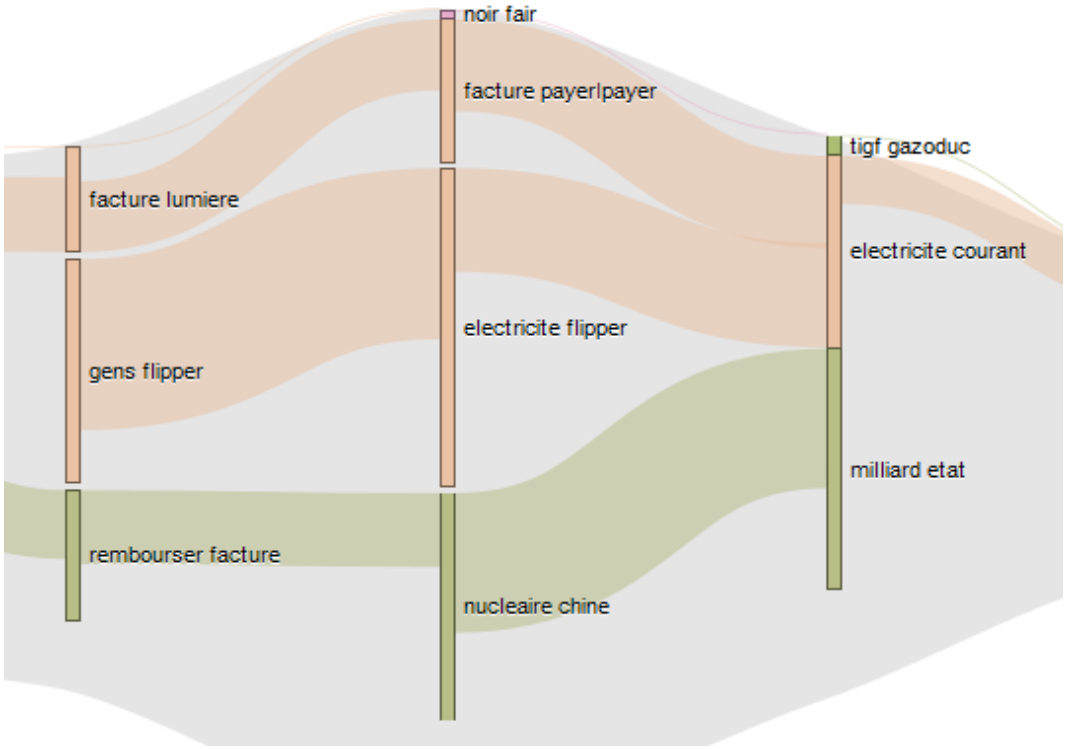


FIGURE 5.25 – Zoom sur le diagramme de Sankey construit avec la méthodologie Dynamic-HDP illustrant la construction de chaînes distinctes avec peu de croisements de liens.

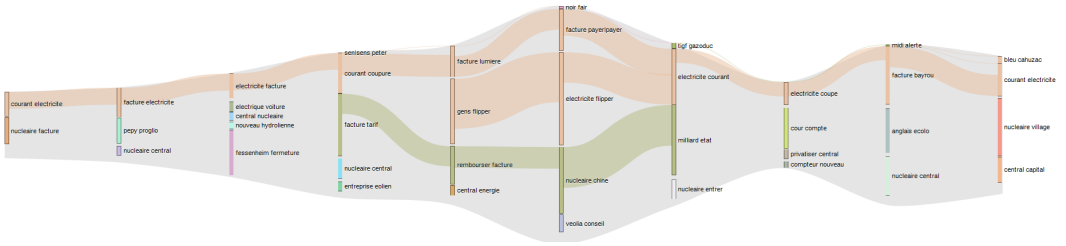


FIGURE 5.26 – Diagramme de Sankey illustrant les chaînes de thèmes HDP ainsi que les principaux thèmes de chaque période. La juxtaposition des chaînes et des thèmes isolés permet d'expliquer l'intégralité du flux avec peu de thèmes.

6 Conclusion

Dans ce chapitre, nous avons débuté par une présentation de quelques visualisations classiques adaptées aux données textuelles et mettant l'accent sur la représentation de la dimension temporelle. Parmi ces visualisations, les diagrammes de Sankey nous ont semblé les plus adaptés à l'analyse de l'évolution temporelle de flux de données pouvant fusionner ou se séparer. En effet, cette description permet, à la fois, d'illustrer l'évolution temporelle des thématiques et leurs importances relatives en assignant leurs poids aux volumes des flux, mais également de visualiser des motifs de fusion et de division entre flux.

Nous avons ensuite présenter l'outil de fouille exploratoire que nous avons dé-

veloppé au cours de cette thèse et qui permet à un analyste d'explorer les chaînes de thématiques construites avec notre méthodologie Dynamic-LDA. Nous avons décrit l'interface intégrant, notamment, un diagramme de Sankey et un nuage de mots, et nous avons précisé les interactions proposées. Suite à cela, nous avons présenté les résultats obtenus avec les trois mesures de divergence retenues afin de les évaluer, en partenariat avec un expert de l'entreprise, en termes de visualisations produites. Les premiers résultats ont montré que le niveau de granularité des thèmes construits avec le modèle LDA est adapté à la détection de chaînes de thématiques, mais que le choix du nombre de thèmes à chaque période est un paramètre crucial de notre visualisation.

Afin d'évaluer l'impact de ce paramètre sur les visualisations produites, nous avons intégré le modèle non-paramétrique HDP à la place du modèle LDA. Nous avons rappelé la définition du modèle HDP et de son processus générateur qui permet d'estimer automatiquement le nombre de thèmes à partir des données. Par la suite, nous avons montré comment la méthodologie Dynamic-HDP pouvait s'intégrer à notre outil de visualisation et nous avons présenté plusieurs visualisations obtenues avec cette approche. Les résultats ont montré l'intérêt de cette modification qui permet d'explorer des chaînes de thématiques plus significatives présentant des motifs de fusion et division plus pertinents.

