l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Rami BELKAROUI

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

École doctorale : Sciences et Technologies de l'Information, et Mathématiques

Discipline : Informatique, section CNU 27
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)
Laboratoire de Recherche Opérationnelle, de Décision et de Contrôle de Processus (LARODEC)

Soutenue le 31 mai 2018

Vers une contextualisation des tweets basée sur une analyse des graphes de conversations

JURY

Rapporteurs :	**M^me^ Chiraz LATIRI**, Maître de conférences, HDR, Université de la Manouba, Tunisie
	M. Jean DIATTA, Professeur des Universités, Université de la Réunion, France
Examinateurs :	**M. Mohand BOUGHANEM**, Professeur des Universités, Université de Toulouse, France
	M^me^ Hanène AZZAG, Maître de conférences, HDR, Université Paris 13, France
Directrices de thèse :	**M^me^ Pascale KUNTZ**, Professeur des Universités, Université de Nantes, France
	M^me^ Rim FAIZ, Professeur des Universités, Université de Carthage, Tunisie

Table des matières

1 Introduction **1**

Introduction **1**

1.1 De l'essor des usages de Twitter à ses limites . . . 1
1.2 Du tweet isolé au tweet contextualisé . . . 2
1.3 Conversations autour d'un tweet . . . 3
1.4 Plan du manuscrit . . . 4

2 Contextualisation des messages courts **6**

2.1 Introduction . . . 6
2.2 Vers la RI contextuelle . . . 7
2.3 Définitions du contexte en RI . . . 8
2.4 Taxonomies des facteurs impliqués . . . 9
2.4.1 Taxonomie de Dey 1999 . . . 10
2.4.2 Taxonomie de Fuhr 2001 . . . 10
2.4.3 Taxonomie de Cool 2002 . . . 10
2.4.4 Taxonomie de Goker et al. 2002 . . . 11
2.4.5 Taxonomie de Briconsouf et al. 2007 . . . 11
2.4.6 Taxonomie de Mansourian 2008 . . . 11
2.4.7 Taxonomie de Tamine et al. 2010 . . . 12
2.5 Contextualisation des tweets . . . 13
2.5.1 Sources d'information pour la contextualisation des tweets . . . 13
2.5.2 Approches de contextualisation des tweets . . . 15
2.6 Discussion : positionnement de notre méthode de contextualisation par rapport à la littérature . . . 19
2.7 Conclusion . . . 20

3 Conversations sur Twitter **21**

3.1 Introduction . . . 21
3.2 Objectifs majeurs de l'analyse des conversations sur Twitter . . . 22
3.3 Définitions des conversations . . . 23
3.4 Méthodes de construction des conversations . . . 24
3.4.1 Méthodes de construction à base de relations . . . 24
3.4.2 Méthodes de construction à base de seuil . . . 26
3.4.3 Méthodes de construction à base de classification . . . 27
3.5 Les indices d'analyse des conversations . . . 28
3.5.1 Notations . . . 29
3.5.2 Indices structurels sur les graphes de conversations . . . 29
3.5.3 Indices associés aux contenus des conversations . . . 30
3.5.4 Indices associés aux participants des conversations . . . 31
3.6 Conclusion . . . 33

4 Caractérisation des conversations sur Twitter **34**
4.1 Introduction . . . 34
4.2 Analyses expérimentales de la littérature . . . 35
4.2.1 Collecte des données sur Twitter . . . 35
4.2.2 Propriétés des conversations sur Twitter . . . 41
4.3 Un nouveau graphe de conversation . . . 42
4.4 Distribution observée des indices . . . 44
4.4.1 Illustration sur un graphe de conversation . . . 44
4.4.2 Résultats expérimentaux observés sur un corpus de graphes de conversations . . . 53
4.5 Conclusion . . . 61

5 Contextualisation basée sur les conversations **62**
5.1 Introduction . . . 62
5.2 Approche proposée : processus de contextualisation . . . 63
5.2.1 Construction des requêtes . . . 63
5.2.2 Collecte des tweets et construction des graphes de conversations . . . 64
5.2.3 Indicateurs d'analyse . . . 64
5.2.4 Génération du contexte . . . 66
5.3 Contextualisation des tweets : notre système proposé . . . 68
5.3.1 Architecture générale du système . . . 69
5.3.2 Contextualisation des tweets : Exemples de scénarios d'utilisation . . . 71
5.4 Conclusion . . . 77

6 Expérimentations **79**
6.1 Introduction . . . 79
6.2 Contextualisation des tweets : État de l'art sur les données et les résultats . . . 80
6.2.1 Les données INEX/CLEF . . . 80
6.2.2 Évaluation de la contextualisation des tweets . . . 82
6.2.3 Discussion sur les résultats de la campagne INEX/CLEF "Tweet Contextualization" . . . 84
6.3 Évaluations de notre approche : protocole expérimental . . . 85
6.3.1 Étude utilisateurs . . . 85
6.3.2 Les entretiens . . . 87
6.4 Résultats . . . 87
6.4.1 Résultats issus du questionnaire . . . 88
6.4.2 Résultats extraits des entretiens . . . 90
6.5 Conclusion . . . 91

Conclusion **92**
6.6 Perspectives . . . 93

Liste des tableaux

4.2 Variabilité de la taille des corpus des données utilisés dans les travaux effectués sur les conversations Twitter 40
4.4 Variabilité et détails des données utilisées dans nos expérimentations 54

Table des figures

2.1 Dimensions du contexte multidimensionnel de Tamine et al. 2010 (Tamine-Lechani et al., 2010) . . . 12
2.2 Exemple de résumé de contextualisation pour un tweet ambigu . . . 14

3.1 Exemple d'un résultat de recherche sous forme d'une conversation . . . 23
3.2 Structure en "étoile" d'un graphe construit en utilisant un seuil de similarité textuelle . . . 26
3.3 Les différentes méthodes de construction des conversations . . . 28

4.1 Exemple des attributs fournis avec chaque tweet . . . 36
4.2 Les différentes méthodes de collecte des données . . . 37
4.3 Nouvelle modélisation du graphe de conversation . . . 43
4.4 Distribution de la taille des composantes connexes dans le graphe de conversation 44
4.5 La taille de la grande composante connexe retenue pour nos expérimentations . 44
4.6 Distribution des scores d'influence des participants de la plus grande composante connexe du graphe de conversation . . . 45
4.7 Distribution des scores d'influence des participants ayant un tweet ou une réponse de la plus grande composante connexe du graphe de conversation . . . 45
4.8 Distribution des scores de popularité par intervalle des participants de la plus grande composante connexe du graphe de conversation . . . 46
4.9 Distribution de la moyenne de score de popularité par participant de la plus grande composante connexe du graphe de conversation . . . 46
4.10 Corrélation entre le score de popularité et le score d'influence des participants de la plus grande composante connexe du graphe de conversation . . . 47
4.11 Distribution des scores d'expertise des participants de la plus grande composante connexe du graphe de conversation . . . 47
4.12 Distribution des scores d'expertise des participants ayant publié un tweet ou une réponse de la plus grande composante connexe du graphe de conversation . . . 48
4.13 Distribution des scores d'activité sur la plus grande composante connexe du graphe de conversation . . . 48
4.14 Distribution des scores d'influence des tweets de la plus grande composante connexe du graphe de conversation . . . 49
4.15 Distribution des scores de hashtag des tweets de la plus grande composante connexe du graphe de conversation . . . 50
4.16 Distribution des scores d'URL des tweets de la plus grande composante connexe du graphe de conversation . . . 50
4.17 Corrélation entre le nombre de retweets et le nombre de favoris de la plus grande composante connexe du graphe de conversation . . . 51
4.18 Distribution des degrés des sommets tweets de la plus grande composante connexe du graphe de conversation . . . 51

4.19 Distribution des degrés des sommets participants de la plus grande composante connexe du graphe de conversation . . . 52
4.20 Distribution des degrés de centralité des sommets tweets de la plus grande composante connexe du graphe de conversation . . . 52
4.21 Distribution des degrés de centralité des sommets participants de la plus grande composante connexe du graphe de conversation . . . 53
4.22 Distribution des degrés des sommets tweets des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 55
4.23 Distribution des degrés des sommets participants des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 56
4.24 Distribution des degrés de centrailité des sommets tweets des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 56
4.25 Distribution des degrés de centrailité des sommets participants des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 57
4.26 Distribution des scores d'hashtag des tweets des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 57
4.27 Distribution des scores d'URLs des tweets des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 58
4.28 Distribution de la moyenne des scores d'influence des participants des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 59
4.29 Distribution de la moyenne des scores de popularité des participants des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 59
4.30 Distribution de la moyenne des scores d'expertise des participants des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 60
4.31 Distribution de la moyenne des scores d'activité des participants des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 60
4.32 Distribution de la moyenne des scores d'influence des tweets des plus grandes composantes connexes du graphe de conversation (30 sujets) . . . 61

5.1 Chaîne de pré-traitements des tweets bruts et construction de requête . . . 64
5.2 Exemple de construction d'une requête à partir d'un tweet initial . . . 64
5.3 Exemple d'un graphe de conversation . . . 65
5.4 Liste des indicateurs d'analyse . . . 65
5.5 Calcul des indicateurs pour chaque sommet . . . 66
5.6 Modélisation du tweet . . . 67
5.7 Exemple du dictionnaire des hashtags . . . 68
5.8 Système de contextualisation et de visualisation des tweets proposé . . . 68
5.9 Interface du service de manipulation des données : Collecte, import . . . 69
5.10 Architecture du système proposé . . . 70
5.11 L'interface principale de visualisation des graphes de conversations . . . 70
5.12 L'interface principale du service d'analyse des conversations . . . 71
5.13 Exemple d'un tweet ambigu à contextualiser . . . 73
5.14 Requête de recherche . . . 73
5.15 Graphe fourni à l'utilisateur pour contextualiser le tweet . . . 74
5.16 Les éléments d'informations affichés sur les noeuds participants . . . 74
5.17 Les éléments d'informations affichés sur les noeuds tweets . . . 75
5.18 Fonctionnalités de consultation des sujets tendance . . . 75
5.19 Fonctionnalités de consultation des listes de tweets/participants et l'export du graphe vers le logiciel de visualisation Gephi . . . 76
5.20 Un graphe de hashtags . . . 76

5.21 Un graphe de mention 77
5.22 Fonctionnalité des N-meilleures données et des métadonnées 77

6.1 Exemple des données tweets fournies par INEX en 2011 81
6.2 Exemple des données tweets fournies par INEX 2012 81
6.3 Exemple de format de tweet INEX 2014 82
6.4 Exemple d'un tweet ambigu adopté dans notre évaluation 86
6.5 Ensemble des textes de tweets 86
6.6 Âge, sexe et expertise des participants 88
6.7 Résultats de l'évaluation de l'apport des textes de tweets pour la compréhension du tweet ambigu 88
6.8 Résultats de l'évaluation de l'apport des graphes de conversations pour la compréhension du tweet ambigu 89
6.9 Résultats de l'évaluation de l'importance des graphes de conversations par rapport à la lecture des textes de tweets 89
6.10 Résultats de l'évaluation de l'apport des éléments d'informations pour la compréhension du tweet 90
6.11 Résultats de l'évaluation de la compréhension du contexte par rapport au contexte approprié 90

Remerciements

*Je tiens à exprimer ma plus sincère et profonde gratitude au Professeur **Pascale Kuntz** pour sa patience, son soutien et les discussions constructives que nous avons eu tout au long de l'accomplissement de ce travail et au Professeur **Rim Faiz** pour ses encouragements continus, sa motivation, son soutien et ses conseils précieux qui ont mené à la réussite de cette thèse. Leur orientation m'a permis de surpasser des périodes difficiles et a fait renaître en moi la volonté d'aller vers l'avant et de continuer mon chemin vers la réussite. Cette dissertation n'aurait jamais été réalisée sans leur engagement et j'espère continuer à travailler avec ses nobles pensées.*

Je tiens également à remercier tous les membres du jury d'avoir accepté de participer à l'évaluation de ce travail. Je remercie très sincèrement Monsieur Lamjeud Ben said et Monsieur Jean Diatta d'avoir rapporté mon mémoire de thèse. Je remercie également mon examinatrice Madame Hanène Azzag qui m'a fait honneur en acceptant de faire partie de mon jury de thèse. Un spécial remérciment à Monsieur Mohand Boughanem avec qui j'ai fait mes premiers pas en recherche. Merci d'avoir accepté de participer à mon jury de thèse.

*Je tiens à remercier mes collègues de l'équipe **DUKe** et ceux de mon laboratoire **LARODEC** pour leurs amitiés, leur soutien et leurs conseils. Un grand remerciement à mes amis, spécialement **Seif Eddine**, ils ont tous contribué à rendre mon temps plus amusant et intéressant.*

*Les mots ne suffisent pas pour remercier mes parents, mon frère **Ramzi** et toute ma famille pour leur soutien et pour avoir toujours cru en moi. Ce travail n'aurait pas été possible sans votre soutien. Enfin et surtout, je remercie avec grand amour Khouloud qui m'a encouragé à y aller de l'avant.*

Merci infiniment, je vous aime !

Résumé

Les plateformes de microblogging contribuent à l'accroissement continu de la dynamique des échanges sur internet par la diffusion rapide de messages courts à des échelles variées, du cercle d'amis à l'ensemble des utilisateurs. Cependant, cette pratique rédactionnelle peut se heurter à la difficulté de compréhension de certains messages. Même avec le récent passage à 280 caractères, les messages de Twitter considérés dans leur singularité, sans information additionnelle exogène, peuvent confronter leurs lecteurs à des difficultés d'interprétation. L'ajout d'une contextualisation à ces messages s'avère donc une voie de recherche prometteuse pour faciliter l'accès à leur contenu informationnel. Dans la dernière décennie, la majorité des travaux se sont concentrés sur la construction de résumés à partir de sources d'information complémentaires telles que Wikipédia. Nous avons choisi dans cette thèse une voie complémentaire différente qui s'appuie sur l'analyse des conversations sur Twitter. Les fonctionnalités de base de Twitter facilitent les interactions et conduisent à de nouvelles formes d'échanges de type conversationnel modélisables par des graphes. Nous avons proposé une définition qui intègre à la fois les relations entre les messages et les relations entre les individus qui les produisent et nous avons mobilisé des mesures développées en métrologie de réseaux pour caractériser leurs propriétés structurelles. Cette analyse expérimentale, menée sur un ensemble de graphes construits à partir de données réelles, nous a permis d'identifier des mesures pertinentes et des seuils adaptés pour permettre d'extraire des informations utiles à la contextualisation d'un tweet. Ces informations ont été intégrées dans un prototype qui, pour un tweet donné, propose une visualisation d'un sous-graphe du graphe de conversation associé au tweet. Ce sous-graphe extrait automatiquement à partir de l'analyse des distributions des indicateurs structurels, permet de mettre en évidence notamment des individus qui jouent un rôle majeur dans la conversation et des tweets qui ont contribué à la dynamique des échanges. Ce prototype a été testé sur un panel d'utilisateurs, pour valider son apport et ouvrir des perspectives d'amélioration.

Abstract

The microblogging platforms contribute to the continuous increase of the dynamics of information exchange on the Internet by the fast diffusion of short messages on various scales ; from a circle of friends to a whole host of users. However, this editing practice may face difficulty in understanding certain messages. Even with the recent switch to 280 characters, Twitter messages considered in their singularity, without any additional exogenous information, can confront their readers with difficulties of interpretation. The integration of contextualization on these messages is therefore a promising avenue of research to facilitate access to their information content. In the last decade, most works have focused on building summaries from complementary sources of information such as Wikipedia. In this thesis, we choose a different complementary path that relies on the analysis of conversations on Twitter. The basic functionalities of Twitter have facilitated interactions and led to new forms of exchanges mainly conversational ones modeled by graphs. We propose a definition that integrates both the relationships between the messages and the relationships between the individuals who produce them, and we mobilize metrology measures developed to characterize their structural properties. This experimental analysis conducted on a set of graphs built from real data, allows us to identify relevant measures and thresholds adapted to extract conversations from information useful for the contextualization of a tweet. These information were integrated in a prototype which, for a given tweet, offers a visualization of a subgraph of the conversation graph associated with the tweet. This subgraph, automatically extracted from the analysis of structural indicators distributions, allows to highlight particular individuals who play a major role in the conversation and tweets that have contributed to the dynamics of exchanges. This prototype was tested on a panel of users to validate its efficiency and open up prospects for improvement.

1

Introduction

Si, comme le rappelait récemment F. Gandon dans un article consacré au fondateur du World Wide Web Tim Berners-Lee (Gandon, 2017), la recherche de techniques d'organisation et de moyens d'accès efficients aux masses d'information que nous collectons a été une motivation omniprésente dans la préhistoire du Web, la construction d'un nouvel espace social aux interactions facilitées par les technologies est devenue peu à peu un enjeu majeur. Le Web 2.0 a permis d'étendre l'accès à Internet aux utilisateurs disposant de connexions lentes et peu robustes, et ainsi de multiplier les acteurs et leurs activités. Dès le milieu des années 1990, apparaissent les premiers réseaux sociaux. Citons à titre illustratif le site Classmates né en 1995 qui cherchait à référencer les individus ayant suivi une même formation. Son successeur "Sixdegrees", apparu deux ans plus tard, permettait à ses membres de créer du contenu et d'interagir avec d'autres membres. En deux décennies, l'essor du Web social a considérablement modifié le champ du Web. La nouvelle génération de sites associés au Web social a facilité à la fois la publication de contenus et les interactions. Les plateformes de microblogging, qui ont pris leur essor plus récemment, ont complété l'offre en permettant une diffusion rapide des informations à des échelles variées, de l'espace personnel à l'espace planétaire. Twitter est l'exemple le plus populaire des plateformes de microblogging. Il permet de diffuser des messages courts, de 140 caractères à son origine à 280 caractères depuis le 8 novembre 2017 [1], à toutes les échelles. Lancé en juillet 2006, Twitter comptait 94 millions utilisateurs en avril 2007 [2] pour atteindre les 330 millions en 2017 [3]. Et le nombre de tweets est passé de 20 000 tweets par jour en mars 2007 à plus de 500 millions de tweets échangés chaque jour en 2017.

1.1 De l'essor des usages de Twitter à ses limites

Dès ses premiers pas, des usages variés de Twitter ont été identifiés. Des travaux datant de ses premières années d'utilisation (Phelan et al., 2009; Sankaranarayanan et al., 2009; Dong et al., 2010; Diakopoulos and Shamma, 2010; Yardi and Boyd, 2010) ont mis en évidence trois grands motifs d'utilisation : (i) la recherche d'actualités en temps réel, (ii) la recherche d'informations sur des sujets spécifiques et (iii) la recherche de liens sociaux. Dès l'origine, la recherche et le suivi d'informations sur des évènements récents ou en cours ou des sujets "tendance" ont

[1] http ://www.digitalisons-nous.com/2017/11/08/twitter-a-280-caracteres/
[2] https ://usatoday30.usatoday.com/tech/webguide/2007-05-28-social-sites.htm
[3] https ://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

été la motivation majeure de l'usage de Twitter (près de la moitié des usages répertoriés dans plusieurs enquêtes (Phelan et al., 2009; Dong et al., 2010; Diakopoulos and Shamma, 2010)). En complément, la recherche sur des sujets spécifiques a été également souvent citée (Sankaranarayanan et al., 2009; Yardi and Boyd, 2010). Mais, au-delà de la recherche d'informations ciblées à la fois thématiquement et temporellement, Twitter s'est révélé être également un espace d'échanges sociaux : (Teevan et al., 2011) ont observé au début des années 2010 qu'un quart des usages de Twitter étaient motivés par une intention sociale, et notamment par la recherche d'utilisateurs avec des intérêts similaires. Les informations personnelles, les opinions personnelles et le partage de simples idées apparaissent également parmi les contenus fréquents des tweets -représentant plus d'un quart des contenus- (Naaman et al., 2010). Aujourd'hui, l'usage de Twitter n'a pas guère changé ; le partage de l'information reste parmi les usages principaux (Wen et al., 2014). Ainsi, la recherche d'actualités et des news reste encore la plus grande motivation d'utilisation (59 % des utilisateurs) de cette plateforme (Greenwood et al., 2016; Griffin et al., 2017) [4].

Cependant, même si le nombre d'utilisateurs de Twitter continue de progresser (+3% par an en 2017), la pente de cet accroissement commence à se réduire significativement depuis deux ans [5]. Le succès de sa diffusion a été basé sur des innovations technologiques permettant une création de contenus, une interaction aisée et une diffusion très rapide des messages. Mais, aujourd'hui, la recherche d'informations se heurte à la volumétrie des contenus, à la forte volatilité des informations émises, et à la qualité des messages diffusés. Des efforts importants sont déployés en recherche dans le domaine de la science des données pour tenter de catégoriser automatiquement les contenus et améliorer l'analyse textuelle des messages (Manning et al., 1999; George, 2004; Li, 2010) mais des difficultés persistent. La contrainte liée à la taille limitée du message a été détournée par l'utilisation d'un vocabulaire particulier contenant des termes mal orthographiés (Hu et al., 2013; Poblete et al., 2011), des émoticônes et de multiples conventions linguistiques (Naveed et al., 2011). S'ajoute à ceci l'emploi de syntaxes spécifiques porteuses de contenu informationnel tels que les hashtags et les mentions (Carter et al., 2013). Ces spécificités, qui éloignent l'interprétation des tweets de l'interprétation de messages rédigés en langage plus conventionnel, peuvent rendre la compréhension de certains tweets difficile, voire inextricable lorsqu'ils sont considérés isolément.

1.2 Du tweet isolé au tweet contextualisé

L'ajout d'éléments d'informations au tweet brut est donc devenu un challenge aujourd'hui pour faciliter sa compréhension et ainsi contribuer à la poursuite de l'essor de l'appropriation de Twitter en ne le réservant pas aux couches les plus informées de la population -les journalistes restant parmi les utilisateurs les plus actifs sur Twitter (Mercier, 2013; Bagdouri, 2016; Powers and Vera-Zambrano, 2017)-. La contextualisation d'un message court consiste en un apport de connaissances qui peuvent être déclinées selon des aspects liés à l'utilisateur (profil, historique de production ou d'interaction, localisation...) ou à l'environnement du message (liens entre documents, type de document...). Elle a pour objectif d'aider le lecteur à comprendre le contexte du texte court (définition de termes, principales thématiques, visualisation des éléments externes en lien avec le texte, etc).

Les recherches sur la contextualisation des tweets sont assez récentes. Elles ont véritablement pris leur essor avec le challenge "Tweet Contextualization" à INEX ("Initiative for the Evaluation of XML Retrieval") en 2011 (Meij et al., 2010) qui est devenu un laboratoire CLEF ("Conference and Labs of the Evaluation Forum") en 2012 (Bellot et al., 2013; SanJuan

[4] http ://www.pewinternet.org/2016/11/11/social-media-update-2016/
[5] https ://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

et al., 2012). Dans le cadre de ce challenge, qui se renouvelle tous les ans, la contextualisation d'un tweet prend la forme d'un résumé, rédigé en langue naturelle, de 500 mots. Le résumé est construit automatiquement à partir des ressources textuelles extraites de Wikipédia. Différentes approches ont été explorées : extraction de passages de textes de Wikipédia basée sur des algorithmes de modélisation thématique, reconnaissance d'entités nommées dans des phrases candidates, enrichissement sémantique et expansion de messages. Les résultats expérimentaux obtenus sont très prometteurs, mais cette voie de recherche n'exploite pas la richesse des informations contenues dans les échanges liés à un tweet sur Twitter. Nous avons donc opté dans cette thèse pour une voie complémentaire qui contextualise les tweets à partir des informations extraites des conversations associées.

1.3 Conversations autour d'un tweet

Dès le début du siècle dernier, le sociologue G. Tarde soulignait l'intérêt majeur que pourrait représenter une "histoire complète de la conversation chez tous les peuples et à tous les âges" (de Tarde, 1901). Son projet s'est partiellement concrétisé à partir des années 1960, où sous l'impulsion des recherches relevant de différents champs disciplinaires -linguistique mais aussi sociologique, anthropologique et psychologique-, l'analyse des conversations "ordinaires" à partir de l'observation des pratiques a connu un véritable essor. L'analyse des conversations renvoie alors à "tout type d'échange verbal, quelles qu'en soient la nature et la forme" (Traverso, 1999). Mais, à notre connaissance, Tarde n'imaginait pas encore les conversations médiatisées ou médiées ou instrumentées selon les auteurs (Panckhurst, 1997). La variété des pratiques n'a cessé de croître avec l'essor des technologies de l'information et de la communication. Dans un ouvrage qui ne date que de 2011, (Develotte et al., 2011) énumèrent les dispositifs les plus courants -dont certaines dénominations semblent déjà désuètes- : le courrier électronique, le forum de discussion, la liste de diffusion, le tchat, la messagerie instantanée, le téléphone par Internet, la visio-conférence, le blog et les réseaux sociaux. Rappelons que dès sa création, Twitter a été conçu pour "permettre aux utilisateurs de partager facilement des petits moments de vie avec leurs amis" (interview de J. Dorsey sur YouTube 2007) [6]. Et les enquêtes des premières années ont confirmé que près d'un quart de tous les utilisateurs de Twitter tiennent des conversations avec d'autres utilisateurs (Java et al., 2007; Steiner, 2009; Honey and Herring, 2009) et que plus d'un tiers des tweets sont conversationnels (Ritter et al., 2010). Aujourd'hui, encore la pratique des conversations persiste parmi les usages principaux de Twitter (Wen et al., 2014; Dann, 2015). Si le "partage" n'est pas tout à fait "l'échange" qui est sous-jacent à la définition familière d'une conversation telle qu'on peut la trouver dans le Littré, les interactions ont rapidement émergé grâce aux trois fonctionnalités "reply", "retweet" et "mention" offertes par Twitter pour assurer les activités de conversations.

Ces nouvelles formes de conversation mettent en jeu des contenus -portés par les tweets- et aussi des relations entre les tweets et entre les individus qui les produisent, les propagent, et y répondent. Ces individus peuvent être regroupés en trois catégories majeures (Java et al., 2007) : les fournisseurs d'informations appelés aussi sources d'informations, les amis, et les demandeurs d'informations. Plus précisément, les sources d'informations publient des nouvelles et ont tendance à avoir une grande base de suiveurs (followers) ; ces sources peuvent être des individus ou des services automatisés. Les "amis" représentent une vaste catégorie qui englobe la plupart des utilisateurs, y compris la famille et les collègues de travail. Les demandeurs d'informations ont tendance à être les utilisateurs qui partagent rarement des tweets mais qui suivent d'autres utilisateurs régulièrement. Ainsi, un tweet se retrouve inséré dans un réseau relationnel dont les composantes et la structuration des liens entre elles ne sont pas aléatoires et nous avons

[6] https ://www.youtube.com/watch ?v=uRXsgEBFy6A

fait l'hypothèse dans cette thèse que leur analyse permettrait de contribuer à la compréhension du tweet. Dans notre travail nous n'avons pas considéré le contenu textuel des tweets et nous nous sommes focalisés sur les liens qui sous-tendent les conversations. L'importance des relations a été soulignée par les pionniers de l'analyse des réseaux sociaux dès les années 1960. L'explosion des travaux sur les réseaux sociaux des deux dernières décennies ont confirmé leur rôle dans l'organisation sociale et ont contribué à développer, ce qu'un des fondateurs français de l'analyse des réseaux sociaux A. Degenne, appelait "l'économétrie des réseaux". Dans cette thèse, nous avons donc cherché à exploiter ces relations pour contextualiser les tweets et faciliter ainsi leur interprétation.

1.4 Plan du manuscrit

Ce manuscrit est organisé en six chapitres complétés par une conclusion qui ouvre des perspectives. Les deux premiers (chapitres 2 et 3) sont consacrés à un état de l'art sur la contextualisation et l'analyse des conversations sur Twitter. Les deux suivants (chapitre 4 et 5) décrivent l'apport de notre travail : l'analyse expérimentale d'un ensemble de conversations à partir d'un panel d'indicateurs permettant d'en déduire des invariants comportementaux, et la proposition d'une nouvelle approche de contextualisation des tweets intégrant les informations relationnelles. Le dernier chapitre (chapitre 6) présente une validation expérimentale du système développé.

- Le *chapitre 2* présente un état de l'art sur la notion du contexte. Il permet de montrer la diversité des définitions qui ont été proposées pour dresser les contours de cette notion protéiforme. Pour l'aborder ensuite d'un point de vue opérationnel, nous précisons les différentes dimensions associées à la contextualisation mobilisées dans les processus de Recherche d'Information. Ces dimensions sont extraites de différentes taxonomies qui ont été proposées dans la littérature. Puis, nous nous focalisons sur les travaux qui portent précisément sur la contextualisation des tweets en les organisant autour de trois grandes familles : (i) les approches basées sur du résumé automatique, (ii) les approches de modélisation thématique et (iii) les techniques d'expansion des tweets.

- Le *chapitre 3* porte sur les travaux consacrés d'analyse des conversations sur Twitter. Nous présentons les deux grands modèles de conversation, basés sur des graphes, utilisés dans la littérature. Puis, nous décrivons les principales étapes des algorithmes de construction des graphes de conversations. Et nous présentons une famille d'indices classiquement utilisés pour caractériser leurs propriétés structurelles.

- Le *chapitre 4* est consacré à une analyse expérimentale des graphes de conversations associés à des tweets. La première partie est consacrée à la collecte des données ; nous présentons différentes méthodes de collecte décrites dans la littérature et synthétisons les caractéristiques des corpus recueillis pour les expérimentations. Nous complétons cet état de l'art par les principales conclusions issues de ces expérimentations. Si les conclusions des différents auteurs permettent de relever quelques invariants, la variété des corpus utilisés et l'absence de protocole partagé par toutes les recherches ont été un frein à une caractérisation suffisamment précise du comportement de l'ensemble des indicateurs que nous avons sélectionnés. La deuxième partie du chapitre est donc consacrée à nos propres expérimentations. Nous avons extrait plus de 100 000 graphes de conversations à partir de 30 sujets (ex : #Elections2017) et nous avons étudié les distributions statistiques des indicateurs présentés au chapitre 3. Cette analyse permet de sélectionner les indicateurs pertinents pour l'extraction d'informations utiles à la contextualisation ainsi que leurs intervalles de valeurs représentatifs des comportements que nous souhaitons mettre en évidence (ex : noeuds associés aux participants les plus actifs).

- Le *chapitre 5* présente notre nouvelle approche de contextualisation des tweets basée sur l'analyse des conversations Twitter. Il décrit les quatre phases majeures de l'approche : construction des requêtes, collecte des tweets et construction des graphes de conversations, filtrage des graphes par les indicateurs d'analyse, et génération du contexte avec une présentation détaillée des composantes de l'interface de visualisation interactive. Des exemples illustrent chaque phase et deux scenarios d'usage complètent la présentation.

- Le *chapitre 6* est consacré à la validation expérimentale de notre approche. La première partie rappelle les principales caractéristiques des différents protocoles d'évaluation mis en place lors du challenge INEX "Tweet contextualization" qui s'étend sur les six dernières années. La deuxième partie du chapitre présente une étude utilisateurs qui s'appuie sur un questionnaire que nous avons menée pour évaluer la pertinence de notre approche. Les questionnaires ont été complétés par des entretiens semi-directifs qui ont permis d'approfondir les statistiques recueillies par ces derniers.

  Enfin, une conclusion générale clôt le document en résumant le travail réalisé au cours de cette thèse et en présentant les perspectives de recherche.

2

Contextualisation des messages courts

Sommaire

2.1 Introduction 6
2.2 Vers la RI contextuelle 7
2.3 Définitions du contexte en RI 8
2.4 Taxonomies des facteurs impliqués 9
2.4.1 Taxonomie de Dey 1999 10
2.4.2 Taxonomie de Fuhr 2001 10
2.4.3 Taxonomie de Cool 2002 10
2.4.4 Taxonomie de Goker et al. 2002 11
2.4.5 Taxonomie de Briconsouf et al. 2007 11
2.4.6 Taxonomie de Mansourian 2008 11
2.4.7 Taxonomie de Tamine et al. 2010 12
2.5 Contextualisation des tweets 13
2.5.1 Sources d'information pour la contextualisation des tweets . . 13
2.5.2 Approches de contextualisation des tweets 15
2.6 Discussion : positionnement de notre méthode de contextualisation par rapport à la littérature 19
2.7 Conclusion 20

2.1 Introduction

Comme le rappelaient récemment (Brézillon and Gonzalez, 2014) dans un ouvrage collectif, le contexte a toujours joué un rôle important dans l'intelligence humaine. Mais, en dépit de cette importance mise en évidence dans de nombreux domaines, la notion de contexte n'est encore que partiellement cernée et les nombreuses définitions proposées dans la littérature témoignent de l'absence de consensus. Citons à titre illustratif quelques

définitions proposées dans différents domaines. Dans le dictionnaire "Oxford Advanced Learner's", un contexte est "une situation dans laquelle quelque chose se produit et qui vous aide à la comprendre" (Crowther, 1995). Une définition assez proche se retrouve dans le Webster : "the situation in which something happens : the group of conditions that exist where and when something happens" (Merriam-Webster, 2006). Le lien avec "la situation" se retrouve en effet chez différents auteurs qui le précisent : "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves" (Dey, 2001). Des définitions plus spécifiques dépendent du domaine applicatif : par exemple "context is what contrains a problem solving without intervening in it explicitly" (Brézillon, 1999). Une base de données de plus de 150 définitions provenant de disciplines diverses (informatique, philosophie, économie, etc.) avait même été constituée il y a une dizaine d'années dans l'objectif d'en tirer un consensus (Bazire and Brézillon, 2005) mais des propositions récentes (Alegre et al., 2016; Temdee and Prasad, 2018) montrent que le débat est encore ouvert. (McCarthy, 1993) soutenait même que le contexte était de dimension infinie. Cependant, de nombreux auteurs font l'hypothèse qu'un nombre restreint de dimensions est nécessaire en pratique, et que les humains peuvent identifier les aspects contextuels importants rapidement.

Dans ce chapitre, nous tentons de cerner ces dimensions en nous restreignant à un domaine ; celui de la Recherche d'Informations (RI). Nous commençons par la présentation de la notion du contexte ainsi que de son exploitation dans le processus de la RI. Puis, nous présentons la notion de contextualisation et les différentes sources d'informations exploitées dans les travaux de la littérature. Suite à cela, les principales approches de contextualisation des tweets sont introduites. Enfin, une discussion sur les principales conclusions tirées à partir des travaux de la littérature est proposée.

2.2 Vers la RI contextuelle

La Recherche d'Information (RI), qui a une longue histoire en informatique (Baeza-Yates and Ribeiro-Neto, 1999; Singhal, 2001), vise à développer des solutions automatiques, dont les premières remontent aux années 50 (Luhn, 1957), pour faciliter l'accès à l'information pertinente et pour satisfaire un besoin exprimé par un utilisateur. Ce besoin est généralement formulé à l'aide d'une requête décrite par un ensemble de mots clés. Pour une requête utilisateur, un système de RI permet de retrouver un sous-ensemble de documents susceptibles d'être pertinents, à partir d'une collection de documents, en réponse à cette requête.

L'une des premières définitions pour les systèmes de RI est proposée par Salton en 1968 : "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information" (Salton, 1968). Les composantes de base d'un système de RI se retrouvent deux décennies plus tard chez le même auteur (Salton and McGill, 1983) : "An information retrieval system is an information system, that is, a system used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations". D'autres définitions des pionniers du domaine s'axaient plus sur les besoins des utilisateurs : "the goal of an information [retrieval] system is for the user to obtain information from the knowledge resource which helps her/him in problem management" (Belkin, 1984).

L'essor des besoins associés à celui de l'essor de l'informatique a conduit à de très nom-

breux progrès dans les premières décennies de la RI. La puissance des moteurs de recherche s'est démultipliée mais les résultats proposés sont restés longtemps - et encore aujourd'hui pour beaucoup - indépendants des profils des utilisateurs et des contextes dans lesquels la recherche est effectuée. Cette indépendance aux situations de recherche des utilisateurs peut rendre les résultats peu pertinents. Pour reprendre un exemple classique (Sugiyama et al., 2004), pour la requête "Java", certains utilisateurs sont intéressés par des documents en rapport avec le langage de programmation Java alors que d'autres souhaitent des informations concernant l'île Java, ou bien relatifs au café Java.

Pour tenter de palier ces limites, des travaux en RI s'orientent vers une nouvelle génération de moteurs de recherche basés sur l'accès contextuel à l'information (Recherche d'information contextuelle). Selon (Allan et al., 2003), la RI contextuelle consiste à exploiter le contexte de l'utilisateur ainsi que des connaissances liées à la requête dans le but de mieux répondre aux besoins en information par une information pertinente et appropriée au contexte de l'utilisateur. La RI contextuelle se déroule selon un processus d'interaction homme-machine (H-M) où plusieurs facteurs interviennent et influencent la perception de pertinence de l'information du côté utilisateur. Celui-ci apporte son jugement sur les documents renvoyés par le système selon des critères liés au contexte dans lequel la recherche est effectuée. Ces recherches nécessitent donc une définition opérationnelle de la notion de "contexte".

2.3 Définitions du contexte en RI

Nous retrouvons ici à nouveau une variété de définitions comme nous l'avons déjà souligné dans le cadre plus général de l'introduction de ce chapitre. Le contexte est défini comme "l'ensemble des facteurs cognitifs et sociaux ainsi que les buts et les intentions de l'utilisateur au cours d'une session de recherche" (Ingwersen, 1996). Une tentative de distinction entre ces notions a fait l'objet d'autres travaux (Sonnenwald, 1999) qui précisent que le contexte couvre des aspects larges tels que l'environnement cognitif, social et professionnel dans lesquels s'inscrivent des situations liées à des facteurs tels que le lieu, le temps et l'application. C'est ce sens générique du contexte qui a été largement exploré cette dernière décennie en RI (Ingwersen and Järvelin, 2005).

Plus précisément, le contexte peut être défini, selon (Dourish, 2004) suivant deux aspects : un aspect de représentation, et un aspect interactionnel. Dans le premier cas, il est considéré comme une "forme d'information" qui est délimitée, stable et indépendante de l'activité. Il se compose des notions qui décrivent l'utilisateur et l'environnement dans lequel les activités d'information se produisent. Le deuxième cas définit le contexte comme un résultat de l'activité de l'utilisateur, à partir duquel il ne peut pas être séparé.

Nous retrouvons essentiellement le premier cas chez de nombreux auteurs :

- "Un cadre qui a des composantes sociales/physiques et conceptuels, y compris si la tâche se fait en collaboration ou seul et les demandeurs d'information physiques et les états psychologiques" (Marchionini, 1995).
- "La description des aspects d'une situation" (Göker and Myrhaug, 2002), similaire à celle de (Dourish, 2004).
- "Le contexte a le potentiel d'être pratiquement rien ... [il est] une sorte de conteneur dans lequel le phénomène réside" (Dervin, 1997).

Plus rares sont les définitions qui ont abordé le deuxième cas :

- "Il n'y a pas de contexte sans contexte", ceci dit, le contexte n'existe pas seul, il émerge pour une finalité bien précise (Brézillon, 2002).
- "Le contexte est un ensemble d'informations", cet ensemble est structuré et partagé ; il évolue et sert l'interprétation (Winograd, 2001).

Les définitions proposées dans la littérature en RI diffèrent essentiellement par leurs éléments constitutifs. Mais nous retrouvons toutefois des dimensions descriptives communes telles que l'environnement cognitif, le besoin mental en information, l'interaction liée à la recherche d'information, la tâche de recherche, les préférences de recherche liées à la qualité de l'information (fraîcheur de l'information, genre du document, crédibilité de la source de l'information, etc.), les ressources disponibles, etc. Plus rarement, nous observons l'utilisation des éléments concernant les émotions, les états d'esprit et les données culturelles (Brézillon and Marquois, 2004).

De façon générale, les définitions du contexte en RI, proposées dans la littérature, portent principalement sur des facteurs ou des caractéristiques liés à différentes entités impliquées dans le processus de recherche. Nous distinguons principalement trois entités qui interviennent dans la RI contextualisée :

- L'utilisateur : il peut être représenté par ses connaissances, ses buts et ses intentions de recherche d'information ;
- Le système : il se modélise par des caractéristiques liées au processus de recherche tel que : le temps de réponse, le cout, etc ;
- L'environnement de recherche : il présente des caractéristiques liées à des critères sociaux, organisationnels et situationnels.

Si le contexte renvoie à un concept "statique" décrit par plusieurs dimensions, la contextualisation renvoie à un processus de construction qui, lui aussi, a donné lieu à de nombreuses propositions. Par exemple :

"Contextualizing is a process of adding new meaning to a given situation in order to characterize this situation in terms of what could (or should) be done, and by the same token to exclude (for the time being) alternative interpretations of the required mode of acting" (Van Oers, 1998) ;

ou plus récemment, "Contextualization is the process by which people interpret, communicate, and apply the Bible within a particular cultural context" (Wu, 2013).