6

Conclusion et perspectives

Either mathematics is too big for the human mind, or the human mind is more than a machine.

—Kurt Gödel

Sommaire

1 Conclusion

Dans cette thèse, nous avons étudié, sans connaissances *a priori*, un corpus inédit de textes extraits de Twitter et ciblant l'entreprise EDF. Malgré l'accès à ses données proposé par Twitter, les contraintes techniques et méthodologiques limitent fortement l'établissement de corpus de référence. La collecte de gros volumes de données reste encore très contrainte, et la variété des problématiques soulevées par ce nouveau média restreint largement l'établissement de valeurs de vérités. Les thématiques extraites relèvent en fait de plusieurs concepts, détection d'événements (du monde réel), détection de nouvelles tendances et suivi de tendances, et sont modélisées à différents niveaux de granularité, aussi bien sémantiques que temporels. La finalité des méthodes proposées couvrent un large éventail, allant de la recherche d'information à l'extraction de connaissances, en passant par les systèmes d'alerte en temps réel. Dans ce cadre, il est difficile de comparer les résultats des méthodes proposées puisqu'il n'existe pas de mesure de qualité absolue. L'évaluation des thématiques se fait parfois au travers de tâches secondaires (par exemple en mesurant leur apport à une tâche de clustering des documents associés) qui ne présagent pas nécessairement de leur capacité à produire de l'information utile. Les thèmes construits par les meilleurs modèles selon ce critère ne seront pas forcement les plus exploitables. Par ailleurs, la nature fortement bruitée et le contexte de forte nouveauté qui caractérise Twitter contribuent à la détection de nombreuses thématiques dont la pertinence est délicate à évaluer sans l'appui d'un analyste.

En outre, le corpus étudié dans cette thèse présente la particularité de s'étendre sur une période particulièrement longue au regard de ce qui peut se faire classiquement dans la littérature. Même si les volumes de données concernant l'entreprise restent raisonnables, les besoins d'analyse s'inscrivent dans une démarche de suivi de la gestion de la relation client sur le long terme et vont au delà de la description du corpus à un instant donné. Dans ce contexte, l'analyse de la dimension temporelle s'est révélée d'une importance cruciale. En effet, la diversité des thématiques abordées ainsi que la vitesse à laquelle elles se succèdent limitent la mise en relation de concepts immédiatement dépassés. L'opposition entre contexte de forte nouveauté et suivi de l'évolution des thématiques nous a alors incité à mettre en place des outils d'analyse mettant en valeur la continuité temporelle.

Afin d'étudier de manière itérative un jeu de données original, nous nous sommes placés dans une logique de fouille exploratoire intégrant un expert de la gestion de la relation client et du traitement du langage naturel au processus de fouille. Nos premières analyses ont reposé initialement sur deux approches classiques de la fouille de données textuelles : la classification non-supervisée (clustering) et l'extraction de motifs fréquents. Elles nous ont amenés par la suite à proposer une méthodologie originale analysant la coévolution temporelle des termes. Ces approches, qui

analysent les tweets à la fois à l'échelle des documents et à l'échelle des termes individuels, ont confirmé deux particularités des données Twitter : premièrement, le système de retweets, qui permet à chaque utilisateur de diffuser à l'ensemble de ses abonnés une information qui lui semble pertinente, favorise l'émergence de thématiques saillantes atteignant leur pic de popularité très rapidement. Deuxièmement, l'architecture de Twitter, qui permet d'indexer les tweets et de les rendre disponibles dans des délais s'approchant du temps réel, a développé une culture de l'instantanéité chez les utilisateurs du service. Les sujets populaires s'enchaînent à un rythme journalier ne laissant pas de place aux échanges conversationnels. Twitter apparaît alors davantage comme une source d'information en continu que comme un réseau social.

De manière à proposer à l'analyste une information plus riche qu'une simple succession de pics, il est nécessaire de lui fournir des éléments de contexte en modélisant des thématiques au niveau de granularité adéquate pour permettre un suivi temporel. Pour ce faire, nous proposons d'exploiter la structure latente extraite d'un corpus de documents par les modèles de thèmes ("topic models") en développant une méthodologie permettant d'adapter les modèles d'analyse statique aux corpus de textes dynamiques. Ces modèles probabilistes ont prouvé leur intérêt dans leur application à la recherche d'information et, comme en témoigne l'imposante bibliographie maintenue à jour par l'université de Cornell[1], appartiennent désormais à un domaine de recherche prolifique. Dans cette thèse, nous sommes concentrés sur les variantes les plus élémentaires de ces modèles, qui apportent le moins de connaissances a priori sur les données tout en permettant d'enrichir l'analyse textuelle en détectant des éléments de synonymie et de polysémie.

Toutefois, une des hypothèse fondamentale de ces modèles est que les documents sont supposés interchangeables au sein du corpus. En effet, l'hypothèse d'indépendance des variables nécessaire à l'estimation des paramètres impose, à la fois, que les documents soient interchangeables au sein du corpus, et que les mots soient interchangeables au sein des documents. Les thématiques extraites sont alors représentatives du corpus dans son ensemble, et ces modèles ne sont pas directement applicable pour l'étude de flux de textes dynamiques, et à plus forte raison dans un contexte de forte nouveauté, où l'hypothèse de stabilité des thématiques n'est pas raisonnable.

L'approche proposée dans cette thèse consiste à diviser la période de temps en sous-intervalles, et à entraîner un modèle sur chacun d'eux. Ainsi les documents ne sont pas supposés interchangeables sur l'ensemble du flux, mais seulement au sein d'une sous-période. La similarité entre les thèmes extraits par ces modèles, définis en termes de distributions de probabilités, est ensuite calculée par une mesure de divergence afin de construire des chaînes sémantiques dans lesquelles les thèmes

[1] mimno.infosci.cornell.edu/topics.html

peuvent fusionner ou se diviser. Dans un premier temps, nous avons utilisé le modèle de Latent Dirichlet Allocation (LDA) qui est l'un des modèles de thèmes les plus populaires, et nous avons montré dans un second temps comment notre approche, suffisamment générale, permettait d'intégrer facilement des modèles plus complexes tel que le processus de Dirichlet Hiérarchique ("Hierarchical Dirichlet Process" (HDP)).