Mais, qu'il s'agisse du cadre contextuel ou de sa mise en oeuvre, la difficulté majeure d'un point de vue opérationnel est de cerner les différentes dimensions en jeu et les relations qu'elles entretiennent.

2.4 Taxonomies des facteurs impliqués

Pour mieux appréhender les facteurs contextuels nécessaires à intégrer dans les systèmes de RI contextuels, on peut s'appuyer sur différentes tentatives de taxonomies (Dey and Abowd, 1999; Cool and Spink, 2002; Göker and Myrhaug, 2002; Bricon-Souf and Newman, 2007; Ingwersen and Järvelin, 2005; Mansourian, 2008; Tamine-Lechani et al., 2010). Nous les présentons synthétiquement ci-dessous selon un ordre chronologique qui rend compte à la fois de la variété des dimensions mises en oeuvre mais aussi de leur importance.

2.4.1 Taxonomie de Dey 1999

La classification proposée par (Dey and Abowd, 1999) repose sur les dimensions dans lesquelles le contexte est évalué. Elle définie trois dimensions : les "endroits" comme les régions de l'espace géographique, tels que chambres ou bureaux, les "personnes" y compris des individus ou des groupes, co-localisés ou distribués, et les "objets" comme par exemple des objets physiques, des composants logiciels et des fichiers informatiques.

2.4.2 Taxonomie de Fuhr 2001

De nouvelles caractéristiques liées à l'aspect temporel du besoin en information d'une part et au type de recherche demandé d'autre part sont ajoutées dans la définition de (Fuhr, 2001). Le contexte multidimensionnel possède trois principales dimensions : sociale, applicative et temporelle.

1. La dimension sociale définit la composante d'appartenance possible de l'utilisateur à un groupe ou à une communauté, ou la non appartenance, ce qui souligne un aspect individuel.
2. La dimension applicative définit le contexte applicatif du besoin exprimé : recherche adhoc, résolution de problème ou workflow.
3. La dimension temporelle permet de décrire les éléments temporels relatifs au besoin exprimé : temps passé, instant courant ou à court terme (interactivité), intention ou long terme (personnalisation).

   Sous l'angle de la dimension temporelle, nous pouvons distinguer deux types de contextes avec des démarches de personnalisation appropriées : le contexte courant ou à court terme est associé à des besoins et à des préférences instantanées de l'utilisateur lors d'une session d'une recherche et contexte à long terme traduit les besoins à long terme de l'utilisateur sur diverses sessions de recherche qui pourront être intégrées dans un processus de personnalisation.

2.4.3 Taxonomie de Cool 2002

(Cool and Spink, 2002) définissent quatre dimensions considérées les plus significatives en RI afin de dissocier les entités intervenant dans le processus de recherche :

1. L'environnement de recherche lié aux facteurs cognitifs, sociaux ou professionnels qui influencent le comportement de recherche de l'utilisateur et sa perception de la pertinence ;
2. Les connaissances de l'utilisateur, ses buts et ses intentions de recherche ;
3. L'interaction utilisateur-système. Cette dimension met en évidence l'impact des situations ou de l'environnement sur la rétroaction ou les jugements de pertinence de l'utilisateur ;
4. Le niveau linguistique explore la performance du Système de Recherche d'Information (SRI) dans l'interprétation des requêtes des utilisateurs et leur habilité à les désambiguïser.

2.4.4 Taxonomie de Goker et al. 2002

(Göker and Myrhaug, 2002) présentent une taxonomie organisée selon cinq dimensions principales.

1. La dimension tâche concerne l'utilisateur, ses objectifs, ses tâches et ses activités ;
2. La dimension sociale se réfère aux aspects sociaux de l'utilisateur, tels que des informations sur les amis et la famille ;
3. La dimension du contexte personnel regroupe les informations mentales et physiques sur l'utilisateur telles que l'humeur et l'expertise ;
4. La dimension spatio-temporelle concerne des attributs comme le temps et le lieu ;
5. La dimension de l'environnement concerne des composantes variées de l'environnement de l'utilisateur telles que les pensées, la lumière et l'information accessible par l'utilisateur.

2.4.5 Taxonomie de Briconsouf et al. 2007

Cette taxonomie a été principalement définie dans le cadre d'applications liées à la santé mais ses composantes peuvent se généraliser à d'autres domaines. Le contexte est ici défini selon trois dimensions principales :

1. Le but de l'utilisation du contexte qui reprend les trois types de contexte d'utilisation proposés par (Dey, 2001) ;
2. Les éléments de la représentation du contexte identifie trois classes principales (les personnes, l'environnement et les activités) ;
3. L'organisation des caractéristiques du contexte renvoie à différents modèles tels que l'organisation hiérarchique (du général aux aspects locaux du contexte), une organisation en fonction de la dimension interne et externe du contexte et une organisation en fonction de l'utilité du contexte (pertinent ou non pertinent).

2.4.6 Taxonomie de Mansourian 2008

Cette taxonomie s'organise autour de cinq dimensions principales :

1. La dimension utilisateur qui regroupe des concepts divers : sentiments, pensées (attitudes et style cognitif), actions (utilisateurs passifs vs actifs) et compétences ;
2. L'outil de recherche et le sujet de recherche ;
3. La situation de la recherche qui contient le lieu de la recherche, le type de la recherche (lié au travail ou à la recherche de la vie quotidienne), l'immédiateté de la recherche et son importance ;
4. Les ressources d'information qui renvoient à la consultation, l'accessibilité, le niveau de prestation (la disposition du public d'accès/restreint) et le niveau de convivialité.

2.4.7 Taxonomie de Tamine et al. 2010

La figure 2.1 représente les quatre dimensions décrites dans la taxonomie proposée par (Tamine-Lechani et al., 2010). Parmi les dimensions proposées, nous précisons celles qui sont les plus importantes pour notre travail qui portent sur le contexte utilisateur et le contexte document.

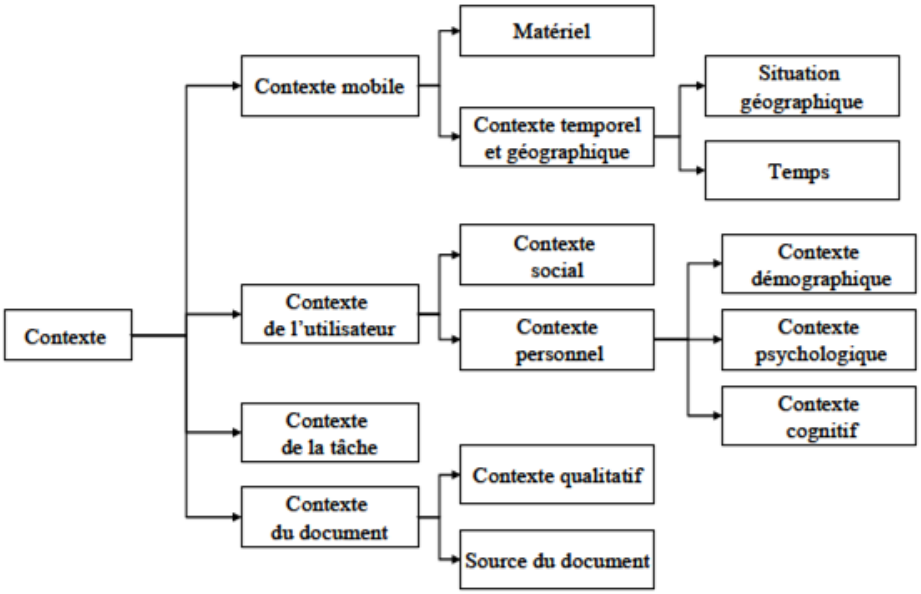

FIGURE 2.1 – Dimensions du contexte multidimensionnel de Tamine et al. 2010 (Tamine-Lechani et al., 2010)

1. **Contexte de l'utilisateur**
   Le contexte de l'utilisateur est la dimension centrale en RI contextuelle. Cette dimension contient deux sous-dimensions liées respectivement au contexte personnel de l'utilisateur et son environnement social.
   - **Contexte personnel** : ce contexte est associé aux sous-dimensions suivantes :
     * Contexte démographique : ce contexte désigne des préférences personnelles comme le langage préféré des résultats de recherche ou le genre du document utilisé dans le but d'adapter les résultats de recherche aux préférences de l'utilisateur (Hupfer and Detlor, 2006; Frias-Martinez et al., 2007).
     * Contexte psychologique : ce contexte désigne des caractéristiques affectives de l'utilisateur tels que l'anxiété et la frustration qui influencent son comportement de recherche ainsi que ses jugements de pertinence (Bilal, 2000; Kim, 2008).
     * Contexte cognitif : c'est la sous-dimension la plus abordée dans le domaine, et désigne le niveau d'expertise de l'utilisateur, ses centres d'intérêts qui peuvent être à court ou à long terme.
   - **Contexte social** : ce contexte désigne une communauté d'utilisateurs comme des amis, des collègues ou des voisins. Selon la dimension sociale, l'adaptation du processus de RI consiste à retourner de l'information qui répond aux préférences de la communauté des utilisateurs plutôt que de répondre aux préférences d'un seul utilisateur (Lang, 1995; Smyth and Balfe, 2006). Le contexte social est souvent exploité dans les systèmes de recommandations (Leung et al.,

2006) basé sur le filtrage collaboratif, utilisé sous le nom du profil collaboratif. Actuellement, plusieurs travaux (Kazai and Milic-Frayling, 2009; Kazai and Milic-Frayling, 2008) exploitent le contexte social en RI (Smith et al., 2008; Yahia et al., 2008).

2. **Le contexte du document**

Le contexte du document est issu du principe de la poly-représentation présenté par (INGWERSEN, 1996). Le principe de la poly-représentation est basé sur l'hypothèse que les documents peuvent être classifiés dans un espace à variables multiples (structure du document, genre du document, style du contenu, structure des hyperliens, etc.) et que cette classification augmente la probabilité que ces documents soient utilisés dans certaines situations de recherche (Ingwersen and Järvelin, 2005).

Trois sous-dimensions sont identifiées sous le contexte du document :

- La première dimension concerne la représentation du document tels que les éléments structurels, la forme, les citations et des méta-données (Tombros et al., 2005).
- La deuxième dimension concerne les caractéristiques de la source des données (Xie, 2008), telles que sa crédibilité, sa fiabilité, etc. et leur perception par les utilisateurs.
- La troisième dimension concerne la qualité de l'information. En effet, la mesure de pertinence de l'information ne dépend pas seulement du contenu des données mais aussi de sa qualité. En général, la qualité de l'information concerne sa fraîcheur, sa précision, sa cohérence, sa complétude, sa sécurité, etc.

2.5 Contextualisation des tweets

Comme nous l'avons indiqué dans l'introduction, Twitter représente un canal fécond et instantané de publications. Mais la limite de ses 140 caractères conduit à l'utilisation d'un vocabulaire particulier qui peut rendre le contenu des tweets peu explicite ; ce qui rend leur contextualisation particulièrement utile pour la poursuite de l'essor de Twitter. Cela a motivé la proposition du laboratoire CLEF issu du challenge INEX qui vise à fournir aux lecteurs des tweets des informations pertinentes qui facilitent leur compréhension. Dans le cadre du challenge, l'information est fournie au lecteur sous la forme d'un résumé de 500 mots, construit à partir d'extraits de ressources textuelles, spécifiquement de Wikipédia. L'exemple suivant illustre le type de résultat produit pour un tweet considéré comme ambigu.

Le tweet : "Bobby Brown-Fighting #WhitneyHouston's Family to See Bobbi Kristina"

Le résumé : La figure 2.2 représente le résumé fourni pour contextualiser ce tweet. Un tel résumé devrait fournir quelques informations générales sur tous les éléments du tweet (Bellot et al., 2016).

Ce recours à des sources d'informations exogènes pour la contextualisation des tweets est commun à la plupart des approches (Deveaud and Boudin, 2013a; Abel et al., 2011). Nous présentons ci-dessous les sources les plus utilisées.

2.5.1 Sources d'information pour la contextualisation des tweets

Wikipédia représente la source la plus exploitée dans les travaux de la littérature. Mais elle peut être complétée par d'autres sources, notamment : DBpedia, les sites de presse et Twitter lui-même par extraction de tweets similaires au tweet cible à contextualiser.

Whitney Elizabeth Houston (August 9, 1963 February 11, 2012) was an American recording artist, actress, producer, and model. Houston was one of the world's best-selling music artists, having sold over 170 million albums, singles and videos worldwide. Robert Barisford "Bobby" Brown (born February 5, 1969) is an American R&B singer-songwriter, occasional rapper, and dancer. After a three-year courtship, the two were married on July 18, 1992. On March 4, 1993, Houston gave birth to their daughter Bobbi Kristina Houston Brown, her only child, and his fourth. With the missed performances and weight loss, rumors about Houston using drugs with her husband circulated. Following fourteen years of marriage, Brown and Houston filed for legal separation in September 2006. Their divorce was finalized on April 24, 2007, with Houston receiving custody of their then-14-year-old daughter. On February 11, 2012, Houston was found unresponsive in suite 434 at the Beverly Hilton Hotel, submerged in the bathtub.

FIGURE 2.2 – Exemple de résumé de contextualisation pour un tweet ambigu

– **Wikipédia** [1]

C'est une encyclopédie en ligne universelle généraliste lancée en 2001, éditée en collaboration, et devenue la plus grande édition linguistique avec 45 millions articles dans 295 langues [2], dont la majorité sont des entités (personnes, lieux, événements). La version anglaise est la plus riche avec 5,4 millions d'articles décrivant chacun une entité ou un concept. Généralement, le texte de l'article commence par une définition, un résumé ou une brève description de l'entité. Les avantages pouvant être apportés par une telle ressource sont principalement liés au fait que les articles sont entièrement construits manuellement. L'information, par conséquent, est construite avec la participation de plusieurs personnes. Derrière cette source d'information se trouve de la connaissance des humains qui ont pris le temps d'acquérir et de retranscrire afin d'améliorer en permanence la qualité des informations. Wikipédia est une source riche de connaissances et une cible commune pour les travaux de contextualisation. La quasi-totalité des approches de contextualisation ont utilisé Wikipédia comme source d'information. Néanmoins, cette source ne garantit pas la "fraîcheur" des informations et n'est donc pas toujours bien adaptée à l'aspect temporel et social des tweets. Les situations d'urgence ou les événements d'actualité peuvent donc représenter un problème d'accès à l'information dû au manque d'actualisation de la base de connaissances Wikipedia.

– **DBpedia** [3]

DBpedia est un ensemble de données "Linked Open Data (LOD)" [4]. Les données sur DBpedia sont extraites à partir des infoboxes Wikipédia. Elles sont structurées sous forme de triplets RDF (Resource Description Framework) et sont accessibles sur le Web (Lehmann et al., 2015). La version anglaise de cette base de connaissances comporte 6 millions [5] entités, dont 5.2 millions sont classées d'une manière cohérente sous forme des ontologies. Cette source gigantesque de connaissances a plusieurs avantages sur les bases de connaissances existantes : elle couvre de nombreux

[1] https ://fr.wikipedia.org/
[2] https ://stats.wikimedia.org
[3] http ://wiki.dbpedia.org/
[4] LOD est un projet collaboratif visant à publier des jeux de données ouverts et liés sur le Web en respectant les principes du Web sémantique édictés par Tim Berners-Lee (Bizer et al., 2009)
[5] http ://wiki.dbpedia.org/Datasets/DatasetStatistics

domaines et elle est multilingue. Néanmoins, comme pour Wikipedia, on retrouve pour cette source la difficulté d'intégration des informations en temps réel. Cette source ne peut pas supporter les spécificités des plateformes de microblogs, où les informations diffusées sont courtes et publiées en temps réel. Une étape nécessaire pour l'acquisition de l'information sur les microblogs est d'identifier les entités présentes dans les messages. Une telle identification peut être difficile en utilisant cette source parce que les entités mentionnées peuvent être ambiguës.

– **Les sites de presse et d'actualité**
  Les sites de presse et d'actualité, dont les plus utilisés dans la littérature sont ceux de BBC, de CNN et du New York Times, représentent des sources d'information traditionnelles qui ont été sollicitées pour améliorer et faciliter la compréhension des tweets (Guo et al., ; Abel et al., 2011). Ils couvrent différentes catégories d'actualité telles que la politique, le sport ou la culture et sont disponibles via le flux RSS ou via leur site Web. En complément, les médias sociaux tels que Twitter fournissent un flux constant d'informations qui est utilisé par beaucoup comme une source de nouvelles, en particulier dans les événements inattendus (actualité).
– **Les tweets similaires**
  En plus de son apport pour l'accès à l'information en continu, Twitter fournit également une source de données qui peut être utilisée pour la contextualisation : (Guo et al., 2013) ont proposé d'exploiter les tweets similaires pour l'expansion du contexte. Ils utilisent des tweets supplémentaires qui sont liés à un tweet donné dont le contenu est ambigu et insèrent des caractéristiques explicites pour la désambiguïsation.

2.5.2 Approches de contextualisation des tweets

La très grande majorité des approches développées pour la contextualisation des tweets se sont basées sur la recherche et l'extraction de passages de corpus exogènes provenant de sources telles que celles citées précédemment en utilisant des algorithmes de modélisation thématique (Genc et al., 2011) et de reconnaissance des entités nommées (Ermakova and Mothe, 2012a). De façon générale, ces approches peuvent être regroupées en trois catégories. La première, qui est la plus utilisée, contient les approches combinant les techniques de Recherche d'Information (RI) avec les techniques de Résumé Automatique (RA), pour fournir le contexte d'un tweet ambigu. La deuxième concerne les approches qui ont utilisé les techniques de modélisation thématique et l'enrichissement sémantique des tweets et la troisième regroupe les approches basées sur l'expansion des tweets. Nous les précisons ci-dessous.

2.5.2.1 Approches combinant la recherche d'information et le résumé automatique

La construction d'un résumé automatique, souvent issu d'informations provenant de Wikipédia, est l'approche certainement la plus présente dans la littérature (ex : (Deveaud and Boudin, 2013a; Ansary et al., 2013; Ganguly et al., 2012; Ermakova and Mothe, 2012a)). Selon (Saggion and Lapalme, 2002), un résumé est défini comme une "version condensée d'un document ayant un genre et un but spécifique : donner au lecteur une idée exacte et concise du contenu du document source (original)". Nous distinguons classiquement deux types de résumés : (i) le résumé extractif résulte d'un processus d'extraction des phrases pertinentes dans un corpus de documents, (ii) le résumé abstractif peut contenir des phrases qui n'existent pas dans le corpus. Pour Twitter, la majorité des approches produisent un résumé extractif. Quelques expérimentations ont néanmoins été conduites avec

des résumés abstractifs (Olariu, 2013) dont la construction repose sur un regroupement préalable de messages selon une mesure de similarité. Dans la suite, nous rappelons les principes de quelques méthodes classiques pour déterminer la pertinence d'une phrase pour la construction de résumés extractifs. Les approches historiquement les plus fréquentes sont non supervisées mais l'apprentissage automatique a connu un grand essor cette dernière décennie.

Méthodes basées sur la fréquence des mots Les premiers travaux, menés par (Luhn, 1958),ont exploré l'utilisation de la fréquence lexicale des mots comme un indicateur de leur importance. Cette fréquence lexicale se déduit de la distribution de la "probabilité des mots" dans le document. Cependant, tous les mots n'ont pas la même valeur informative. Différentes mesures ont été proposées pour surmonter ce problème. La méthode de pondération "TF*IDF" permet d'ajuster le poids des mots de sorte que les mots ayant les poids les plus élevés sont les plus susceptibles d'être les plus descriptifs. La méthode de "Log-likelihood ratio" associe des "mots signatures" à chaque sujet ; ils permettent de filtrer les phrases. Nous les précisons ci-dessous :

– **La probabilité des mots**
  C'est la méthode la plus simple à utiliser pour estimer l'importance d'un mot en calculant sa fréquence. La probabilité *p(w)* d'un mot *w*, est calculée à partir de l'entrée, qui peut être un groupe de documents ou un seul document. Elle est calculée comme le nombre d'occurrences d'un mot *c(w)*, divisé par le nombre de tous les mots d'entrée *N*.

$$\mathrm{p(w) = c(w)/N.}$$

– **TF*IDF mesures (Term Frequency-Inverse Document Frequency)**
  Une limitation forte de l'usage de la probabilité des mots est due à la distribution de la fréquence des mots dans un corpus qui suit souvent une loi de Zipf : de nombreux mots ont une fréquence faible et peu ont une fréquence élevée ; et les mots les plus fréquents sont des mots peu informatifs comme les déterminants, les prépositions et les verbes auxiliaires. La détermination d'un seuil de sélection des mots utiles pour une variété de documents n'est pas une question triviale et différentes approches ont tenté de mesurer la représentativité des termes. La pondération des mots TF*IDF est la méthode la plus utilisée dans la Recherche d'Information. Elle permet d'évaluer l'importance d'un mot contenu dans un document par rapport à une collection des documents ou un corpus.
  TF= la fréquence *c(w)* d'un mot w est le nombre d'occurrence de ce mot dans le document considéré *d* ; *IDF*= est une mesure de l'importance du terme dans l'ensemble du nombre total de documents *D* dans le corpus.

$$TF * IDF_w = c(w) * log\frac{D}{d(w)} \tag{2.1}$$

  Les mots descriptifs sont ceux qui apparaissent souvent dans un document avec un poids élevé mais ne sont pas très communs dans d'autres documents. En outre, les mots "vides" apparaissent pratiquement dans tous les documents, de sorte que leur poids sera proche de zéro. Ces propriétés, associées à sa facilité de calcul, expliquent pourquoi TF*IDF est l'une des mesures les plus utilisées pour l'extraction de résumé (Erkan and Radev, 2004; Filatova and Hatzivassiloglou, 2004; Fung and Ngai, 2006; Galley, 2006; Hovy and Lin, 1999).