En complément, une analyse quantitative de sept mesures de divergence nous a permis d'identifier une mesure particulièrement adaptée à la comparaison de thèmes successifs construits par un modèle de thème. L'analyse de la corrélation entre ces différentes mesures a en effet montré que, même si les distances absolues entre thèmes étaient potentiellement différentes et reparties de manière inégale sur l'espace des valeurs, les proximités relatives entre thèmes étaient respectées. Cette corrélation forte signifie, du point de vue de notre méthodologie, que les liens créés entre thématiques, selon différentes mesures, se font dans le même ordre et conduisent à la création de chaînes similaires. Il est alors possible de sélectionner la mesure la plus adaptée à la comparaison de thèmes en analysant les distributions empiriques de leurs valeurs calculées sur nos données. La majorité des mesures étudiées concentrent leurs valeurs sur un intervalle restreint qui limite leur pouvoir de discrimination entre thèmes, alors que la divergence de Kullback-Leibler, très répandue dans la littérature, présente une répartition plus homogène de ses valeurs. Cette divergence apparaît alors comme le représentant des mesures étudiées le plus à même d'exposer les relations complexes entre thèmes.

Afin de permettre à l'analyste d'explorer les chaînes de relations entre thèmes construites par notre méthodologie, nous avons développé un prototype de visualisation basé sur les graphes de Sankey. Ce paradigme de visualisation, inspiré de l'analyse des flux énergétique, étend les classiques aires empilées pour permettre la visualisation de motifs de fusion et de division des flux. Une expérimentation conduite avec un expert de l'entreprise a permis de détecter des thématiques récurrentes et des glissements de concepts dans un contexte de forte nouveauté, et a également mis en avant l'importance du choix du niveau de granularité dans l'analyse des thématiques. En effet, les nombreuses imbrications entre thèmes calculés pour certaines périodes témoignent d'un découpage arbitraire de la structure des documents. Le nombre de thèmes imposés à chaque modèle n'est pas toujours adapté et conduit parfois à la séparation, forcée, des thèmes majoritaires, ou au regroupement de thèmes minoritaires. Cette observation nous a poussés à intégrer un modèle non paramétrique (HDP) qui estime automatiquement le nombre de thèmes à partir des données. Nos premiers résultats avec ce modèle sont prometteurs et ont confirmé le découpage arbitraire des thèmes construits avec LDA. Ils ouvrent toutefois la voie à des expérimentations plus poussées afin de s'assurer de sa robustesse, notamment en étudiant l'influence des hyper-paramètres sur la granularité des thèmes estimés.

2 Perspectives

Les recherches présentées dans ce manuscrit s'inscrivent dans le domaine de la fouille exploratoire de données textuelles qui, bien que les problématiques sous-jacentes ne soient pas nouvelles, a connu un nouvel essor ces dernières années, porté à la fois par la multiplication des sources de données disponibles, et les avancées technologiques permettant de traiter des volumes de données toujours plus importants. Ces nouveaux enjeux confrontent les méthodes existantes, non seulement au problème du passage à l'échelle, mais également au contexte de forte nouveauté favorisé par les besoins d'instantanéité des médias sociaux et au bruit sémantique introduit par le paradigme du consommateur/producteur.

Dans ce contexte, nous avons étudié plusieurs approches de modélisation de thématiques permettant de synthétiser un flux de textes, et les résultats intermédiaires nous ont amenés à proposer une variante dynamique des modèles de thèmes. A en croire le nombre de modèles proposés, ce domaine de recherche est très actif et les variantes sont nombreuses : intégration de différentes distributions *a priori* [CBA12], de différents processus générateur [GSBT04, EFL04, KPC14] ou de structures plus complexes [LBM12, TJ10].

Dans un futur proche nous privilégions parmi les perspectives variées deux directions de recherche : sur le plan théorique il s'agit d'intégrer d'autres modèles d'extraction de thèmes qui intègrent de la connaissance sur les données ; et concernant la fouille visuelle, il s'agit de compléter les fonctionnalités du prototype développé en suivant notamment les remarques que nous avons recueillies lors des premiers tests.

Vers l'intégration d'autres modèles de thèmes Le modèle de LDA est en quelque sorte le modèle apportant le moins d'information *a priori* sur les données. Le choix de la distribution de Dirichlet comme distribution *a priori* n'est justifié que par des raisons opérationnelles, puisqu'elle est la distribution conjuguée de la distribution de Bernoulli multivariée et simplifie fortement les calculs. En outre, le choix, souvent fait dans la littérature, d'utiliser une version symétrique de la distribution de Dirichlet confirme la volonté de ce modèle de n'introduire aucune connaissance sur les thèmes *a priori*. Il n'est pas étonnant alors que plusieurs variantes aient été proposées afin d'intégrer des connaissances sur les données. Ces modèles, qui comme LDA, décrivent la structure sous-jacente d'un corpus de documents à l'aide de distributions de probabilités peuvent être facilement intégrés dans notre approche. Ils pourraient, par exemple, permettre de prendre en compte des connaissances sur les auteurs des documents [RZGSS04], sur la structure de leur réseau social[LB12] ou sur la longue traîne des distributions linguistiques [Teh06]. Par ailleurs, puisque le niveau de granularité des thématiques extraites apparaît comme

un enjeu majeur de la modélisation de thématiques, plusieurs modèles ont été développés afin d'extraire des hiérarchies de thèmes permettant de décrire un corpus sous forme d'un arbre et ses documents sous forme de branches composés de thèmes de différents niveaux de granularité [WLD+14]. Finalement, plusieurs approches ont été proposées afin d'introduire la dimension temporelle dès la conception du modèle [ABD08, WBH12, WM06] et l'intégration de ces modèles dynamiques pourrait permettre une analyse plus approfondie de l'évolution temporelle des thématiques abordées dans un flux de textes.