– **Test du ratio log-vraisemblance (Log-likelihood ratio test for topic signatures)**
La distribution de la fréquence des mots est également utilisée dans une autre approche populaire basée sur le log de vraisemblance (Dunning, 1993). Contrairement à la mesure TF*IDF, le test du ratio log-vraisemblance est utilisé pour déterminer si un mot est indicatif du sujet des documents d'entrée. Ces mots descriptifs ont été appelés "signatures thématiques" et sont explorés et utilisés pour le résumé automatique (Lin and Hovy, 2000). Lin et Hovy définissent les signatures thématiques comme un "ensemble de mots généralement fortement corrélés, avec des poids associés, organisés autour d'un sujet cible dans le corpus" qui identifie la présence de concepts complexes dans les documents. Par exemple : le concept "visite-restaurant", implique au moins les concepts de menu, de manger, de payer et éventuellement de serveurs.

Pour calculer la fréquence d'occurrence d'un mot dans un large corpus de données, il est nécessaire de déterminer une base de "signatures thématiques".

Une base de signature thématique pour un sujet donné est créée comme suit :

1. Classer les documents comme pertinents ou non pertinents selon un sujet donné ;
2. Calculer la valeur du ratio de vraisemblance pour chaque terme dans la collection de documents ;
3. Classer les termes en fonction de leur valeur du ratio de vraisemblance ;
4. Sélectionner un niveau de seuil en utilisant la table de distribution X^2 pour déterminer le nombre de termes à inclure dans les signatures.

Les "signatures thématiques" sont plus puissantes que l'utilisation directe de la fréquence, car elles donnent un seuil pour décider quels mots doivent être considérés, sur la base d'une distribution de probabilité (Gupta et al., 2007).

Approches de regroupement des phrases (Sentence clustering) Pour limiter la présence de phrases similaires dans le document, une approche de regroupement des phrases a été proposée (McKeown et al., 1999; Hatzivassiloglou et al., 2001; Siddharthan et al., 2004). Les phrases similaires sont classées dans une même classe et un résumé peut alors s'appuyer sur l'extraction d'une phrase représentative de chaque classe. Les approches proposées ont été basées sur un partitionnement de l'ensemble des phrases qui ne permet pas de prendre en compte les sens multiples ; mais des extensions qui autorisent les chevauchements pourraient être mis en oeuvre.

Méthodes de résumé à base de graphes Contrairement aux approches précédentes qui considèrent comme entités uniquement les phrases, les méthodes à base de graphes pour le classement de phrases exploitent les répétitions au niveau des mots et des phrases dans les documents d'entrée. Ainsi, elles combinent les avantages des méthodes basées sur la fréquence des mots et celles basées sur des classifications de phrases (Qazvinian and Radev, 2008). Elles décrivent les relations entre les différentes entités par un graphe où les sommets représentent les phrases d'un corpus et les poids associés aux arêtes une similarité entre phrases. La similarité généralement choisie est la "similarité cosinus". Ainsi, les poids sont élevés pour des phrases qui partagent de nombreux éléments lexicaux. Sur ce graphe, l'importance des sommets -c.à.d des phrases- peut être calculée via des algorithmes classiques développés en analyse de réseaux tels que Page Rank (Erkan and Radev, 2004; Mihalcea and Tarau, 2004). Ces approches semblent parmi les plus efficaces. Ainsi, la considération de la similarité syntaxique dans l'étape de construction du graphe a conduit à des résultats meilleurs (Mihalcea and Tarau, 2004).

2.5.2.2 Approches de modélisation thématique de microblogs

Afin de surmonter le problème de l'ambiguïté et de l'absence de l'aspect sémantique dans les textes courts (Genc et al., 2011; Michelson and Macskassy, 2010), plusieurs travaux (Banerjee et al., 2007; Meij et al., 2012; Morchid et al., 2013; Guo et al., 2013) se sont basés sur l'enrichissent sémantique en utilisant des informations provenant de sources telles que celles citées précédemment dans le paragraphe 2.5.1. Dans ce cadre, parmi les techniques utilisées, on trouve notamment la classification par modèle de mélanges via l'allocation latente de Dirichlet. Cette approche permet d'associer les tweets à un ensemble de thèmes. D'autres travaux ont renforcé l'exploitation des sources d'informations par les informations concernant les activités des utilisateurs sur Twitter afin d'enrichir la sémantique des tweets (Abel et al., 2011).

2.5.2.3 Approches utilisant la technique d'expansion de microblogs

La troisième catégorie concerne les approches utilisant la technique d'expansion de messages. Dans ce qui suit, nous présenterons ces approches.

Expansion de requêtes L'expansion de requête est une technique utilisée en Recherche d'Information(RI). Elle sert à ajouter des informations supplémentaires à la requête de recherche. L'objectif est de retourner à l'utilisateur des documents qu'il n'aurait pas trouvés avec sa requête initiale. En outre, les utilisateurs n'emploient habituellement que quelques mots ; généralement moins de 5 selon (Jansen et al., 1998) pour décrire leurs requêtes, ce qui peut donner lieu à des ambiguïtés. Il a été montré que l'enrichissement de la requête pouvait améliorer les résultats de recherche. Face à la quantité croissante d'informations disponibles sur les sites de microblogging, cette technique est susceptible d'être bénéfique pour ce type d'information (Bandyopadhyay et al., 2011), vu que le vocabulaire limité lié à la taille réduite des microblogs représente un obstacle majeur pour restituer des microblogs pertinents en utilisant les SRI.

Expansion de hashtags Le symbole #, appelé hashtag (ou mot-dièse en français), est un caractère utilisé pour préfixer les mots-clés dans les tweets. Les réseaux sociaux utilisent largement cette installation pour indiquer des mots, des groupes ou des sujets pertinents. Depuis décembre 2008 à octobre 2009, les hashtags sont devenus largement adoptés sur le site de Twitter, qui permettent de catégoriser les tweets en sujets de discussion. Les hashtags représentent un moyen de catégorisation et d'étiquetage et sont ainsi des marqueurs d'informations importantes directement fournis par l'auteur. Plusieurs études (Kwak et al., 2010; Mendes et al., 2010; Deveaud and Boudin, 2013b; Vivaldi and da Cunha, 2013; Laniado and Mika, 2010) sur ce nouveau type de marquage, ont montré l'utilité et l'apport des hashtags pour la contextualisation des tweets. Ainsi, les auteurs de ces études ont supposé que les hashtags sont principalement utilisés pour "contextualiser" les tweets, une hypothèse que nous adoptons dans notre travail, qui permet d'ajouter la sémantique aux tweets avec l'utilisation des définitions de hashtags.

Exploitation des URLs Les utilisateurs de Twitter partagent des URLs dans leurs tweets pour attirer l'attention de leurs suiveurs sur une nouvelle information contenue dans une page Web, souvent pas encore indexée par les moteurs de recherche classiques. Ces pages représentent ainsi de l'information enrichissante par rapport au contenu du tweet. Les

URLs améliorent remarquablement la qualité des tweets. Leur présence reflète la pertinence du tweet (Cheng et al., 2013). En effet, un tweet contenant des URLs est plus informatif car il apporte une information additionnelle portée par l'URL. Généralement, le contenu des URLs est utilisé pour enrichir le vocabulaire des tweets, limités en longueur. D'autre part, les URLs ont été utilisés pour mesurer la centralité des tweets, ainsi que leur fiabilité (Ravikumar et al., 2012). En outre, la présence des URLs dans les tweets a montré un fort impact dans l'amélioration de la qualité des résultats (Damak et al., 2013).

2.6 Discussion : positionnement de notre méthode de contextualisation par rapport à la littérature

Comme nous venons de le présenter, les approches proposées pour la contextualisation des tweets sont diverses mais la plupart structurent la tâche de contextualisation autour de trois phases (Bhaskar et al., 2012) :

1. Analyse des tweets,
2. La recherche et l'identification des passages pertinents à partir des documents Wikipédia,
3. Construction du résumé court à partir de ces documents.

La première phase est généralement basée sur un ensemble de pré-traitement appliqué sur les tweets afin de formater le contenu et construire la requête pour la recherche d'information. Dans cette phase de pré-traitement, les hashtags sont considérés comme l'un des principaux sujets du tweet et comme des requêtes courtes utilisées dans la recherche d'information. En effet, les hashtags ont été très exploités dans plusieurs travaux de contextualisation des tweets pour la recherche et l'identification des documents liés au tweet (Deveaud and Boudin, 2013b; Vivaldi and da Cunha, 2013). Un autre type d'information est aussi exploité dans les pré-traitements de tweets : les mentions. Une mention dans le texte d'un tweet peut être considérée comme l'auteur du tweet, ou un autre utilisateur lié au tweet. Ces deux types d'informations sont utilisés, dans les travaux de la littérature, pour améliorer les requêtes utilisées pour récupérer des documents de Wikipédia (Pontes et al., 2016). Néanmoins, les mentions ne sont pas présentes dans Wikipédia, ce qui peut poser un problème d'accès aux informations. Vue leur importance confirmée par les travaux précités, dans notre proposition, nous considérons également les "hashtags" et les "mentions" dans la formulation des requêtes de construction des conversations.

La deuxième phase du processus consiste à rechercher et récupérer les documents issus de Wikipédia les plus pertinents pour la requête (Schiffman et al., 2007; Pakray et al., 2011). Le choix de Wikipédia s'argumente par la popularité de la source et sa fiabilité suffisante. Dans notre proposition, nous n'utilisons pas Wikipedia mais nous exploitons, comme source d'information, les discussions sur Twitter. Notre approche est complémentaire. Elle permet d'intégrer dans la contextualisation des informations sociales absentes de Wikipedia et des informations récentes mises à jour en temps réel.

Dans la troisième phase, la quasi-totalité des approches de contextualisation, proposées dans la littérature, se sont focalisées sur la recherche et l'extraction de passages -généralement des phrases-. Elles se basent donc sur des approches d'extraction visant à sélectionner les phrases pertinentes liées au tweet en évaluant la pertinence par une mesure de similitude entre la requête et les phrases candidates. Ces approches fournissent en sortie des scores

aux documents candidats et construisent le résumé à partir du document ayant le score le plus élevé. Comme indiqué ci-dessus, notre proposition ne s'appuie pas sur la sélection de phrases mais d'éléments (liens ou noeuds) dans un graphe de relations décrivant une conversation. La sélection est basée sur des mesures caractérisant des propriétés structurelles du graphe. Et la sortie du processus n'est pas un résumé textuel mais une restitution visuelle d'un ensemble d'informations "pertinentes" extraites de l'analyse des graphes de conversations.

2.7 Conclusion

L'analyse de différentes approches de contextualisation proposées dans la littérature nous a permis de préciser le lien entre la contextualisation des tweets et l'analyse des graphes de conversations sur Twitter. Dans le chapitre suivant, nous nous focalisons sur l'analyse des conversations sur Twitter en abordant les différentes définitions de graphes, les principales étapes des algorithmes de construction et la présentation des indicateurs d'analyse des graphes de conversations.

3

Conversations sur Twitter

Sommaire

3.1 Introduction 21
3.2 Objectifs majeurs de l'analyse des conversations sur Twitter . . . 22
3.3 Définitions des conversations 23
3.4 Méthodes de construction des conversations 24
3.4.1 Méthodes de construction à base de relations 24
3.4.2 Méthodes de construction à base de seuil 26
3.4.3 Méthodes de construction à base de classification 27
3.5 Les indices d'analyse des conversations 28
3.5.1 Notations 29
3.5.2 Indices structurels sur les graphes de conversations 29
3.5.3 Indices associés aux contenus des conversations 30
3.5.4 Indices associés aux participants des conversations 31
3.6 Conclusion 33

3.1 Introduction

De façon générale, les nouveaux services associés aux médias sociaux qui se sont popularisés dans la dernière décennie, ont profondément modifié la manière dont les utilisateurs communiquent et partagent l'information ; ils offrent la possibilité de communiquer, d'interagir, de collaborer et de créer des conversations avec d'autres utilisateurs, en répondant à partir de messages en temps réel sur les événements, le quotidien et les tendances. Twitter est un exemple paradigmatique de ces nouvelles formes de conversation et son expansion a stimulé le développement de recherches portant sur l'analyse de ces dernières.

Après une présentation synthétique des différents objectifs des études récentes sur les conversations associées à Twitter, ce chapitre se focalise sur l'analyse de la structure des

relations entre les différentes composantes en jeu. Il rappelle les différentes définitions de graphes qui ont été proposées dans la littérature pour modéliser ce système de relations, puis il décrit les principales étapes des algorithmes de construction. La dernière partie du chapitre est consacrée à la présentation des indicateurs qui permettent de caractériser des propriétés, essentiellement structurelles, sur les graphes de conversations. Ces propriétés nous servent ensuite de guide pour l'extraction d'éléments de contexte pertinents.

3.2 Objectifs majeurs de l'analyse des conversations sur Twitter

Les principaux objectifs de l'analyse des conversations sur Twitter peuvent s'organiser selon quatre grandes directions qui exploitent chacune des informations spécifiques : (1) l'analyse structurelle qui porte sur la combinatoire des relations, (2) l'analyse des patterns linguistiques centrée sur le contenu sémantique des échanges, (3) l'analyse des opinions qui tente d'associer des informations linguistiques spécifiques à des catégories d'opinion souvent pré-établies, et (4) la recherche d'informations qui remplacent la réponse classique à une requête de type document par des échanges de messages sur Twitter.

1. **Analyse structurelle des conversations**

   Au cours des dernières années, de nombreux travaux se sont penchés sur l'étude des structures et des propriétés des conversations (Cogan et al., 2012; Nishi et al., 2016; Kumar et al., 2010; Honey and Herring, 2009; Russell et al., 2011; Greetham and Ward, 2014; Romero et al., 2011).

   Dans les travaux de (Tamine et al., 2015; Purohit et al., 2013), les auteurs présentent une analyse des caractéristiques de collaboration des conversations associées à des catastrophes naturelles ou industrielles et des événements de grande envergure. Ils ont identifié les modèles structurels et les patrons de coordination entre les participants aux conversations dont l'objectif est de mieux comprendre les formes de collaboration dans les situations de crise.

   D'autres auteurs (Romero et al., 2011; Bliss et al., 2011) ont travaillé sur la prédiction des liaisons entre les participants. Ils se sont intéressés à étudier l'utilisation du même hashtag pour prédire la réciprocité entre les participants.

2. **Analyse Linguistique : Recherche des patterns linguistiques**

   Dans cette famille, différents auteurs (Danescu-Niculescu-Mizil et al., 2011; Doyle and Frank, 2015) se sont intéressés à analyser les styles linguistiques présents dans les conversations afin de définir les modèles de coordination entre les participants et d'identifier des variations linguistiques.

3. **Analyse d'opinion et de sentiments**

   Cette famille regroupe les travaux qui portent sur l'analyse des sentiments et d'opinions. Dans (Kim et al., 2012), les auteurs ont travaillé sur la découverte et la classification des modèles d'émotion des participants dans les conversations sur Twitter. Ils ont montré que les participants de la même conversation partagent une émotion commune. Mathew et al., 2016 ont analysé les sujets des conversations portant sur les événements promotionnels de E-commerce afin de capturer les opinons des participants et les classer sur la base d'une analyse de sentiment (Mathew et al., 2016).

4. **Recherche d'informations et recommandations**

   Quelques travaux visent à répondre à des requêtes de recherche par des résultats associés à des conversations sociales (Magnani et al., 2012; Herrera et al., 2015; Amer et al., 2015). La figure 3.1 donne un exemple de résultat fourni en réponse à la requête "Francois Mitterrand". D'une façon générale, ces travaux se basent sur l'agrégation de caractéristiques textuelles et de facteurs sociaux. Leurs auteurs ont montré que l'utilisation des facteurs sociaux pouvait améliorer les résultats obtenus, mais ils n'ont pas mesuré l'apport de ces facteurs par rapport à la pertinence thématique. D'autres travaux visent à recommander des conversations aux utilisateurs (Chen et al., 2011) en exploitant les contenus et les propriétés structurelles.

Temps de publication	Nom de l'auteur de message	Le message	Nom du récepteur de message
10:31	@LeaSalame	Il sort rarement de sa réserve. Reecoutez Michel Charasse sur Francois Mitterrand, l'Histoire, et les politiques aujourd'hui	-------
10:46	@lpenet	@LeaSalame le "Pourquoi Mitterrand ?" de Pierre Joxe est également instructif et émouvant https://fr.wikipedia.org/wiki/	@LeaSalame
11:54	@jn31200	@LeaSalame @franceinter personne ne se plaignait du silence de charasse	@LeaSalame @franceinter
11:43	@Gerard_Clech	@LeaSalame Charasse et Dumas ont depuis longtemps fait une OPA sur #Mitterrand mais c'était bien un homme de droite... extrême même !	@LeaSalame
11:42	@apantzer	@LeaSalame @franceinter il est temps de tourner la page marre de l histoire de hier	@LeaSalame @franceinter
11:30	@AymPasdutout	@LeaSalame @franceinter j'adore votre petit sourire quand vous posez une question piège ? Ou du mois qui mérite une vraieréponse #Mitterrand	@LeaSalame @franceinter
11:28	@AtonAha	@LeaSalame Très bon.	@LeaSalame
11:27	@BILBAO83	@LeaSalame @franceinter alors finalement ce sont des comptables ou pas nos présidents depuis Mitterand ?	@LeaSalame @franceinter
10:42	@lagrandefilolie	@LeaSalame Par contre, il était normand...	@LeaSalame

FIGURE 3.1 – Exemple d'un résultat de recherche sous forme d'une conversation

3.3 Définitions des conversations

D'une façon générale, la conversation est un objet d'étude pluridisciplinaire auquel s'intéressent en particulier la linguistique, mais aussi la psychologie, la sociologie, l'anthropologie et aujourd'hui l'informatique. Un ouvrage de référence en linguistique définit la conversation comme *"un type particulier d'interaction verbale"* (Traverso, 1999). Mais comme le note D. Luzzati, *"on est dans un univers d'ambiguïtés où l'interopérabilité des concepts jette un trouble permanent"* (Luzzati, 2007). Cependant, le cadre spécifique des réseaux sociaux permet de préciser les concepts manipulés même s'ils ne couvrent qu'une partie des questionnements posés par les sciences sociales.

Une conversation C sur Twitter se définit à partir d'un ensemble de messages échangés entre utilisateurs, à la suite d'un message initial (Amer et al., 2015). Plus précisément, elle peut se modéliser par un graphe orienté pour lequel différentes définitions ont été proposées dans la littérature selon l'information manipulée. Elles peuvent s'organiser en deux familles :

(1) dans la première qui est la plus fréquemment utilisée, les sommets du graphe sont les tweets impliqués dans la conversation (Ritter et al., 2010; Cogan et al., 2012; Herrera et al., 2015; Magnani and Montesi, 2010; Zubiaga et al., 2015; Amer et al., 2015; Chen

et al., 2011; Nishi et al., 2016; Kumar et al., 2010; Alvarez-Melis and Saveski, 2016; Doyle and Frank, 2015; Danescu-Niculescu-Mizil et al., 2011; Jang et al., 2012; Kim et al., 2012) ;

(2) dans la deuxième (Greetham and Ward, 2014; Honey and Herring, 2009; Purohit et al., 2013; Boyd et al., 2010; Mathew et al., 2016; Tamine et al., 2015; Sousa et al., 2010; Bliss et al., 2011; Romero et al., 2011; Russell et al., 2011), les sommets du graphe sont les participants impliqués dans la conversation.

Dans la première famille, différentes relations binaires sont prises en compte. La relation binaire non symétrique la plus fréquente est la réponse entre deux tweets : un arc (T_i, T_j) modélise la réponse entre un tweet T_j à un tweet T_i. Cette définition conduit généralement à un arbre de réponse orienté (Magnani et al., 2011; Chen et al., 2011; Amer et al., 2015; Cogan et al., 2012; Ritter et al., 2010; Herrera et al., 2015; Nishi et al., 2016; Kumar et al., 2010; Alvarez-Melis and Saveski, 2016; Doyle and Frank, 2015; Kim et al., 2012). D'autres auteurs (Zubiaga et al., 2015) considèrent qu'un tweet dans une conversation peut être en réponse à un ou plusieurs autres tweets, ce qui conduit à un graphe de réponse orienté. On peut, en effet, répondre par un seul tweet à plusieurs tweets simultanément.

Dans la deuxième famille, la relation binaire considérée est la relation non symétrique (P_i, P_j) modélisant la réponse du participant P_i au participant P_j. Cette définition conduit à un multigraphe orienté, chaque participant dans le graphe pouvant avoir plusieurs relations avec les autres participants. De même, d'autres auteurs prennent en compte le "retweet" et la "mention" entre deux participants (Purohit et al., 2013; Boyd et al., 2010; Greetham and Ward, 2014). Cette famille s'élargit aussi aux multigraphes où chaque arc entre deux mêmes sommets P_i et P_j est associé à un horodatage (Greetham and Ward, 2014). En particulier, d'autres auteurs considèrent les multi-arcs entre deux mêmes sommets P_i et P_j comme un arc pondéré (Mathew et al., 2016; Sousa et al., 2010; Bliss et al., 2011; Romero et al., 2011)(pondéré par le nombre de messages entre participants).

3.4 Méthodes de construction des conversations

Différentes méthodes de construction ont été proposées dans la littérature pour construire les graphes de conversations. Ils peuvent se structurer en trois grandes familles qui sont :

1. Méthodes de construction à base de relations.
2. Méthodes de construction à base de seuil.
3. Méthodes de construction à base de classification.

3.4.1 Méthodes de construction à base de relations

Dans cette section, nous présentons et décrivons les méthodes de construction des graphes de conversations en exploitant les relations entre les tweets ou les participants (selon le type de graphe). Elles sont résumées sur la figure 3.3.

La méthode de construction à base de relations est la plus fréquemment utilisée dans la littérature (Ritter et al., 2010; Zubiaga et al., 2015; Chen et al., 2011; Magnani et al., 2012; Amer et al., 2015; Kumar et al., 2010; Cogan et al., 2012; Alvarez-Melis and Saveski, 2016; Doyle and Frank, 2015; Jang et al., 2012; Danescu-Niculescu-Mizil et al., 2011; Kim et al., 2012; Nishi et al., 2016; Bliss et al., 2011; Russell et al., 2011). Elle se base

sur l'exploitation des relations (orientées) entre les tweets ou les participants impliqués dans une conversation, permettant de générer les deux types de graphes définis dans la section précédente. Ces relations sont de trois types : réponse, retweet, mention.

La quasi-totalité des auteurs se sont focalisés sur l'exploitation de la relation de type réponse pour construire leurs graphes de conversations. Le résultat est un graphe orienté de réponses, souvent restreint à un arbre, dont la relation binaire est non symétrique (réponse entre deux tweets). D'autres auteurs (Bliss et al., 2011; Tamine et al., 2015) utilisent la même relation de réponse mais cette fois entre les participants, ce qui permet d'obtenir un graphe de participants dont la relation binaire considérée est non symétrique modélisant la réponse entre deux participants.

Cette méthode de construction exploite l'ensemble des informations relatives aux tweets (date de publication, le texte du message, nom de l'auteur du tweet, etc.) fournies avec le corpus des tweets collectés. Les informations clés fournies avec les tweets collectés sont : le champ "$in-reply-to-status-id$" et le champ "$in-reply-to-user-id$". Le champ "$in-reply-to-status-id$" offre la possibilité de trouver les liaisons de réponse entre les tweets directement, ce qui permet de créer un graphe de réponse entre les tweets. Le champ "$in-reply-to-user-id$" contient la représentation de l'identifiant (ID) de l'auteur du tweet. Ainsi, ce champ donne la possibilité d'identifier la relation des réponses entre les auteurs des tweets et permet la création d'un graphe de participants.

Le choix du champ à exploiter se fait selon le type de graphe à créer (soit graphe de tweets ou graphe de participants). La méthode de création d'un graphe de tweets peut être modélisée en deux étapes majeures (Cogan et al., 2012) :

Á partir d'un corpus D de tweets, la première étape consiste à partir du champ "$in-reply-to-status-id$" à déterminer l'ensemble des tweets $tweet_{0_1}, tweet_{0_2}, tweet_{0_K}$ qui sont les initiateurs des conversations ; chaque tweet $tweet_{0k}$ est «racine» d'un graphe de conversation G_{C_K}. Cette étape se fait d'une manière récursive. Partant d'un tweet initial quelconque, une vérification se fait au niveau du champ "$in-reply-to-status-id$" : si le champ n'est pas nul, on vérifie le champ du tweet qui l'a engendré (par une réponse) et ainsi de suite jusqu'à trouver un tweet dont le champ "$in-reply-to-status-id$" est nul. La procédure est réitérée à partir d'un autre tweet non analysé jusqu'à épuisement de D.

Á chaque fois qu'un tweet racine $tweet_{0k}$ a été identifié, la deuxième étape s'enclenche pour construire l'arbre de conversation G_{C_i} de façon descendante à partir de chaque tweet racine $tweet_{0_k}$. Un balayage du corpus D permet d'identifier les tweets qui sont une réponse au $tweet_{0_k}$ (niveau 1) en vérifiant le champ "$in-reply-to-status-id$" de chaque tweet. Une fois que tous les tweets qui sont en réponse au $tweet_{0_k}$ sont récupérés, on ôte du corpus D les tweets dans l'arbre en cours de construction et on rebalaye l'ensemble restant pour identifier les tweets qui sont en réponse aux tweets du niveau 2, etc. Ce même processus se répète à chaque fois pour identifier les autres niveaux de l'arbre en balayant l'ensemble restant des tweets.

D'autres auteurs (Zubiaga et al., 2015) utilisent le même principe de construction mais sans avoir recours aux informations fournies par l'API de Twitter. Ils exploitent la relation de réponse entre les tweets, mais pas à travers le champ "$in-reply-to-status-id$" ; ils balayent manuellement la page Web de chaque tweet source (racine) pour récupérer les identifiants (ID) des tweets qui répondent à ce tweet. Par la suite, les tweets de type réponse sont collectés par le biais de leurs ID. En particulier, (Tamine et al., 2015) prennent

aussi en compte les relations "mention" et "retweet" pour inclure les relations de mention et retweet dans la construction des conversations. En outre, la considération des trois relations possibles permet de construire un graphe de participants avec multi-arcs entre deux mêmes sommets.

3.4.2 Méthodes de construction à base de seuil

Dans cette section, nous présentons les méthodes de construction qui ont été proposées dans la littérature, en utilisant différents types de seuils pour déterminer l'ensemble des tweets impliqués dans une même conversation.

3.4.2.1 Seuil de similarité textuelle

Afin de résoudre les contraintes de restriction de l'API Twitter pour l'obtention d'un fil de conversation complet, (Herrera et al., 2015) proposent une nouvelle méthode qui se base sur une similarité textuelle entre les tweets. Cette méthode permet de créer un arbre de réponse entre tous les tweets considérés similaires textuellement. Plus formellement, pour un tweet initial donné (représenté par un vecteur de mots), un ensemble des tweets candidats en réponse à ce tweet initial est collecté. Et il existe une arête entre deux tweets T_0 (tweet initial) et T_i (tweet appartenant à l'ensemble des tweets collectés), si et seulement si le score de similarité entre ces deux tweets est supérieur à un seuil de similarité fixé. Les auteurs utilisent la similarité cosinus pour calculer le score de similarité. Cette méthode de construction génère un arbre de réponse sous forme d'une étoile comme le montre la figure 3.2.

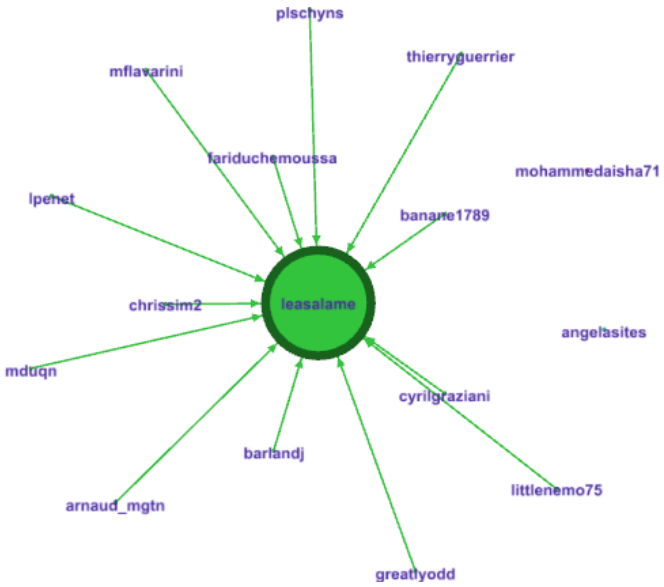

FIGURE 3.2 – Structure en "étoile" d'un graphe construit en utilisant un seuil de similarité textuelle

3.4.2.2 Seuil de similarité temporelle

Le travail de (Greetham and Ward, 2014) présente une méthode simple pour identifier les graphes de conversations entre les participants, sur la base d'un seuil temporel calculé

entre les temps de publication des réponses entre les participants. En particulier, il existe un arc (P_i, P_j) entre deux participants si et seulement si la différence entre le temps de publication du message publié par P_i et le temps de publication de message publié par P_j ne dépasse pas 9 heures. Le seuil a été choisi à la suite des expérimentations réalisées par les auteurs.

3.4.2.3 Seuil de nombre de réponses

(Romero et al., 2011) construisent un graphe de réponse entre participants en se basant sur un seuil défini sur le nombre de tweets échangés. Il existe un nombre de réponses minimum au-dessus duquel un arc peut être créé entre deux participants. Pour envoyer un tweet, l'utilisateur émetteur commence toujours son tweet par @username pour indiquer l'utilisateur qui reçoit son tweet. En particulier, dans le graphe de participants, un arc (P_i, P_j) est uniquement conservé si et seulement s'il y a eu plus de 1 (seuil n=1) tweet échangé entre chaque paire de participants.

3.4.3 Méthodes de construction à base de classification

Dans cette section, nous présentons les différentes méthodes de construction basées sur la classification qui ont été proposées dans la littérature.

3.4.3.1 Classification automatique

(Purohit et al., 2013) font l'hypothèse que l'exploitation des relations fournies par la plateforme Twitter ne suffit pas pour construire des conversations complètes. Pour pallier à cette limitation, ils proposent donc une méthode de classification qui utilise un ensemble d'indicateurs linguistiques basés sur la théorie de coordination. Ils ont identifié pour chaque type de relation (réponse, retweet, mention) un ensemble de modèles linguistiques qui leur correspond et ils ont montré que les modèles linguistiques utilisés pour la coordination entre les participants de la conversation permettent de distinguer les réponses, les retweets et les mentions du reste du trafic sur Twitter.

3.4.3.2 Classification manuelle

(Honey and Herring, 2009) utilisent une méthode d'analyse dynamique du sujet (Dynamic Topic Analysis method) pour analyser la cohérence dans les échanges entre les participants. Le principe de cette méthode est de lire les tweets et de décider s'ils appartiennent (en réponse à un autre tweet précédent) à une conversation ou pas. Cette catégorisation est effectuée par un analyste humain. En particulier, il existe un arc (P_i, P_j) entre deux tweets si et seulement si l'analyste juge que le contenu de deux tweets publiés par P_i et P_j est cohérent.

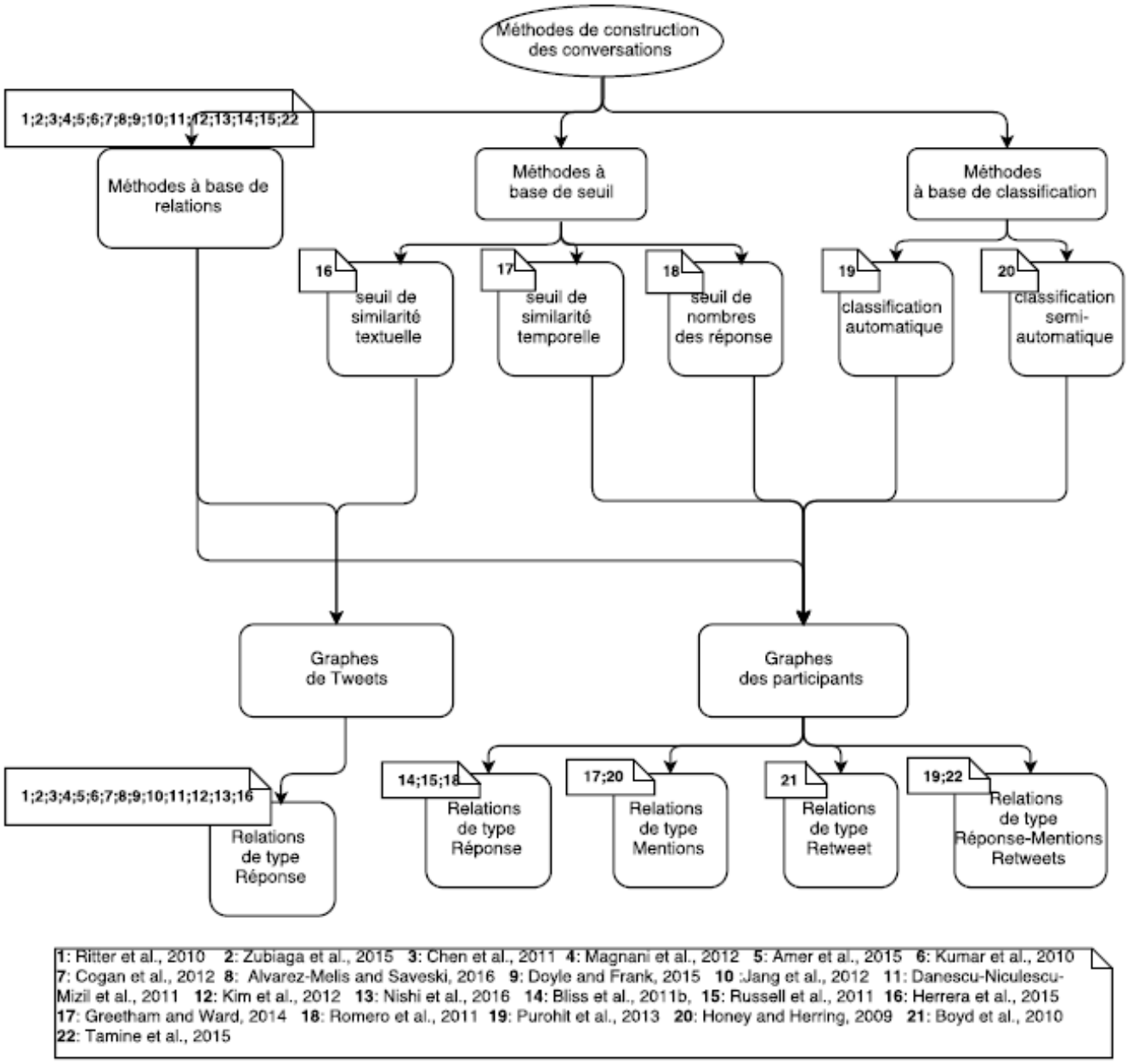

FIGURE 3.3 – Les différentes méthodes de construction des conversations

3.5 Les indices d'analyse des conversations

Différents indices ou critères d'analyse ont été proposés dans la littérature pour analyser les conversations. Ils peuvent se structurer en trois grandes familles qui sont :

1. Indices associés à la structure des graphes qui modélisent les conversations.
2. Indices associés aux contenus textuels de la conversation.
3. Indices associés aux participants impliqués dans la conversation.

3.5.1 Notations

Dans la suite, nous considérerons qu'une conversation C est modélisée par un graphe G_C = (V, A) où V =$\{v_1, v_2,...,v_n\}$ est l'ensemble fini des n sommets et A=$\{a_1, a_2,..., a_m\}$ l'ensemble fini des m arcs. Nous appelons classiquement n l'ordre du graphe et m la taille du graphe. Les sommets et les arcs peuvent correspondre selon les cas définis dans la section 3.3 respectivement aux participants, aux tweets et aux différentes relations.

- C : une conversation C peut se décomposer en "conversations restreintes" C=$\{c_1, c_2,..., c_k\}$ qui sont les sous-graphes connexes maximaux de G_C. Nous rappelons qu'un sous-graphe est connexe s'il existe une chaîne entre n'importe quelle paire de sommets. Un sous-graphe de G_C est connexe maximal s'il n'existe pas de sous-graphe induit connexe plus grand contenant les sommets de ce sous-graphe.
- T_C : l'ensemble total des tweets impliqués dans la conversation C.
- P_C : l'ensemble des participants p de la conversation C.
- H_C : l'ensemble des hashtags h (mots-clés) dans la conversation C.
- URL_C : l'ensemble des urls dans la conversation C.
- $Mention_C$: l'ensemble des mentions dans la conversation C.

Nous définissons les fonctions suivantes :

- Tweets(p) : l'ensemble des tweets publiés par le participant p.
- Retweets(p) : l'ensemble des tweets retweetés par le participant p.
- Favoris(p) : l'ensemble des tweets mis en favoris par le participant p.
- nbrRetweets(t) : le nombre total de retweets dans le tweet t.
- nbrFavoris(t) : le nombre total de favoris dans le tweet t.

Pour une conversation restreinte c_i, nous notons de plus les fonctions suivantes :

- c_i.tweet : l'ensemble des tweets impliqués dans c_i.
- c_i.participant : l'ensemble des participants impliqués dans c_i.
- c_i.sujet : le vecteur des f termes les plus fréquents dans c_i.tweet.
- c_i.h : nombre des hashtags présents dans c_i.
- c_i.url : nombre des urls présentes dans c_i.
- c_i.mention : nombre des mentions présentes dans c_i.
- c_i.retweet : l'ensemble des tweets impliqués dans c_i retweetés par c_i.participant.
- c_i.favoris : l'ensemble des tweets impliqués dans c_i mis en favoris par c_i.participant.

3.5.2 Indices structurels sur les graphes de conversations

Dans cette section, nous présentons les principaux indices basés sur la combinatoire du graphe G_C qui sont utilisés par la majorité des auteurs. Ils permettent de caractériser l'organisation macroscopique de G_C.

- **Densité du graphe** G_C : la densité du graphe G_C est définie par le rapport entre le nombre d'arcs divisé par le nombre d'arcs du graphe complet K_n d'ensemble de sommets V.

- **Diamètre du graphe** G_C : le diamètre du graphe G_C est la plus longue distance (au sens du plus court chemin) entre deux sommets de ce graphe.
- **Degré d'un sommet** : le nombre d'arcs incidents à ce sommet $\delta(v_i)$. Le degré d'un sommet est la taille de son voisinage.
- **Degré du graphe** G_C : le degré du graphe G_C est le degré maximal de tous ses sommets.
- **Degré de centralité d'un sommet** : cette mesure nous permet d'obtenir l'activité du sommet étudié. En effet, il constitue le rapport entre le nombre des arcs sortants du sommet et le degré du graphe K_n. Nous obtenons ainsi, pour un sommet v_i : $Dc(i) = d_{s(v_i)}/(n-1)$.

  Ces définitions sont appliquées à chacun des graphes modélisant les conversations restreintes de C. De plus, lorsque la conversation est modélisée par un arbre, les auteurs utilisent l'indice de profondeur.

- **Profondeur d'un arbre planté** : est la longueur maximale du plus court chemin dans l'arbre de sa racine jusqu'aux feuilles (sommets de degré 1).
  S'ajoute à ces indices structurels, un indice classique qui prend en compte la dimension temporelle de la conversation.
- **Durée d'une conversation restreinte** c_i **:** différence entre le temps de publication du dernier tweet publié et le temps de publication du premier tweet.

3.5.3 Indices associés aux contenus des conversations

Dans cette partie, nous présentons les indices qui sont basés sur le contenu des tweets impliqués dans les conversations. Comme nous l'avons précisé dès l'introduction de la thèse, nous ne tenons pas compte ici du contenu textuel à proprement parlé mais des informations complémentaires intégrées dans le message (URLs, hashtags) ainsi que du statut du tweet dans la conversation.