Vers le déploiement d'un outil interactif L'outil de visualisation interactive développé au cours de cette thèse a servi de preuve de concept afin de valider l'intérêt des modèles de thèmes pour la modélisation de chaînes de thématiques dans un contexte de forte nouveauté. Toutefois, afin de le rendre totalement opérationnel, notamment en vue de son déploiement dans les directions métiers, plusieurs améliorations devront être apportées, aussi bien au niveau de la visualisation, qu'au niveau des interactions offertes à l'analyste.

Tout d'abord, puisque ce prototype a été initialement développé pour confirmer l'existence de chaînes de thèmes, la version actuelle de notre outil ne permet pas d'accéder aux documents à partir du diagramme de Sankey. La nécessité de faire le lien avec les documents ne faisait pas partie des besoins initiaux et s'est pourtant rapidement fait sentir lors des expérimentations. Comme le fait remarquer Van Wijk [Van05], "nous ne savons pas quelle information les données contiennent, et nous en dressons donc un portrait afin d'en obtenir un aperçu". L'analyse des chaînes créées et, en particulier, le besoin d'explications concernant les thèmes qui les composent ont montré la nécessité de revenir aux documents.

Par ailleurs, les interactions proposées à l'analyste ont été développées afin de lui permettre d'explorer les chaînes construites, pour un jeu de paramètres fixé, au travers de mécanismes de zoom et de glisser/déplacer. Un curseur permettant de sélectionner un seuil de similarité entre thématiques a été intégré afin d'étudier des relations plus ou moins fortes et plus ou moins nombreuses entre thèmes et ainsi valider l'intérêt de notre méthodologie dans une tâche de suivi de thématiques. Toutefois, il n'était pas prévu, initialement, de permettre à l'analyste de modifier les paramètres intrinsèques de la modélisation (comme la taille et le nombre de périodes, le nombre de thèmes de chaque modèle et les valeurs des hyper-paramètres). Ce besoin a rapidement émergé, mais ces interactions, plus riches, nécessitent des temps de calcul plus longs, puisqu'elles impliquent une mise à jour de l'ensemble des modèles, et un compromis entre fluidité de la visualisation et puissance de calcul devra être trouvé.

Bibliographie

[AAN+13] Jitendra Ajmera, Hyung-iL Ahn, Meena Nagarajan, Ashish Verma, Danish Contractor, Stephen Dill, and Matthew Denesuk. A CRM System for Social Media : Challenges and Experiences. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW'13, pages 49–58, 2013. 17

[ABD08] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. Online LDA : adaptive topic models for mining text streams with applications to topic detection and tracking. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 3–12. IEEE, 2008. 138

[ACD+98] James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study final report. 1998. 27

[AIS93] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993. 58

[ALJ00] James Allan, Victor Lavrenko, and Hubert Jin. First Story Detection in TDT is Hard. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, CIKM'00, pages 374–381, 2000. 37

[AOGS13] Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, and Frederic Stahl. TRCM : a Methodology for Temporal Analysis of Evolving Concepts in Twitter. In *Artificial Intelligence and Soft Computing*, pages 135–145. Springer, 2013. 42, 46, 53

[APL98] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval*, SIGIR'98, pages 37–45, 1998. 38

[APM+13] L.M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes. Sensing Trending Topics in Twitter. *IEEE Transactions on Multimedia*, 15(6) :1268–1282, Oct 2013. 49

[AS+94] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proceedings of the 20th internationnal confeference on very large data bases (VLDB)*, volume 1215, pages 487–499, 1994. 59

[AS12] Charu C Aggarwal and Karthik Subbian. Event Detection in Social Streams. In *SDM*, volume 12, pages 624–635. SIAM, 2012. 50

[BER67] Jaques BERTIN. Semiologie Graphique : Les diagrames-Les Réseaux-Les Cartes. *Ed. Gauthier-Villars, Paris*, 2, 1967. 22

[BGL+12] Michael Busch, Krishna Gade, Brian Larson, Patrick Lok, Samuel Luckenbill, and Jimmy Lin. Earlybird : Real-Time Search at Twitter. In *Proceedings of the 28th International Conference on Data Engineering*, ICDE'12, pages 1360–1369, 2012. 20

[BHJ09] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi : An Open Source Software for Exploring and Manipulating Networks, 2009. 87

[BJGJ01] Ziv Bar-Joseph, David K Gifford, and Tommi S Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl 1) :S22–S29, 2001. 63

[BL09] David M Blei and John D Lafferty. Topic models. *Text Mining : Classification, Clustering, and Applications*, 10 :71, 2009. 27

[Ble12] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4) :77–84, 2012. 11, 75

[BM73] David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, pages 353–355, 1973. 125

[BNG11] Hila Becker, Mor Naaman, and Luis Gravano. Beyond Trending Topics : Real-World Event Identification on Twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM'11, pages 438–441, 2011. 32, 33

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3 :993–1022, 2003. 21, 34, 37, 72, 74, 76, 78, 94, 104

[BS15] Alina Beck and Philippe Suignard. Procédé d'identification d'une donnée comme pertinente ou hors sujet, FR3016712 (A1). 24-07-2015. 52

[BSH+10] Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H. Chi. Eddi : Interactive Topic-based Browsing of Social Status Streams. In *Proceedings of the 23Nd Annual Symposium on User Interface Software and Technology*, UIST'10, pages 303–312, 2010. 29, 33

[CA13] Freddy Chong Tat Chua and Sitaram Asur. Automatic Summarization of Events From Social Media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media*, ICWSM'13, 2013. 34, 37, 49, 51, 53

[CALC13] Yan Chen, Hadi Amiri, Zhoujun Li, and Tat-Seng Chua. Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th International Conference on Research and Development in Information Retrieval*, SIGIR'13, pages 43–52, 2013. 49, 51

[CBA12] Karla L. Caballero, Joel Barajas, and Ram Akella. The generalized dirichlet distribution in enhanced topic detection. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 773–782. ACM, 2012. 137

[CDS10] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD'10, pages 24–33, 2010. 51

[CG95] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4) :327–335, 1995. 79

[Che78] Emile Cheysson. Les méthodes de statistique graphique à l'exposition universelle de 1878. *Journal de la Société de Statistique de Paris*, 12 :323–333, 1878. 22

[CWT13] Simon Carter, Wouter Weerkamp, and Manos Tsagkias. Microblog language identification : Overcoming the limitations of short, unedited

and idiomatic text. *Language Resources and Evaluation*, 47(1) :195–215, 2013. 19

[CYL+99] Jaime Carbonell, Yiming Yang, John Lafferty, Ralf D Brown, Tom Pierce, and Xin Liu. CMU report on TDT-2 : Segmentation, detection and tracking. In *Proceedings of the DARPA Broadcast News Workshop*, pages 117–120, 1999. 42

[CZL+12] Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. Discover Breaking Events with Popular Hashtags in Twitter. In *Proceedings of the 21st International Conference on Information and Knowledge Management*, CIKM'12, pages 1794–1798, 2012. 28

[CZYZ12] Steven P Crain, Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. Dimensionality reduction and topic modeling : From latent semantic indexing to latent dirichlet allocation and beyond. In *Mining Text Data*, pages 129–161. Springer, 2012. 27

[D+92] Ingrid Daubechies et al. *Ten lectures on wavelets*, volume 61. SIAM, 1992. 31

[Dar11] William M Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics : Human language technologies*, pages 642–647, 2011. 78

[DDL+90] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6) :391–407, 1990. 72

[DHS12] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012. 32

[DJY11] Anish Das Sarma, Alpa Jain, and Cong Yu. Dynamic relationship and event discovery. In *Proceedings of the Fourth International Conference on Web Search and Data Mining*, WSDM'11, pages 207–216, 2011. 27

[DNKS10] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. Diamonds in the rough : Social media visual analytics for journalistic inquiry. In *Proceedings of the Conference on Visual Analytics Science and Technology*, VAST'10, pages 115–122, 2010. 35, 37, 49, 51, 53

[DWS+12] Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X Zhou. Leadline : Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the Conference on Visual Analytics Science and Technology*, VAST'12, pages 93–102, 2012. 12, 27, 34, 37, 49, 51, 111, 112

[EFL04] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1) :5220–5227, 2004. 137

[EW95] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430) :577–588, 1995. 124

[FDGM99] Jon Fiscus, George Doddington, John Garofolo, and Alvin Martin. NIST's 1998 Topic Detection and Tracking evaluation (TDT2). In *Proceedings of the 1999 DARPA Broadcast News Workshop*, pages 19–24, 1999. 27

[Fer73] Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973. 125, 126

[FFP05] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005. 74

[FGK+05] Patrick Flaherty, Guri Giaever, Jochen Kumm, Michael I Jordan, and Adam P Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21(15) :3286–3293, 2005. 74

[FHL+03] Dieter Fox, Jeffrey Hightower, Lin Liao, Dirk Schulz, and Gaetano Borriello. Bayesian filters for location estimation. *IEEE Pervasive Computing*, 2(3) :24–33, 2003. 38

[FYYL05] G.P.C. Fung, J.X. Yu, P.S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB'05, pages 181–192, 2005. 31

[GGJ06] Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and*

the 44th annual meeting of the Association for Computational Linguistics, pages 673–680. Association for Computational Linguistics, 2006. 125

[GGZ+13] Hansu Gu, Mike Gartrell, Liang Zhang, Qin Lv, and Dirk Grunwald. AnchorMF : Towards Effective Event Context Identification. In *Proceedings of the 22Nd International Conference on Information & Knowledge Management*, CIKM'13, pages 629–638, 2013. 51, 53

[GHN12] Emden R. Gansner, Yifan Hu, and Stephen North. Visualizing Streaming Text Data with Dynamic Graphs and Maps. In *Proceedings of the 20th International Conference on Graph Drawing*, GD'12, pages 439–450. Springer-Verlag, 2012. 12, 43, 46, 49, 51, 53, 110, 111

[GL12] Pritam Gundecha and Huan Liu. Mining social media : a brief introduction. *Tutorials in Operations Research*, 1(4), 2012. 18

[GLD12] Wei Gao, Peng Li, and Kareem Darwish. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st International Conference on Information and Knowledge Management*, CIKM'12, pages 1173–1182, 2012. 36, 37, 54

[Gri02] Tom Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. 2002. 78

[GRS96] Walter R Gilks, Sylvia Richardson, and David J Spiegelhalter. Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1 :19, 1996. 73

[GS04] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1) :5228–5235, 2004. 78

[GSB+13] Luigi Grimaudo, H. Song, Mario Baldi, Marco Mellia, and M. Munafò. TUCAN : Twitter User Centric ANalyzer. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, ASONAM'13, pages 1455–1457, 2013. 51, 53

[GSBT04] Thomas L Griffiths, Mark Steyvers, David M Blei, and Joshua B Tenenbaum. Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544, 2004. 137

[Has70] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, 1970. 78

[HCZ+13] Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. STED : Semi-supervised Targeted-interest Event Detection in Twitter. In *Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1466–1469, 2013. 30, 33, 49, 51, 53

[HD10] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88, 2010. 19, 49, 53

[Hei05] Gregor Heinrich. Parameter estimation for text analysis. Technical report, Technical report, 2005. 11, 77, 78

[HJWK12] Yuheng Hu, Ajita John, Fei Wang, and Subbarao Kambhampati. ET-LDA : Joint Topic Modeling for Aligning Events and their Twitter Feedback. In *AAAI*, volume 12, pages 59–65, 2012. 35, 37, 53, 54

[HK07] Martin J Halvey and Mark T Keane. An assessment of tag presentation techniques. In *Proceedings of the 16th international conference on World Wide Web*, pages 1313–1314. ACM, 2007. 109