- **Nombre d'occurrence d'un hashtag dans une conversation restreinte** c_i : nombre d'apparition d'un hashtag dans l'ensemble c_i.tweet des tweets de c_i (Russell et al., 2011).
- **Score hashtags d'une conversation restreinte** c_i : un hashtag présent dans un tweet peut donner un contexte implicite du tweet et augmente la visibilité de ce tweet (Duan et al., 2010). Ce score est calculé comme le rapport entre le nombre de hashtags présents dans les tweets impliqués dans c_i et le nombre de hashtags de la conversation C (Amer et al., 2015) :

$$scorehashtags(c_i) = \frac{c_i.h}{H_C}. \tag{3.1}$$

- **Score URL d'une conversation restreinte** c_i : un tweet contenant des URLs est plus informatif car il apporte une information additionnelle portée par l'URL (Zhao et al., 2011; Nagmoti et al., 2010). Ce score est calculé par le rapport entre le nombre des URLs présents dans l'ensemble c_i.url des tweets impliqués dans c_i et dans l'ensemble total des URLs de C (Amer et al., 2015) :

$$scoreURL(c_i) = \frac{c_i.url}{URL_C} \tag{3.2}$$

- **Longueur d'un tweet** : nombre de caractères qui composent le tweet. Le score d'un tweet est défini par le ratio entre sa longueur et 140 (longueur maximale) (Magnani and Montesi, 2010).
- **Influence d'un tweet** : nombre de fois que le tweet a été mis en favoris ou retweeté (Amer et al., 2015).

$$Influence(t) = \alpha * nbrRetweets(t) + (1 - \alpha) * nbrFavoris(t) \tag{3.3}$$

Où α, dans l'intervalle [0;1], est un paramètre qui dénote l'importance relative des deux critères considérés.

3.5.4 Indices associés aux participants des conversations

Dans cette partie, nous présentons les indices qui mesurent des informations sur les participants (activité, influence, expertise, etc).

- **Popularité du participant** : la popularité d'un participant est estimée par le nombre de ses abonnés sur son profil Twitter (Magnani et al., 2012).
- **La pertinence d'un participant p à une conversation restreinte** c_i : est évaluée par rapport à la maîtrise du sujet par le participant. Elle est définie par le rapport entre la somme des occurrences des f termes de c_i.sujet de la conversation et le nombre de tweets de Tweets(p) dans la conversation (Amer et al., 2015).

Exemple :

- Vecteur c_i.sujet <politique, président, Obama>
- La liste des tweets du participant p :
  1. $Tweet_1$: je vais voter Obama comme président parce que je suis convaincu de sa stratégie politique.
  2. $Tweet_2$: selon les statistiques Obama sera le prochain président.
- Nb occurrence de terme f_1 (politique)=1
- Nb occurrence de terme f_2 (président)=2
- Nb occurrence de terme f_3(obama)=2
- Le nombre de termes significatifs dans $tweet_1$ (après filtrage de "je", "vais", etc) est 3
- Le nombre de termes significatifs dans $tweet_2$ est 2
- La pertinence de p est donc défini par (3+2)/ 2

- **Activité d'un participant p d'une conversation restreinte** c_i : désigne le degré d'implication de p dans c_i et est mesurée par le ratio entre le nombre de tweets publiés par p dans la conversation c_i sur le nombre de tweets de C (Amer et al., 2015) :

$$Activit_{participantp} = \frac{|Tc \bigcap Tweets(p)|}{|Tc|} \tag{3.4}$$

– **Influence d'un participant p dans une conversation restreinte** c_i : L'influence de p est évaluée par deux mesures : le nombre de fois que ses messages ont été retweetés par d'autres participants de c_i, et le nombre de fois que ses messages ont été mis en favoris (j'aime) par d'autres participants de c_i (Amer et al., 2015).

$$Influence(p) = \beta * nbrRetweets(Tweets(p)) + (1-\beta) * nbrFavoris(Tweets(p)) \tag{3.5}$$

Où β, dans l'intervalle [0;1], est un paramètre qui dénote l'importance relative des deux critères considérés.

– **L'expertise d'un participant p à une conversation restreinte** c_i :

L'expertise du tweet dépend de l'activité du participant p sur Twitter. Elle mesure si le participant p participe à une conversation restreinte c_i pertinente à son profil général sur Twitter. Elle se calcule à partir de deux vecteurs : celui contenant les termes des tweets de c_i.tweet impliqués dans la conversation c_i et celui contenant les termes les plus utilisés par p sur son profil Twitter.

Le premier vecteur résume les sujets de la conversation restreinte c_i, et le deuxième les centres d'intérêt du participant p sur Twitter. La comparaison des deux permet de vérifier si le participant est bien représentatif des sujets traités dans la conversation restreinte c_i.

Les termes de chacun des vecteurs sont pondérés par la fonction TF-IDF (Term Frequency-Inverse Document Frequency) qui permet d'évaluer l'importance d'un terme dans un document relativement à un corpus ou une collection de documents. Le premier vecteur est donc pondéré par un TF-IDF calculé sur c_i et le deuxième par un TF-IDF calculé sur le profil de l'auteur sur Twitter. Le score de pertinence du tweet est la mesure de similarité cosinus calculée entre les deux vecteurs. Techniquement, comme les deux vecteurs n'ont pas nécessairement la même taille, des termes sont ajoutés dans le vecteur c_i.tweet (le deuxième vecteur qui porte sur l'ensemble de Twitter est considéré par les auteurs (Magnani and Montesi, 2010; Chen et al., 2011) de l'indice comme plus grand que c_i.tweet). Une recherche est effectuée pour chaque tweet de c_i.tweet en utilisant la technique TweeTopic qui permet d'interroger les moteurs de recherche tel que Yahoo afin de récupérer des documents et générer une représentation thématique pour chaque tweet donné (liste des termes pertinents qui représentent le sujet). Les termes les plus pertinents sont extraits et ajoutés au vecteur de la conversation.

– **Score de communication d'un participant p d'une conversation restreinte** c_i :

Ce score est calculé entre les participants de c_i qui se suivent mutuellement sur Twitter. Considérons deux participants p_1 et p_2 de la conversation restreinte c_i qui sont mutuellement abonnés. Le score relatif de communication de p_1 (resp p_2) par rapport à p_2 (resp p_1) est défini par le logarithme du nombre des réponses (@replies) de p_1 (resp. p_2) postées par p_2(resp. p_1).

Le score de communication d'un participant p est défini par une double somme : somme de ses scores relatifs + somme des scores relatifs de ses suiveurs. Le score final de chaque participant est calculé comme la somme de son score de communication et le score de communication de ses suiveurs (Chen et al., 2011).

Exemple :

– Alice et Bob sont deux suiveurs sur Twitter

- David et Bob sont deux suiveurs sur Twitter

- Alice a échangé 3 messages avec Bob
- Bob a échangé 4 messages avec Alice
- Bob a échangé 2 messages avec David

- Le score relatif de communication de Alice= $\log(4)$
- Le score relatif de communication de David= $\log(2)$
- Le score relatif de communication de Bob= $\log(3)$

- Le score de communication de Alice= $\log(4) + \log(3)$
- Le score de communication de David= $\log(2) + \log(3)$
- Le score de communication de Bob= $\log(3) + \log(4) + \log(2)$

3.6 Conclusion

Ce chapitre a présenté les différents modèles de graphes utilisés pour modéliser des conversations sur Twitter. Il a permis d'identifier les différentes composantes qui structurent une conversation. De plus, nous avons présenté les principaux indicateurs qui permettent d'une part de caractériser la structure combinatoire des graphes et d'autre part de mesurer le rôle joué par ses différentes composantes (tweets et participants). Ces indicateurs sont mis en oeuvre dans les analyses expérimentales présentées dans le chapitre suivant.

4

Caractérisation des conversations sur Twitter

Sommaire

4.1 Introduction 34
4.2 Analyses expérimentales de la littérature 35
4.2.1 Collecte des données sur Twitter 35
4.2.2 Propriétés des conversations sur Twitter 41
4.3 Un nouveau graphe de conversation 42
4.4 Distribution observée des indices 44
4.4.1 Illustration sur un graphe de conversation 44
4.4.2 Résultats expérimentaux observés sur un corpus de graphes de conversations 53
4.5 Conclusion 61

4.1 Introduction

L'hypothèse qui a initialement guidé notre travail est que la compréhension d'un tweet ambigu peut être améliorée par la présentation des informations complémentaires issues d'une conversation dans laquelle ce tweet est impliqué. Cette hypothèse soulève donc le problème de l'extraction de ces informations "utiles". Et un préalable consiste à cerner les propriétés des graphes de conversations sur Twitter et les rôles joués par certains composants.

Dans ce chapitre, nous commençons par établir une synthèse des principaux résultats obtenus dans la littérature concernant les indicateurs présentés dans le chapitre précédent. Cette synthèse est précédée d'un résumé des différents protocoles expérimentaux sur lesquels s'appuient les résultats publiés. Il n'existe pas, à notre connaissance, de protocole

standardisé partagé par toute la communauté de chercheurs et les grandes variations exposées peuvent expliquer en partie la variabilité des résultats obtenus. Nous montrons cependant quelques invariants. Les travaux de la littérature ne présentent pas des résultats sur les sous-ensembles des indicateurs que nous avons retenus, nous avons choisi de compléter l'analyse bibliographique par nos propres expérimentations. Nous rapportons ici les résultats obtenus à partir de 30 requêtes, ce qui correspond à plus de 100 000 conversations restreintes.

4.2 Analyses expérimentales de la littérature

Dans cette section, nous présentons et décrivons les différentes méthodes de collecte de données et les variations exposées au niveau des différents corpus de données utilisés dans les travaux de la littérature. Ainsi, nous discutons les principales conclusions tirées de ces travaux.

4.2.1 Collecte des données sur Twitter

Les méthodes de collecte reposent sur des APIs fournies par Twitter que nous rappelons brièvement. Elles se déclinent ensuite essentiellement selon quatre stratégies majeures : collecte par mots-clés, par nom d'utilisateur, par zone géographique et par date.

4.2.1.1 Les APIs Twitter

Twitter dispose de plusieurs APIs permettant d'interroger sa base de données. La recherche Twitter est assez puissante et accessible grâce aux APIs Search, REST et Streaming. Ces APIs retournent un ensemble d'informations par requête ; les données concernent les tweets *(date de publication, le texte du message)*, l'auteur (date de création du compte, username), les métadonnées contenues dans les tweets *(hashtags, mentions, URLs)* et les informations de localisation *(pays, timezone, longitude/latitude)* (cf.figure 4.1). Il y a certaines différences subtiles lors de l'utilisation de chaque API pour récupérer les résultats de recherche.

- API Search : permet d'interroger Twitter pour récupérer une collection de tweets pertinents correspondant à une requête spécifiée. Cependant, cette API de recherche est axée sur la pertinence et non sur l'exhaustivité.
- API REST : extension avancée qui permet d'accéder à des fonctionnalités avancées de twitter : chercher des utilisateurs, des followers, des status, des informations sur les comptes utilisateurs, etc. Cette API remonte dans le temps *(jusqu'à environ une semaine)* pour trouver les tweets qui ont déjà été publiés. La recherche sur cette API prend les mots-clés comme des requêtes. La réponse à une requête ressemble plus ou moins à ce que l'on obtiendrait en tapant un mot-clé dans le moteur de recherche de Twitter. La recherche est limitée en vitesse par le nombre d'interrogations possibles *(450 requêtes toutes les 15 min)*.
- API Streaming : permet de communiquer avec Twitter en mode streaming pour capturer de nouveaux tweets en temps réel. Avec l'utilisation de l'API Streaming, la recherche des mots-clés, des hashtags, ID utilisateur et les zones géographiques simultanément est possible. L'API facilite cette recherche et fournit un flux continu de

tweets correspondant aux critères de recherche. L'API ne renvoie pas d'anciennes données, mais une fois que la requête est mise en place, un ensemble des messages répondant à cette requête est collecté. La particularité de cette API, est de permettre l'accès à de gros volumes de données Twitter et présente moins de contraintes pour les limites d'accès sur Twitter. Cependant, cette API nécessite une infrastructure très solide pour gérer l'accès aux données.

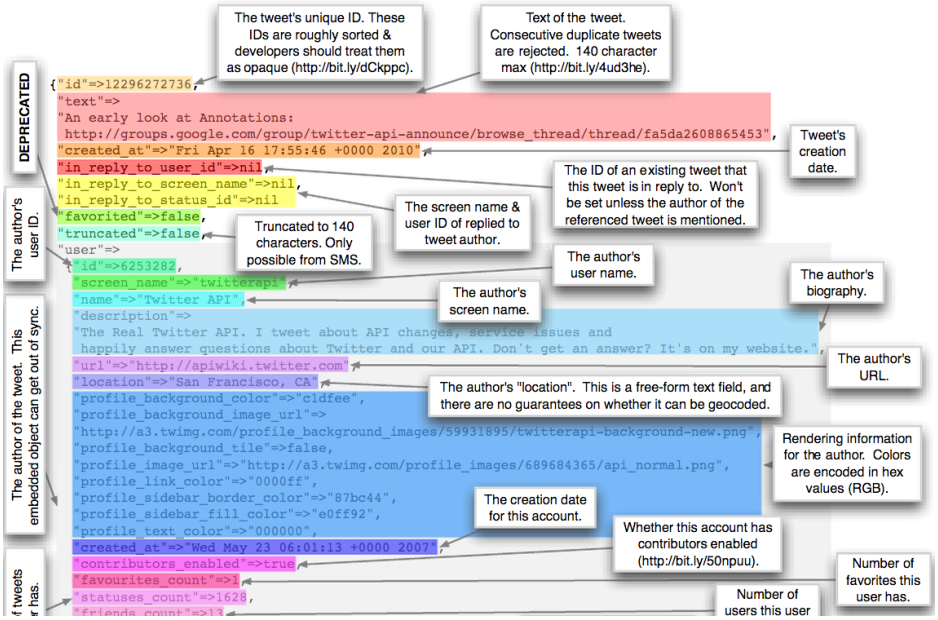

FIGURE 4.1 – Exemple des attributs fournis avec chaque tweet

La collecte de données sur Twitter est facilitée par l'utilisation des paramètres. Les valeurs des paramètres acceptables pour la recherche comprennent des mots-clés, des hashtags, des phrases, des régions géographiques, le nom d'utilisateur ou ID utilisateur et l'intervalle de temps *(date début, date fin)*.

4.2.1.2 Les Différentes méthodes de collecte de données

Dans les travaux de la littérature portant sur l'analyse des conversations, les corpus ne sont pas identiques et chaque chercheur construit son propre jeu de données. De la variété des corpus on peut néanmoins identifier quatre stratégies qui sont résumées dans la figure 4.2.

1. Collecte par mots-clés :

   La collecte par mots-clés est la plus fréquemment utilisée dans les travaux de recherche (Amer et al., 2015; Chen et al., 2011; Cogan et al., 2012; Herrera et al., 2015; Zubiaga et al., 2015; Purohit et al., 2013; Magnani and Montesi, 2010; Boyd et al., 2010; Tamine et al., 2015; Mathew et al., 2016; Alvarez-Melis and Saveski, 2016; Romero et al., 2011). Cette stratégie se divise en deux sous-catégories :

   – La collecte complète : les tweets collectés contiennent un ou plusieurs mots d'une liste établie dans la requête de recherche. Les tweets sélectionnés sont extraits d'un flux contenant l'ensemble des tweets et sont sélectionnés en fonction des mots qu'ils contiennent.

– La collecte filtrée : ce type de collecte n'est utilisé que dans le travail de (Cogan et al., 2012). Chaque tweet collecté doit contenir exactement tous les mots-clés mentionnés dans la requête de recherche. Le point faible de cette collecte est que les tweets collectés ne peuvent pas créer une conversation complète ; ainsi, tous les tweets au sein d'une seule conversation peuvent ne pas correspondre à un même mot-clé. En outre, avec cette collecte, le corpus final sera non exhaustif.

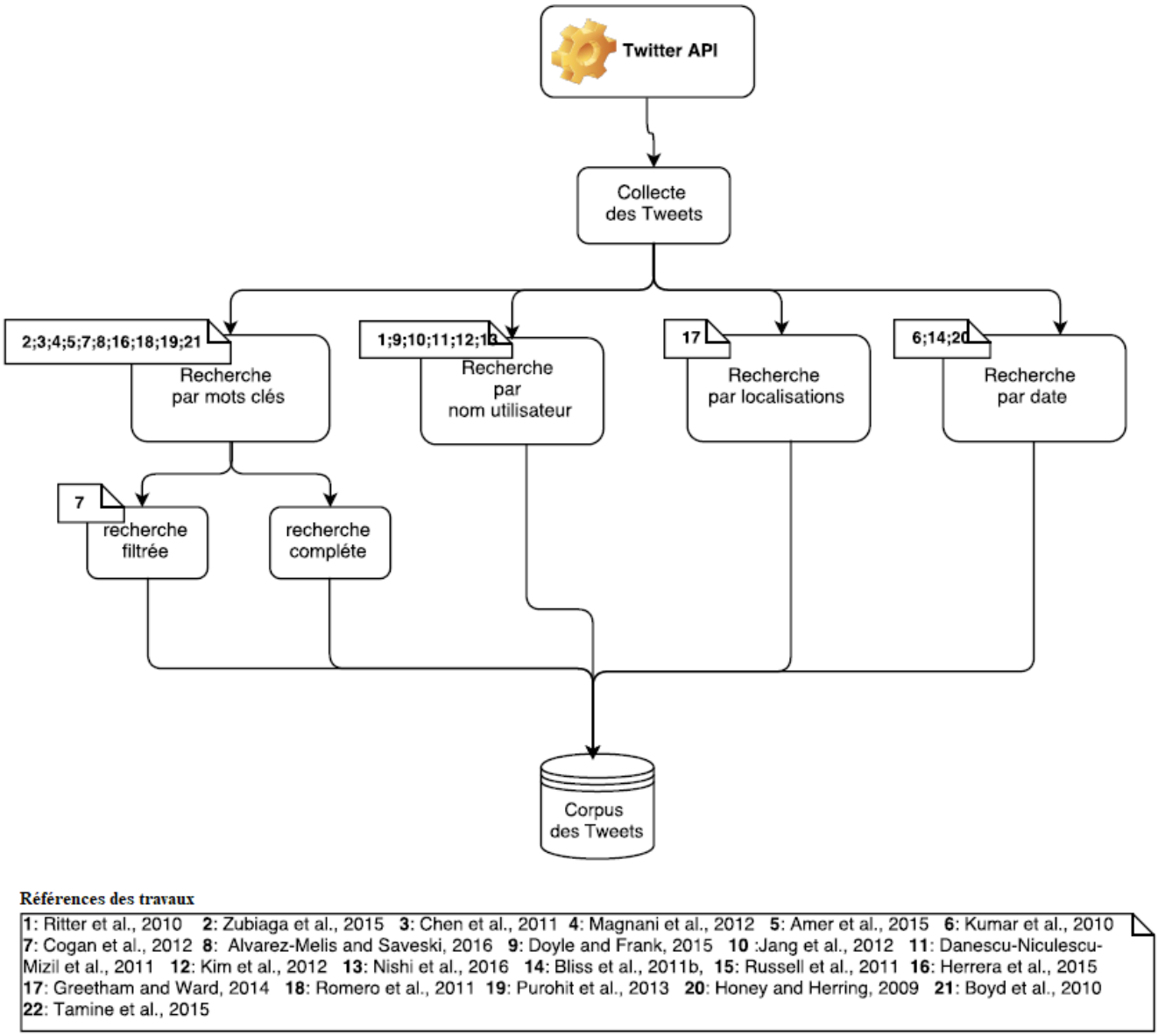

FIGURE 4.2 – Les différentes méthodes de collecte des données

2. Collecte par nom d'utilisateur "Timeline" :

   La collecte par nom d'utilisateur se traduit par une liste de noms, ou d'identifiants d'utilisateurs séparés par des virgules, mentionnée dans la requête de recherche. Le résultat de recherche renvoie tous les tweets publics des utilisateurs mentionnés dans le flux de Twitter (Ritter et al., 2010; Danescu-Niculescu-Mizil et al., 2011; Nishi et al., 2016; Doyle and Frank, 2015; Sousa et al., 2010; Jang et al., 2012; Kim et al., 2012).

3. Collecte par zone géographique :

   La collecte par zone géographique est une autre méthode pour collecter les tweets. Cette méthode a été utilisée dans le travail de (Greetham and Ward, 2014). La requête de collecte est représentée par des informations de localisation *(pays, longitude, latitude, etc.)*.

4. Collecte par date :

   Avec cette méthode, les tweets sont collectés en fonction de leur date de publication. Ainsi, les tweets collectés appartiennent à un intervalle de temps bien défini pendant la requête de recherche. Cette méthode est utilisée dans les travaux de (Honey and Herring, 2009; Kumar et al., 2010; Bliss et al., 2011).