[HMS14] Martin Harrysson, Estelle Métayer, and Hugo Sarrazin. The strength of weak signals. *McKinsey Quarterly*, 2014. 18

[Hof99] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, UAI'99, pages 289–296. Morgan Kaufmann Publishers Inc., 1999. 72

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM, 2000. 59

[HTK13] Yuheng Hu, Kartik Talamadupula, and Subbarao Kambhampati. Dude, srsly ? : The surprisingly formal nature of twitter's language. In *Proceedings of the Seventh International Conference on Weblogs and Social Media*, ICWSM'13, 2013. 19

[IM98] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors : towards removing the curse of dimensionality. In *Proceedings of the 30th Annual Symposium on Theory of Computing*, STOC'98, pages 604–613, 1998. 39

[JLZ+11] Ou Jin, Nathan N Liu, Kai Zhao, Yong Yu, and Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th International Conference on Information and Knowledge Management*, CIKM'11, pages 775–784, 2011. 36, 37, 53

[JM00] Daniel Jurafsky and James H. Martin. *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000. 29

[Joa98] Thorsten Joachims. *Text categorization with support vector machines : Learning with many relevant features*. Springer, 1998. 38

[Jon72] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1) :11–21, 1972. 30

[JP73] Raymond A Jarvis and Edward A Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, 100(11) :1025–1034, 1973. 31

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter : understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, pages 56–65, 2007. 19

[KCLC13] Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes : A study of geotagged tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW'13, pages 667–678, 2013. 20

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media ? In *Proceedings of the 19th International Conference on World Wide Web*, WWW'10, pages 591–600, 2010. 20, 64

[KMBS11] Shiva Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee, and Vikas Sindhwani. Emerging topic detection using dictionary learning. In *Proceedings of the 20th International Conference on Information and Knowledge Management*, CIKM'11, pages 745–754, 2011. 39, 41

[KMS+08] Daniel A Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. *Visual data mining. chap. Visual analytics : Scope and challenges*. Springer, 2008. 12, 108, 109

[KMSZ09] Daniel A Keim, Florian Mansmann, Andreas Stoffel, and Hartmut Ziegler. *Visual analytics*. Springer, 2009. 22

[KPC14] Dongyeop Kang, Youngja Park, and Suresh N Chari. Hetero-Labeled LDA : A Partially Supervised Topic Model with Heterogeneous Labels. In *Machine Learning and Knowledge Discovery in Databases*, pages 640–655. Springer, 2014. 137

[Kru64] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1) :1–27, 1964. 43

[KWT07] Kenichi Kurihara, Max Welling, and Yee Whye Teh. Collapsed variational Dirichlet process mixture models. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2796–2801. Morgan Kaufmann Publishers Inc., 2007. 125

[LB12] Himabindu Lakkaraju and Chiranjib Bhattacharyya. Dynamic multi-relational chinese restaurant process for analyzing influences on users in social media. In *Proceedings of the 12th International Conference on Data Mining*, ICDM'12, 2012. 44, 46, 137

[LBM12] Wei Li, David Blei, and Andrew McCallum. Nonparametric bayes pachinko allocation. *arXiv preprint arXiv :1206.5270*, 2012. 137

[LCB12] Jey Han Lau, Nigel Collier, and Timothy Baldwin. On-line Trend Analysis with Topic Models :# twitter Trends Detection Topic Model Online. In *COLING*, pages 1519–1534, 2012. 40, 41, 42, 49

[Lin91] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1) :145–151, 1991. 40

[LLM13] Pei Lee, Laks V.S. Lakshmanan, and Evangelos Milios. KeySee : Supporting Keyword Search on Evolving Events in Social Streams. In *Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1478–1481. ACM, 2013. 12, 110

[LMD10] Marie Liénou, Henri Maître, and Mihai Datcu. Semantic annotation of satellite images using latent dirichlet allocation. *Geoscience and Remote Sensing Letters, IEEE*, 7(1) :28–32, 2010. 74

[LRB09] Marie-Jeanne Lesot, Maria Rifqi, and Hamid Benhadda. Similarity measures for binary and numerical data : a survey. *International Jour-*

nal of Knowledge Engineering and Soft Data Paradigms, 1(1) :63–84, 2009. 95

[LSD12] Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent : segment-based event detection from tweets. In *Proceedings of the 21st International Conference on Information and Knowledge Management*, CIKM'12, pages 155–164, 2012. 31, 33, 53

[LSLL09] Duen-Ren Liu, Meng-Jung Shih, Churn-Jung Liau, and Chin-Hui Lai. Mining the change of event trends for decision support in environmental scanning. *Expert Systems with Applications*, 36(2) :972–984, 2009. 42

[Luh58] Hans Peter Luhn. A business intelligence system. *IBM Journal of Research and Development*, 2(4) :314–319, 1958. 18

[LWH+12] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. Twiner : named entity recognition in targeted twitter stream. In *Proceedings of the 35th International Conference on Research and Development in Information Retrieval*, SIGIR'12, pages 721–730, 2012. 31

[LY12] Rong Lu and Qing Yang. Trend Analysis of News Topics on Twitter. *International Journal of Machine Learning and Computing, June 2012*, Vol. 2 No. 3, 2012. 111

[Mal99] Stéphane Mallat. *A wavelet tour of signal processing*. Academic Press, 1999. 39

[MBB+11] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo : aggregating and visualizing microblogs for event exploration. In *Proceedings of the Annual Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, 2011. 35, 37, 51

[MC85] Glenn W Milligan and Martha C Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2) :159–179, 1985. 50

[Mit97] Tom M Mitchell. Machine Learning (McGraw-Hill International Editions Computer Science Series). 1997. 32

[ML02] Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on*

Uncertainty in artificial intelligence, pages 352–359. Morgan Kaufmann Publishers Inc., 2002. 78

[MMJ13] Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. Building a Large-scale Corpus for Evaluating Event Detection on Twitter. In *Proceedings of the 22Nd International Conference on Information & Knowledge Management*, CIKM'13, pages 409–418, 2013. 27, 50, 52

[Mon85] Douglas C Montgomery. *Statistical quality control*. John Wiley, 1985. 34

[MPLC13] Fred Morstatter, Jurgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough ? Comparing data from twitter's streaming api with twitter's firehose. *Proceedings of ICWSM*, 2013. 109

[MRR+53] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6) :1087–1092, 1953. 78

[MS72] Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2) :176–187, 1972. 42

[MSBX13] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International Conference on Research and dDevelopment in Information Retrieval*, SIGIR'13, pages 889–892. ACM, 2013. 53

[MSH+13] Sana Malik, Alison Smith, Timothy Hawes, Panagis Papadatos, Jianyu Li, Cody Dunne, and Ben Shneiderman. TopicFlow : visualizing topic alignment of Twitter data over time. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, ASONAM'13, pages 720–726, 2013. 12, 44, 46, 73, 105, 112, 114

[NASW09] David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *The Journal of Machine Learning Research*, 10 :1801–1828, 2009. 78

[NBG11] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy : Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, 62(5) :902–918, May 2011. 28

[New06] M E Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23) :8577–8582, June 2006. 31, 43

[NGKA11] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Searching microblogs : coping with sparsity and document quality. In *Proceedings of the 20th International Conference on Information and Knowledge Management*, CIKM'11, pages 183–188, 2011. 19

[PBFC13] Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. Emoticon Style : Interpreting Differences in Emoticons Across Cultures. In *International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*. AAAI, 2013. 110

[PBG+14] Lambert Pépin, Julien Blanchard, Fabrice Guillet, Pascale Kuntz, and Philippe Suignard. Visual Analysis of Topics in Twitter Based on Co-Evolution of Terms. In *Data Science, Learning by Latent Structures, and Knowledge Discovery*. Springer, 2014. 51

[PBS11] Daniel J Power, Frada Burstein, and Ramesh Sharda. Reflections on the past and future of decision support systems : Perspective of eleven pioneers. In *Decision Support*, pages 25–48. Springer, 2011. 17

[PG09] Michael Paul and Roxana Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP'09, pages 1408–1417, 2009. 36

[PG10] Michael Paul and Roxana Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. *Urbana*, 51 :61801, 2010. 36

[PGMJ11] Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. Do all birds tweet the same ? : characterizing twitter around the world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1025–1030. ACM, 2011. 19

[PK13] Ruchi Parikh and Kamalakar Karlapalem. ET : Events from Tweets. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 613–620, 2013. 31, 33, 50, 51, 53

[PNI+08] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM, 2008. 78

[POL10] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming First Story Detection with Application to Twitter. In *Proceedings of the 2010 Conference on Human Language Technologies*, HLT'10, pages 181–189, 2010. 39, 41, 50

[PSD00] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2) :945–959, 2000. 74

[QWW+14] Zhuolin Qiu, Bin Wu, Bai Wang, Chuan Shi, and L Yu. Collapsed Gibbs Sampling for Latent Dirichlet Allocation on Spark. *Journal Machine Learning Research*, 36 :17–28, 2014. 78

[RDL10] Daniel Ramage, Susan Dumais, and Daniel Liebling. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM'10, pages 130–137, 2010. 19, 109

[REC+12] Alan Ritter, Oren Etzioni, Sam Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining*, SIGKDD'12, pages 1104–1112, 2012. 27, 30, 33, 49, 51, 53

[RH10] Philip Resnik and Eric Hardisty. Gibbs sampling for the uninitiated. Technical report, DTIC Document, 2010. 78

[RHNM09] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA : A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1*, EMNLP '09, pages 248–256. Association for Computational Linguistics, 2009. 74

[RZGSS04] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004. 74, 137

[San96] HR Sankey. The Thermal Efficiency Of Steam-Engines (Including Appendixes. In *Minutes of the Proceedings*, volume 125, pages 182–212. Thomas Telford, 1896. 22

[San98] HR Sankey. Introductory note on the thermal efficiency of steam-engines. In *Minutes of Proceedings of The Institution of Civil Engineers*, pages 278–283, 1898. 111

[Sch94] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer, 1994. 54

[Set91] Jayaram Sethuraman. A constructive definition of Dirichlet priors. Technical report, DTIC Document, 1991. 125

[SG07] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7) :424–440, 2007. 78

[Shn96] Ben Shneiderman. The eyes have it : A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996. 116

[SK+01] Hee Seok Song, Soung Hie Kim, et al. Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications*, 21(3) :157–168, 2001. 42

[SK13] Tobias Schreck and Daniel Keim. Visual analysis of social media data. *Computer*, 46(5) :68–75, 2013. 22, 108

[SKK+00] Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000. 32

[SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users : real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW'10, pages 851–860, 2010. 38, 41, 49, 51

[SS12] Ankan Saha and Vikas Sindhwani. Learning evolving and emerging topics in social media : a dynamic nmf approach with temporal regularization. In *Proceedings of the Fifth International Conference on Web Search and Data Mining*, WSDM'12, pages 693–702, 2012. 40, 41, 51

[SSP12] Alina Stoica, Philippe Suignard, and Lambert Pepin. Twitter : Extraction, Regroupement et Visualisation pour la Veille Stratégique. In *Actes de la conférence Veille Stratégique et Technologique*, Ajaccio, France, 2012. VSST 2012. 52, 55

[SST+09] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. TwitterStand : News in Tweets. In *Proceedings of the 17th International Conference on Advances in Geographic Information Systems*, GIS'09, pages 42–51, 2009. 32, 33, 53, 109

[TC05] James J Thomas and Kristin Cook. Illuminating the pass : The research and development agenda for visual analytics. *IEEE Computer Society Press*, 2005. 108

[Teh06] Yee Whye Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics, 2006. 137

[TGG07] Yee W Teh, Dilan Görür, and Zoubin Ghahramani. Stick-breaking construction for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 556–563, 2007. 125