4.2.1.3 Les caractéristiques des corpus de données

Le tableau 4.2 présente les caractéristiques des corpus utilisés dans les travaux cités dans notre état de l'art. Il permet de confirmer la variabilité des stratégies employées ; et par conséquent la difficulté d'établir ensuite des conclusions consensuelles sur les résultats obtenus.

Auteurs	Type de collecte	Détails de la collecte	Taille du corpus
(Amer et al., 2015)	mots-clés	15 sujets 9 requêtes liées à TREC Microblog 2011-2012-2013 6 requêtes sur des sujets d'actualité 2014	7806 conversations
(Chen et al., 2011)	mots-clés	35 sujets collectés sur une période de 3 semaines	2100 conversations
(Cogan et al., 2012)	mots-clés	2 sujets collectés sur une période de 5 jours (1 février jusqu'au 5 février 2012)	3114 conversations
(Herrera et al., 2015)	mots-clés	collecte sur une période de 28 jours (31 mars jusqu'au 27 avril 2015)	152721 conversations
(Zubiaga et al., 2015)	mots-clés	3 sujets : #ferguson, #ottawashootings, #ebola	291 conversations pour #Ferguson 475 conversations pour #Ottawashootings 18 conversations pour #Ebola
(Purohit et al., 2013)	mots-clés	6 sujets : #Japan Earthquake2011, #HaitiEarthquake2010, #HurricaneIreneStorm2011, #DebtCeilingDebate2011, #SkypeMicrosoftDeal2011, #GlennBeckRally2010	809692 conversations
(Magnani and Montesi, 2010)	mots-clés	10 sujets	20000 tweets
(Boyd et al., 2010)	mots-clés	1 sujet (#Iranian election) collecté sur une période de 6 mois	720000 tweets
(Tamine et al., 2015)	mots-clés	2 sujets : #Sandy collecté sur une période de 2 jours (29 octobre jusqu'au 31 octobre 2012) #Ebola collecté entre la période du 29 juillet et du 28 août 2014	240991 conversations pour #Sandy 196005 conversations pour #Ebola
(Mathew et al., 2016)	mots-clés	3 sujets : #BigBillionDay; #BigAppShoppingDays; #BASD2andBASD3	135593 tweets pour#BigBillionDay 51323 tweets pour #BigAppShoppingDays 160322 tweets pour #BASD2andBASD3
(Alvarez-Melis and Saveski, 2016)	mots-clés	14 sujets collectés sur une semaine du mois d'avril 2014	61.795 conversations
(Ritter et al., 2010)	nom d'utilisateur	collecter les tweets publics des utilisateurs sur une période de 2 mois	1.3 million de conversations
(Danescu-Niculescu-Mizil et al., 2011)	nom d'utilisateur	collecter les tweets publics de 7800 utilisateurs	215000 conversations
(Nishi et al., 2016)	nom d' utilisateur	collecter les tweets publics de 505557 utilisateurs entre la période du 15 mars jusqu'au 9 décembre 2011	2170021 conversations

(Doyle and Frank, 2015)	nom d'utilisateur	collecter les tweets de 14 utilisateurs	63.673 conversations
(Sousa et al., 2010)	nom d'utilisateur	collecter les tweets de 49303 utilisateurs entre la période du 17 mars au 16 mai 2010	612556 tweets
(Jang et al., 2012)	nom d'utilisateur	collecter les tweets publics de 1414 utilisateurs entre la période du 9 septembre au 4 octobre 2011	1,338,002 conversations
(Kim et al., 2012)	nom d' utilisateur	collecter les tweets publics de 136730 utilisateurs sur une période de 4 mois (de maiaoût 2011)	153054 conversations
(Greetham and Ward, 2014)	localisation	tweets collectés sur une période de 28 jours	4 million de tweets
(Honey and Herring, 2009)	dates	tweets collectés entre 6h du matin et 18h du soir sur une période de 1 jour (11 janvier 2008)	1472 tweets
(Kumar et al., 2010)	dates	collecter tous les tweets publiés pendant le mois de septembre 2009	$36.24 * 10^6$ conversations
(Bliss et al., 2011)	dates	tweets collectés entre septembre 2008 et février 2009	40 million de réponses

TABLE 4.2 – Variabilité de la taille des corpus des données utilisés dans les travaux effectués sur les conversations Twitter

De plus, notons que 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- restreignent, également, la mise en place des jeux de données volumineux. En effet, l'accès standard ne permet de récupérer qu'un pourcentage limité de l'ensemble des messages publiés. Un petit nombre de chercheurs ont eu accès à une part plus significative du flux afin de constituer de grands jeux de données et mener des études approfondies (Purohit et al., 2013; Tamine et al., 2015; Ritter et al., 2010; Nishi et al., 2016).

Cette synthèse bibliographique, présentée par le tableau 4.2, met en avant la difficulté à laquelle on est confronté lors de la collecte des données sur Twitter. Malgré l'accès à ces données proposées par Twitter, les contraintes techniques et méthodologiques limitent fortement l'établissement de corpus de référence.

4.2.2 Propriétés des conversations sur Twitter

Dans cette section, nous présentons quelques conclusions tirées des travaux de la littérature. Nous discutons les propriétés observées sur les indices d'analyse en les classant en trois familles : la première famille contient les travaux qui ont étudié la modélisation des structures des conversations et ont exploité les indices structurels des conversations, la deuxième famille regroupent les travaux qui ont exploré le contenu textuel des conversations et la troisième famille regroupe les travaux qui ont étudié conjointement les indices associés au contenu textuel et les indices structurels.

Les travaux de (Cogan et al., 2012; Kumar et al., 2010; Nishi et al., 2016; Kwak et al., 2010) s'inscrivent dans la **première famille**. Ils ont porté sur les structures des conversations et plus particulièrement sur l'interprétation sociologique des patrons des conversations les plus fréquentes (Tamine et al., 2015; Cogan et al., 2012; Kumar et al., 2010; Nishi et al., 2016). L'analyse de ces patrons montre qu'il y a principalement deux types de patrons dominants :

(1) Structure aplatie *(forme en chemin)* : la structure est essentiellement sous forme d'un chemin et les tweets apparaissent successivement comme une cascade de réponses. Cette structure correspond généralement à un large nombre de sommets et une large profondeur. Selon (Cogan et al., 2012), l'analyse topologique des patrons montre que les structures en chemins dominent largement l'ensemble des graphes de conversations.

(2) Structure en étoile : cette structure est représentée par plusieurs participants qui ont répondu à un tweet unique populaire. Cette structure correspond à un large nombre de sommets et un degré de profondeur très faible.

Un troisième type de patrons a été représenté dans les travaux de (Nishi et al., 2016). Ce type n'est pas fréquent. Il est représenté par un grand arbre sous forme d'une structure irrégulière. Cette forme irrégulière correspond généralement à une structure avec un grand nombre de sommets et un degré de profondeur variable.

Peu de travaux de la littérature ont porté sur la comparaison des deux types de graphes de conversations présentés dans la section 3.3 *(graphe de tweets et graphe de participants)*. Un des rares travaux est celui de (Cogan et al., 2012), dont les auteurs ont analysé les graphes des tweets *(en considérant la liaison de type réponse entre les tweets)* et les graphes des participants *(en considérant la liaison de type mention entre les participants)* extraits d'un même corpus collecté en utilisant deux mots-clés "Gop" et "flu" comme requêtes de recherche. L'analyse souligne une différence importante entre les structures des graphes. Les graphes de tweets contiennent un grand nombre de composantes connexes ; ce qui indique la présence de nombreuses conversations courtes. Et, les composantes

connexes des graphes de tweets se réduisent bien souvent à des chemins alors que celles des participants sont généralement structurées en étoile. Une analyse complémentaire a été effectuée en considérant pour les tweets la liaison de type réponse et pour les participants les liaisons de type réponse, retweet et mention. Elle a confirmé les résultats précédents.

Des travaux portant sur les indicateurs classiques *(taille, degré, diamètre, etc)* ressortent quelques résultats communs importants. La majorité des conversations sur Twitter sont très courtes (Ritter et al., 2010; Honey and Herring, 2009). La durée moyenne de la grande majorité des conversations est entre 6 heures (Cogan et al., 2012) et 9 heures (Greetham and Ward, 2014). Le nombre moyen des participants d'une conversation est faible : entre 2 et 7 participants (Tamine et al., 2015). Les degrés des sommets des graphes de conversations de petite taille sont faibles et 30 % des messages sont non réciproques dans la plupart des graphes de conversations, ce qui coïncide avec le fait que les conversations sont souvent dominées par quelques participants (Greetham and Ward, 2014).

Dans la **deuxième famille**, différents auteurs (Ritter et al., 2010; Boyd et al., 2010; Honey and Herring, 2009; Alvarez-Melis and Saveski, 2016; Jang et al., 2012; Kim et al., 2012; Sousa et al., 2010) se sont intéressés au contenu textuel des conversations en exploitant différentes propriétés : par ex. vocabulaire, présence de hashtags, sujets des conversations. Cependant, à notre connaissance, ces travaux n'ont pas fait le lien entre la structure des conversations et leurs contenus. Nous ne les approfondissons donc pas dans la suite de ce manuscrit.

Dans la **troisième famille** qui associe les indices structurels et textuels, nous retenons principalement les résultats des travaux de (Magnani and Montesi, 2010; Magnani et al., 2012; Tamine et al., 2015; Amer et al., 2015). Magnani et Montesi ont comparé dans un cadre de RI des résultats de requêtes effectués avec la prise en compte des deux types d'indices. La comparaison a été faite sur l'ensemble des 10 meilleures conversations retournées pour chaque requête de recherche. Les résultats semblent confirmer l'intérêt de la combinaison des indices mais des travaux complémentaires restent nécessaires pour valider la robustesse des conclusions et identifier précisément la valeur ajoutée de chaque type d'information. Des critiques ont d'ailleurs été récemment émises (Amer et al., 2015) sur les fonctions d'agrégation utilisées par (Magnani and Montesi, 2010; Magnani et al., 2012). En effectuant des analyses approfondies, Amer et al. ont mesuré une amélioration significative de la qualité des résultats obtenus lors de l'intégration des informations associées aux liens sociaux par comparaison avec une analyse basée uniquement sur le contenu textuel. Récemment,(Tamine et al., 2015) ont combiné des indices associés à l'analyse structurelle et des indices associés à l'analyse textuelle. Cependant, leur objectif n'était pas une comparaison des résultats obtenus sur les graphes avec le filtrage de chacun mais l'amélioration des résultats à des requêtes obtenue par la combinaison.

4.3 Un nouveau graphe de conversation

Les travaux cités sur l'analyse structurelle ont confirmé la complémentarité des deux modèles de graphes (entre tweets et entre participants). Cependant, d'un point de vue opérationnel, la manipulation en parallèle de deux graphes s'est avérée délicate. D'une part elle se heurte à la question de l'agrégation des indicateurs et d'autre part elle pose des problèmes de surcharge de carte mentale pour les utilisateurs lors de la manipulation des graphes dans le cadre de la tâche de contextualisation. Il nous a donc semblé pertinent de proposer une nouvelle définition qui permettait de rassembler dans un seul objet

combinatoire les deux aspects. Le nouveau graphe est donc défini comme suit.

Une conversation est représentée par un graphe orienté (V, A) où V= $U \cup T$ représente les sommets du graphe qui peuvent être, soit un tweet $t \in T$, soit un participant $u \in U$, et $A \in T * T \cup U * T$ sont les arcs du graphe qui représentent le type de relation entre les sommets du graphe (cf. figure 4.3).

Ainsi, T*T est défini par la règle suivante :

– Si un tweet t_1 répond à un tweet t_2 alors un arc relie le sommet représentant le tweet t_2 au sommet représentant le tweet t_1.

Et U*T est défini par les règles suivantes :

– Si un tweet *t* est publié par un participant *u* alors un arc relie un sommet représentant le participant *u* au sommet représentant le tweet *t*.
– Si un tweet *t* mentionne un participant *u* alors un arc relie le sommet représentant le tweet *t* au sommet représentant le participant *u*.
– Si un participant *u* retweet un tweet *t* alors un arc relie le sommet représentant le participant *u* au sommet représentant le tweet *t*.

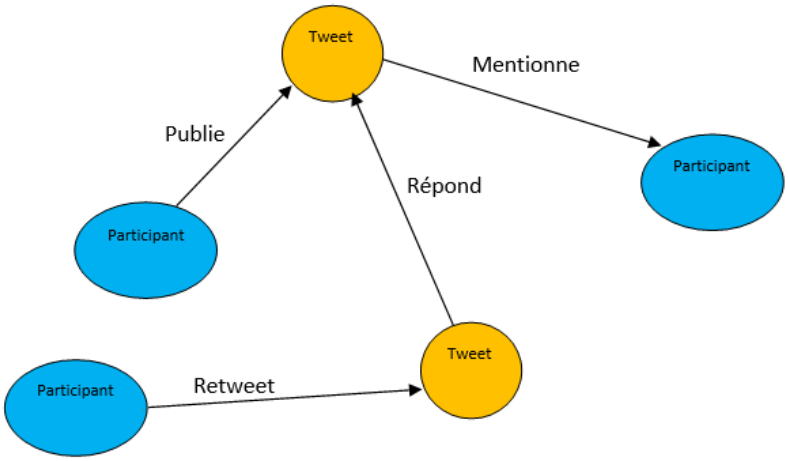

FIGURE 4.3 – Nouvelle modélisation du graphe de conversation

Notons, que les travaux de la littérature cités précédemment considèrent tous trois relations *(réponse, retweet, mention)* entre les sommets du graphe de conversation mais qu'aucun n'a pris en considération la relation "publie" qui modélise la relation entre un tweet et son auteur. Dans notre objectif opérationnel de contextualisation d'un tweet, elle nous a paru pertinente et nous l'avons donc ajoutée dans la définition du graphe retenue.

De plus, une majorité des travaux semble avoir mis l'accent sur les graphes des retweets (Kwak et al., 2010; Bakshy et al., 2011) plutôt que sur ceux des réponses. Cependant, les réponses ont été considérées comme plus informatives sur les relations entre les utilisateurs que sur les relations des retweets par différents travaux (Sousa et al., 2010; Gonçalves et al., 2011; Bliss et al., 2011). Et elles permettent aussi d'être un point d'entrée sur l'analyse des réponses émotionnelles des utilisateurs (Dodds et al., 2011; Bliss et al., 2011) qui jouent au niveau collectif un rôle important sur Twitter ; les émotions collectives et les humeurs de Twitter covarient souvent avec les résultats des actions collectives, probablement induites par la diffusion rapide d'informations à grande échelle.

4.4 Distribution observée des indices

Dans cette section, nous examinons les distributions des indices présentés dans le chapitre précédent sur un corpus de graphes de conversations. Nous commençons par détailler les résultats obtenus sur un exemple paradigmatique construit à partir d'une requête #Macron. Puis, nous présentons les résultats obtenus sur plus de 100 000 conversations restreintes (composantes connexes des graphes de conversations) issues de 30 requêtes.

4.4.1 Illustration sur un graphe de conversation

Nous détaillons les distributions des différents indices sur un graphe de conversation construit à partir de la requête #Macron. La collecte des informations nécessaires à la construction du graphe a été obtenue via l'API Rest de Twitter présenté dans la sous-section 4.2.1 pour la période du mois de mai 2017. Nous avons collecté 8000 tweets et le graphe de conversation correspondant contient donc 5934 sommets de type tweets et 3767 sommets de type participants. Notons que certains tweets collectés ne sont pas impliqués dans le graphe construit d'où la différence entre le nombre de tweets collectés et le nombre de sommets de type tweets du graphe.

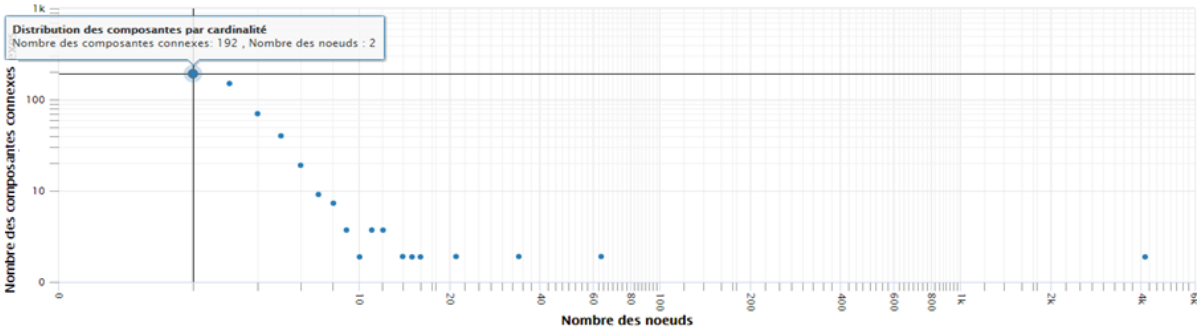

FIGURE 4.4 – Distribution de la taille des composantes connexes dans le graphe de conversation

Ce graphe contient 504 composantes connexes (cf. figure 4.4). En outre, 99 % des composantes connexes (500 composantes connexes) comportent moins de 20 sommets. Il n'y a que 4 composantes connexes qui comportent plus de 20 sommets.

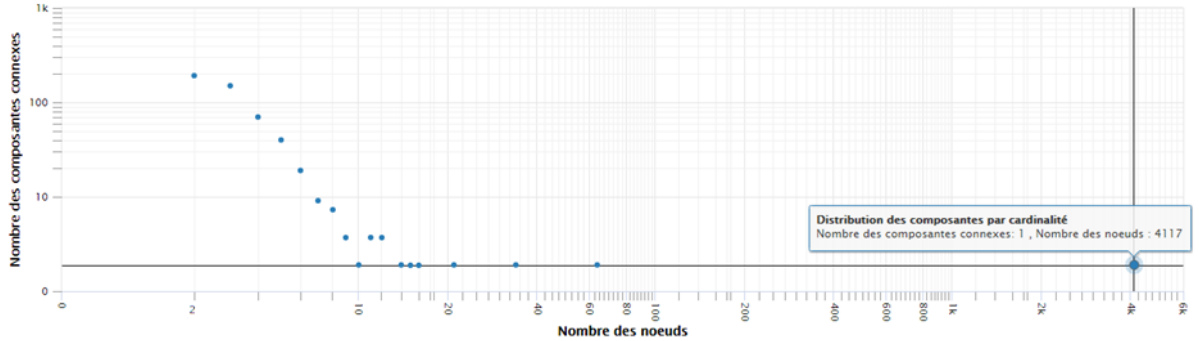

FIGURE 4.5 – La taille de la grande composante connexe retenue pour nos expérimentations

Pour le calcul des indices, nous retenons la plus grande composante connexe (composante contenant 4117 sommets, cf. figure 4.5). Le plus grand nombre des composantes connexes

ont un ordre exactement égal à 2. Ces composantes ont une structure sous forme d'un chemin. Cette structure est dominante sur l'ensemble des composantes connexes.

– **Score d'influence des participants (Amer et al., 2015)**

La figure 4.6 montre la distribution du score d'influence des participants. Le score d'influence évalue si le participant est influent ou pas sur l'ensemble des participants de la conversation. Nous remarquons que près de 80 % des participants ont un score d'influence égal à zéro. Les participants qui ont un score d'influence supérieur à zéro, sont ceux qui ont au moins publié ou ont répondu à un tweet et pour lesquels leurs tweets ont eu des retweets ou des favoris. Pour cette raison, nous ne nous focalisons que sur les participants qui ont un score d'influence supérieur à 0.