[TJ10] Yee Whye Teh and Michael I Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1, 2010. 127, 137

[TJBB06] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006. 124, 125

[Tuk77] John W Tukey. Exploratory data analysis. 1977. 22

[Van05] Jarke J Van Wijk. The value of visualization. In *Visualization, 2005. VIS 05. IEEE*, pages 79–86. IEEE, 2005. 108, 138

[vB10] Artur Šilić and Bojana Dalbelo Bašić. Visualization of text streams : A survey. In *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 31–43. Springer, 2010. 109

[VJLN13] Jan Vosecky, Di Jiang, Kenneth Wai-Ting Leung, and Wilfred Ng. Dynamic Multi-faceted Topic Discovery in Twitter. In *Proceedings of the*

22Nd International Conference on Information & Knowledge Management, CIKM'13, pages 879–884, 2013. 44, 46, 51, 53

[WAB12] Yu Wang, Eugene Agichtein, and Michele Benzi. Tm-lda : efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining*, SIGKDD'12, pages 123–131, 2012. 53

[WBH12] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv :1206.3298*, 2012. 138

[WF05] Ian H Witten and Eibe Frank. *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, 2005. 32

[WL11] Jianshu Weng and Bu-Sung Lee. Event Detection in Twitter. *Proceedings of the International Conference on Weblogs and Social Media*, pages 401–408, 2011. 31, 33

[WLD+14] Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, and Jiawei Han. Constructing Topical Hierarchies in Heterogeneous Information Networks. *Knowledge and Information Systems*, pages 1–30, 2014. 74, 138

[WM06] Xuerui Wang and Andrew McCallum. Topics over time : a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006. 138

[XS10] Han Xiao and Thomas Stibor. Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In *ACML*, pages 63–78, 2010. 78

[XWW+13] Panpan Xu, Yingcai Wu, Enxun Wei, Tai-Quan Peng, Shixia Liu, Jonathan JH Zhu, and Huamin Qu. Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics*, 19(12) :2012–2021, 2013. 42, 46, 51, 112

[XZJ+13] Wei Xie, Feida Zhu, Jing Jiang, Ee-Peng Lim, and Ke Wang. TopicSketch : Real-time Bursty Topic Detection from Twitter. In *Proceedings of the 13th International Conference on Data Mining*, ICDM'13, pages 837–846, 2013. 40, 41, 51

[YA09] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In *Advances in Information Retrieval*, pages 29–41. Springer, 2009. 72

[Zak00] Mohammed J Zaki. Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on*, 12(3) :372–390, 2000. 59

[Zhu92] Jian-Hua Zhu. Issue competition and attention distraction : A zero-sum theory of agenda-setting. *Journalism & Mass Communication Quarterly*, 69(4) :825–836, 1992. 43

l'université nantes angers le mans
COMMUNAUTÉ D'UNIVERSITÉS ET ÉTABLISSEMENTS

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Lambert PÉPIN

Fouille exploratoire de messages publiés sur Twitter pour l'aide à la décision

Visual analytics for decision making in Twitter messages

Résumé

Depuis son lancement en 2006, Twitter n'a cessé de gagner en popularité et s'est maintenant installé dans une position d'acteur incontournable de la diffusion d'information. Son utilisation s'est démocratisée et ce nouveau canal de communication fait maintenant partie intégrante des stratégies décisionnelles de la gestion de la relation client. Si la compréhension du langage naturel est depuis longtemps au cœur des enjeux industriels, les données issues de Twitter introduisent de nouvelles contraintes qui invalident les approches classiques. Ces textes informels sont courts, contiennent de nombreuses innovations linguistiques et traitent de sujets variés qui s'enchaînent à un rythme journalier. Dans cette thèse, nous nous sommes placés dans un contexte de fouille exploratoire permettant, en collaboration étroite avec un expert de l'entreprise, d'évaluer plusieurs approches sans *a priori* sur les données. L'application de méthodes, traitant aussi bien les documents dans leur ensemble que les termes individuellement, a montré la nécessité de se placer à un niveau de granularité intermédiaire afin de construire des connaissances suivies dans ce contexte de forte nouveauté. Pour ce faire, nous proposons une approche adaptant la capacité d'abstraction des modèles de thèmes ("Topic Models") aux flux de textes dynamiques. En effet, ces modèles probabilistes permettent de modéliser des thématiques à différents niveaux de granularité et peuvent être combinés afin de suivre leurs évolutions temporelles. Notre méthodologie, qui en sus de la phase d'extraction des thèmes s'appuie sur une restitution visuelle interactive de l'évolution temporelle de leurs relations fait tout particulièrement sens dans le domaine très actif de la modélisation de thématiques puisqu'elle permet d'intégrer facilement de nouveaux modèles apportant des connaissances supplémentaires.

Mots clés

Topic detection & tracking, Fouille exploratoire, Médias sociaux, Aide à la décision.

Abstract

Since its launch in 2006, Twitter's popularity grows bigger and bigger and the miccroblogging service has now become a major platform of information diffusion. Twitter has reach a massive popularity and this new communication tool is now part of all decision making strategies in customer relationship management systems. If the understanding of natural language is, for a long time, at the core of industrial needs, Twitter's data brings new constraints making classical approaches obsoletes. These informal texts are shorts, they contain many innovative spellings and deal with various subject matter evolving on a daily basis. In this thesis, we use the paradigm of visual analytics in order to evaluate different approaches in collaboration with a domain expert. The application of methods, dealing with both documents and individual terms, shows that an intermediate level of abstraction is needed to model the evolution of informative knowledge in a context of concept drift. To do so, we propose a methodology taking advantage of the latent semantic extracted by topic models to analyze text streams. Indeed, these probabilistic models extract topics at various granularity levels and can be combined to model their temporal evolution. Our methodology, which both extract topics and provide an interactive visualization of their evolution, is particularly well suited in the fast evolving domain of topic modeling since it is able to easily integrate new models bringing additional knowledge.

Key Words

Topic detection & tracking, Visual analytics, Social media, Decision support.