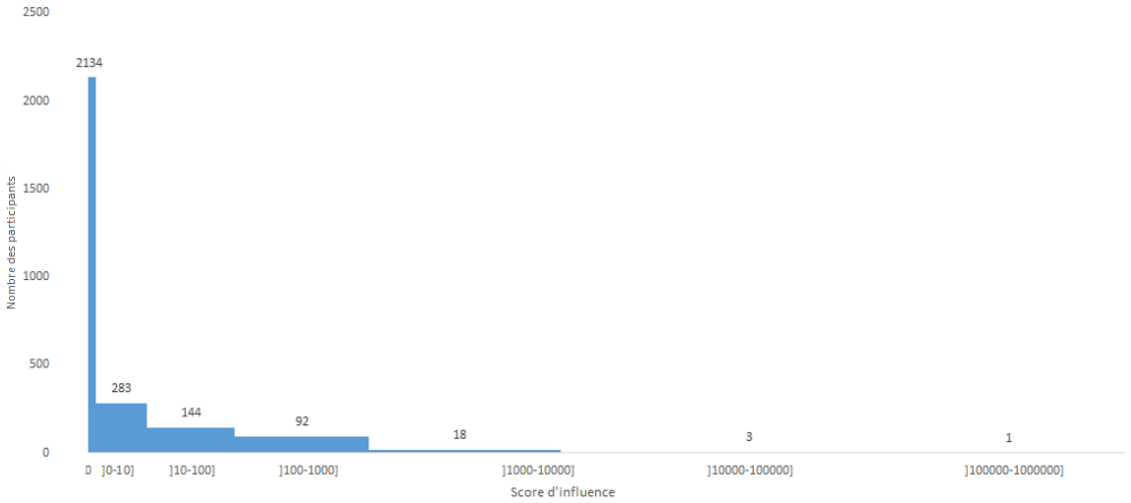

FIGURE 4.6 – Distribution des scores d'influence des participants de la plus grande composante connexe du graphe de conversation

La figure 4.7 montre la distribution des scores d'influence pour les participants qui ont publié ou ont répondu à un tweet. Nous remarquons la persistance des participants avec un score nul ; ce qui est expliqué par le fait que leurs messages n'ont pas attiré l'attention d'autres participants (n'ont pas eu des retweets, ni des favoris).

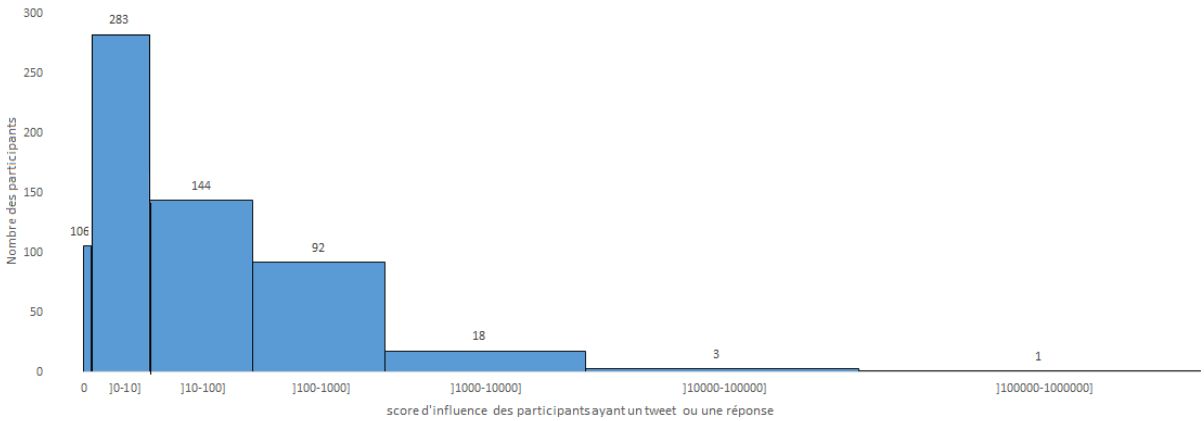

FIGURE 4.7 – Distribution des scores d'influence des participants ayant un tweet ou une réponse de la plus grande composante connexe du graphe de conversation

– **Score de popularité des participants** (Magnani and Montesi, 2010)

La figure 4.8 montre la distribution du score de popularité par intervalle des participants. L'observation de la distribution peut conduire à distinguer deux types de comportements ; les participants qui ont moins de 1000 followers et ceux qui en ont plus de 1000, et qui sont associés à une popularité plus importante.

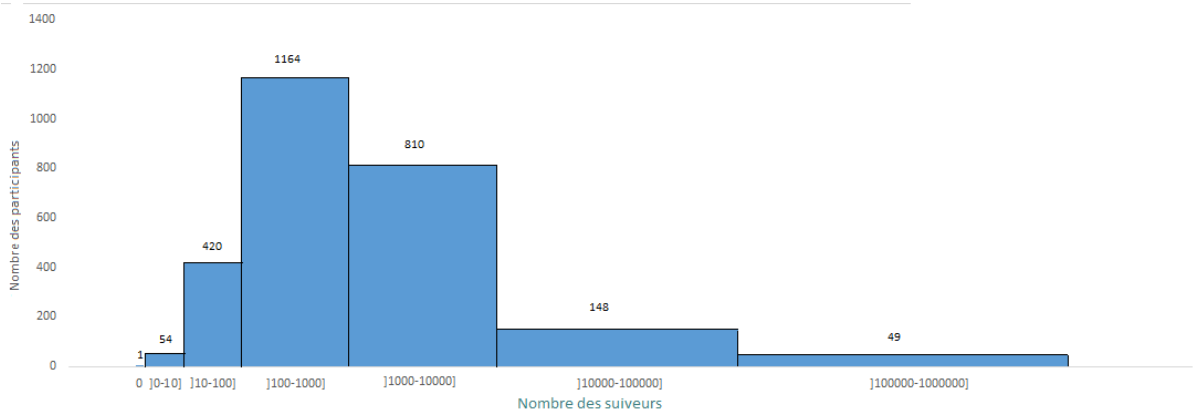

FIGURE 4.8 – Distribution des scores de popularité par intervalle des participants de la plus grande composante connexe du graphe de conversation

– La moyenne de popularité par participant

La moyenne de popularité par participant est estimée par le rapport entre le nombre de followers par participant sur le nombre total des followers des participants. La figure 4.9 illustre la moyenne de popularité par participant. Cette figure montre que la majorité des participants (2673 participants) ont un score compris entre 0 exclu et 0.2.

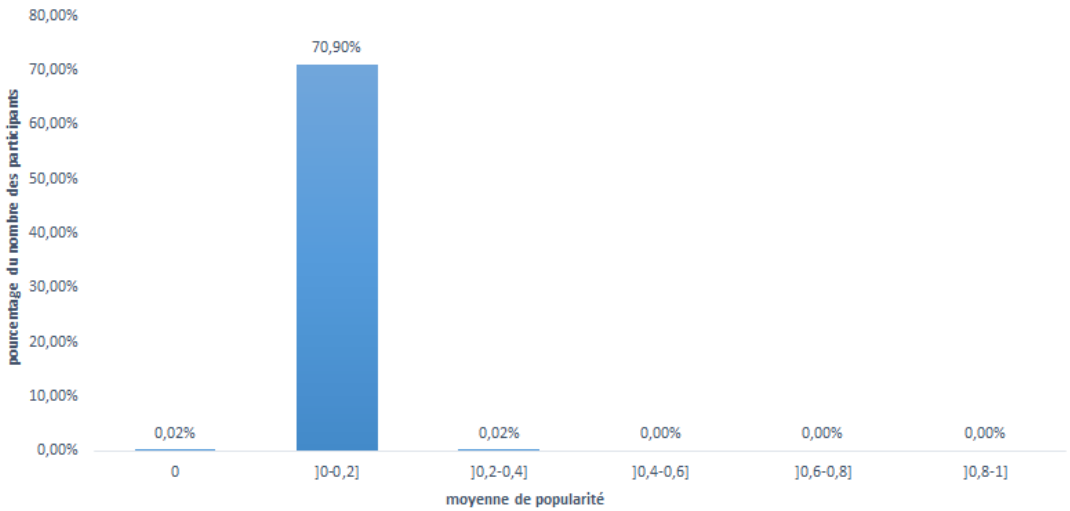

FIGURE 4.9 – Distribution de la moyenne de score de popularité par participant de la plus grande composante connexe du graphe de conversation

– Corrélation entre le score de popularité et le score d'influence

La figure 4.10 analyse la corrélation entre les deux scores : score de popularité et score d'influence. Nous remarquons que les deux scores sont corrélés (score de corrélation égal à 0.6). Ainsi, nous constatons que la relation est forte entre les deux scores, avec une

forme non linéaire monotone et un sens non linéaire positif. Nous privilégions le score d'influence au score de popularité car son calcul ne dépend que d'informations extraites du graphe de conversation alors que le score de popularité dépend de l'activité générale du participant sur Twitter qui est donc d'accès plus coûteux.

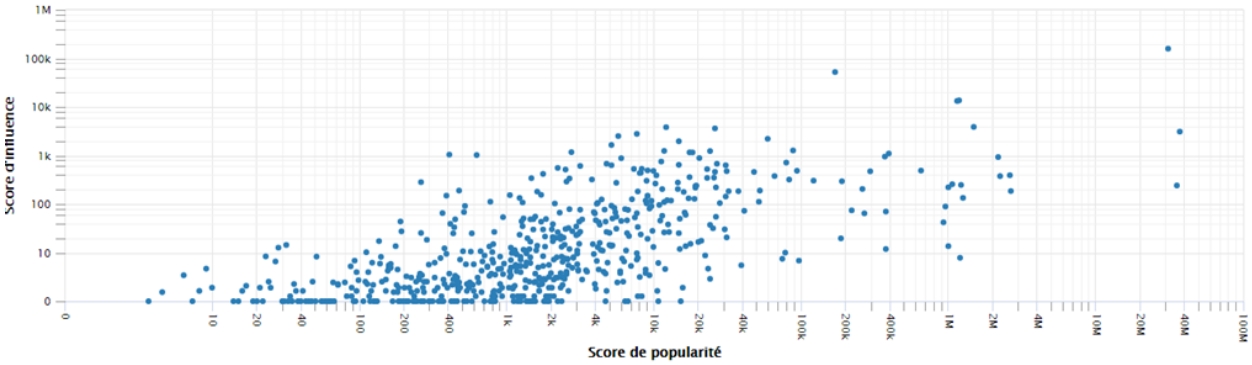

FIGURE 4.10 – Corrélation entre le score de popularité et le score d'influence des participants de la plus grande composante connexe du graphe de conversation

– **Score d'expertise des participants**

La figure 4.11 montre la distribution du score d'expertise des participants. Sur ce graphique, nous remarquons qu'il y a 1019 participants qui ont un score égal à zéro et que la grande majorité des participants ont un score d'expertise faible compris entre]0-0.2]. La présence des participants avec un score égal à zéro est expliquée par le fait que ces participants n'ont pas publié dans la conversation (sont présents dans la conversation par la relation de mention) ou bien n'ont pas publié auparavant (sur leurs profils) sur le sujet de la conversation. Pour cette raison, nous ne calculons (cf. figure 4.12) le score d'expertise que pour les participants qui ont publié des tweets.

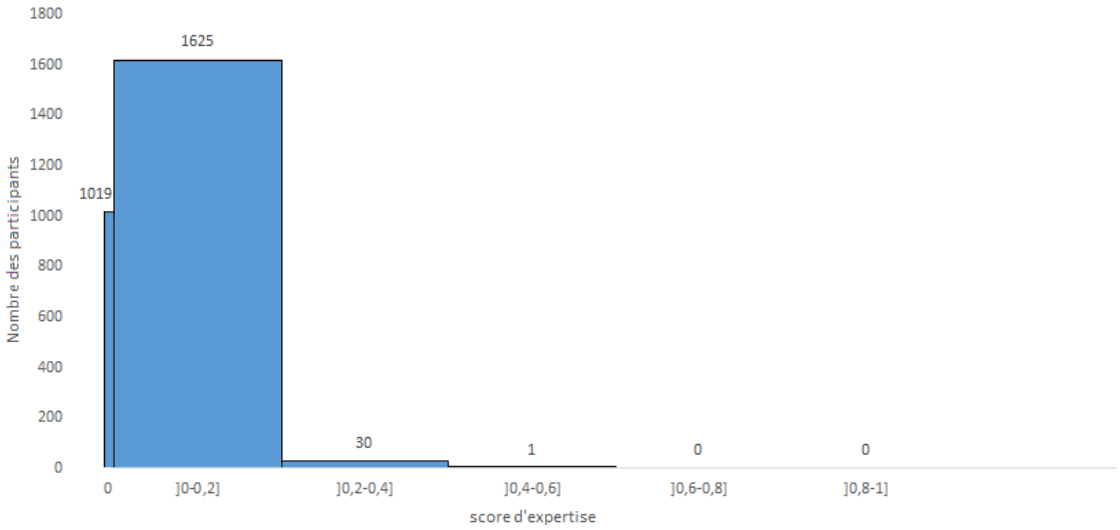

FIGURE 4.11 – Distribution des scores d'expertise des participants de la plus grande composante connexe du graphe de conversation

La figure 4.12 montre la distribution du score d'expertise des participants ayant publié un tweet ou une réponse. Nous remarquons qu'il y a 264 participants avec un score égal à

zéro ; cela s'explique par le fait qu'il y a des participants qui n'ont pas publié sur le même sujet auparavant et qui ne sont donc pas experts. Il y a encore un nombre important de participants avec un score très faible (inférieur ou égal à 0.2). Ceci s'explique par le fait que ces participants ont tendance à s'engager dans des sujets de conversation variés qui ne sont pas forcément liés à leur spécialité (centre d'expertise). Une explication provient également de la variation et de la vitesse de la diffusion des "sujets tendance" sur Twitter.

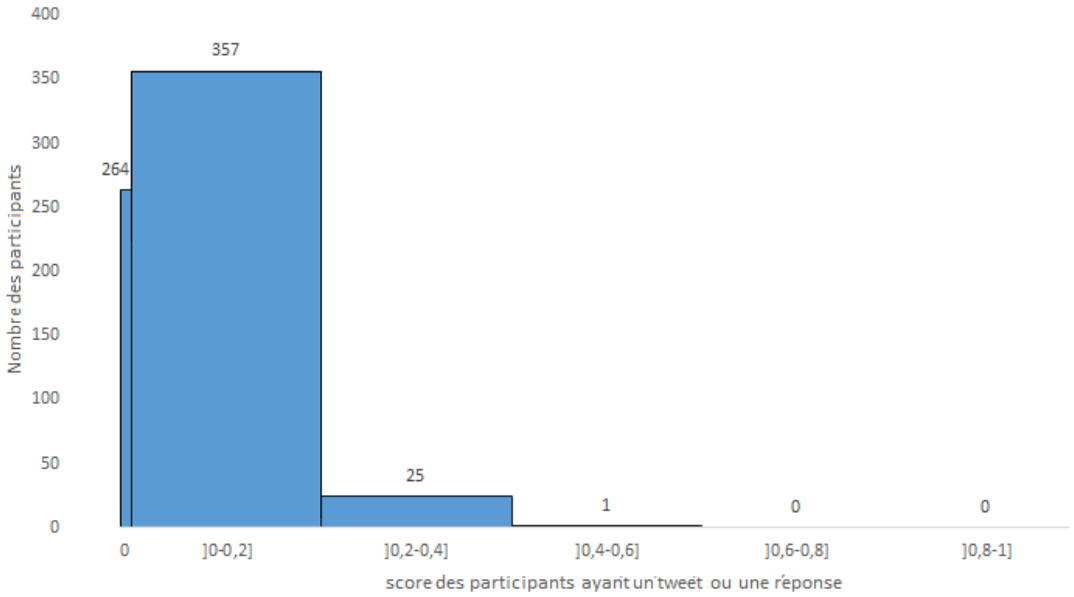

FIGURE 4.12 – Distribution des scores d'expertise des participants ayant publié un tweet ou une réponse de la plus grande composante connexe du graphe de conversation

– Score d'activité des participants

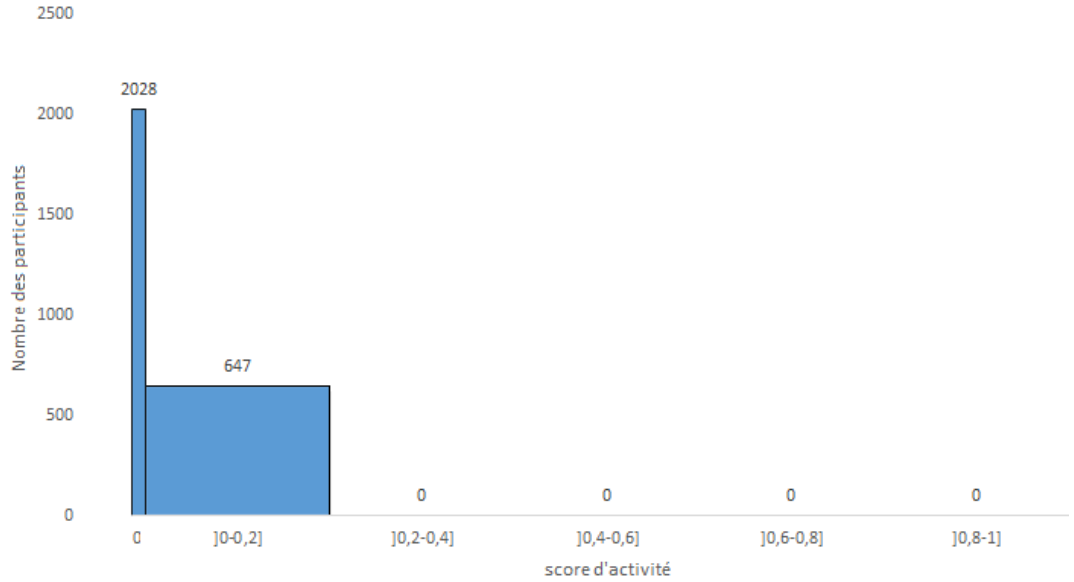

FIGURE 4.13 – Distribution des scores d'activité sur la plus grande composante connexe du graphe de conversation

La figure 4.13 montre la distribution du score d'activité de l'ensemble des participants. Nous remarquons qu'il y a un nombre très élevé de participants avec un score d'activité

nul. Ce score traduit qu'il y a un nombre important (environ 76 %) des participants qui ont participé à la conversation par un retweet ou une mention. Le score d'activité peut être important pour la formation du contexte en exploitant les participants qui sont actifs (qui ont publié ou qui ont répondu à un tweet). Si un participant a publié plusieurs messages dans la conversation, c'est qu'il est intéressé par le sujet et qu'il veut contribuer à la conversation.

– Score de pertinence des participants

La distribution du score de pertinence des participants est singulière : 2659 participants ont un score de pertinence égal à 0 et 16 ont un score inférieur à 0.2. La présence très élevée de participants avec un score nul est dû aux nombreux participants qui n'ont pas publié ou pas répondu à d'autres tweets dans la conversation.

– Score d'influence des tweets

La figure 4.14 montre la distribution des scores d'influence des tweets. Nous remarquons que 23 % des tweets sont non influents ; cela est expliqué par le fait que ces tweets n'ont pas eu des retweets ou des favoris. Cet indice peut être utile pour le contexte afin d'éliminer dans la restitution les tweets non influents, qui jouent un rôle mineur dans la conversation.

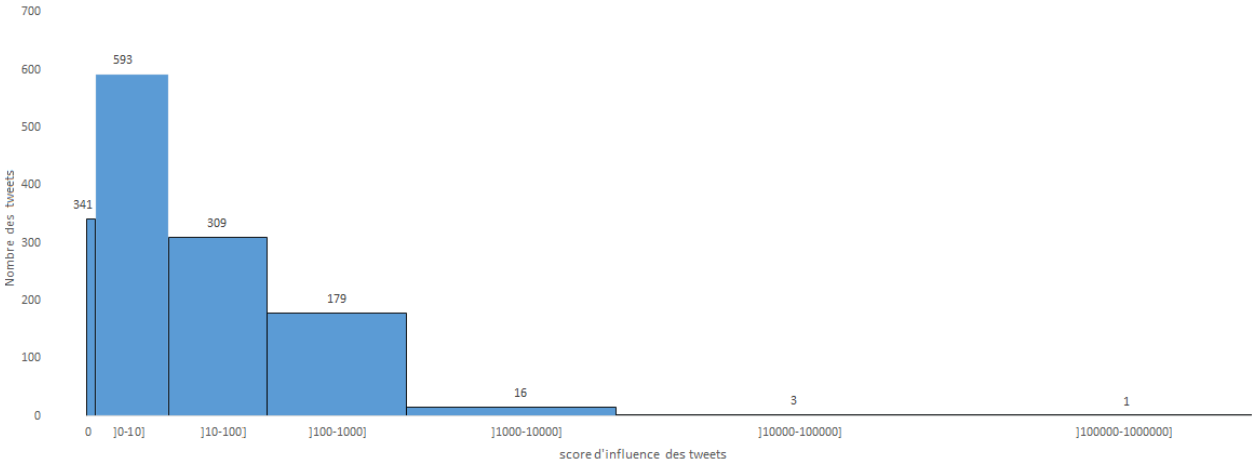

FIGURE 4.14 – Distribution des scores d'influence des tweets de la plus grande composante connexe du graphe de conversation

– Score de hashtag des tweets

Selon la figure 4.15, nous remarquons que presque 76 % des tweets contiennent un ou plusieurs hashtags avec un pic au niveau 1 (410 tweets). Les hashtags sont des marqueurs d'informations et plus le score de hashtag est élevé, plus le tweet est considéré comme informatif. Le score de hashtag peut donc être utile pour la contextualisation.

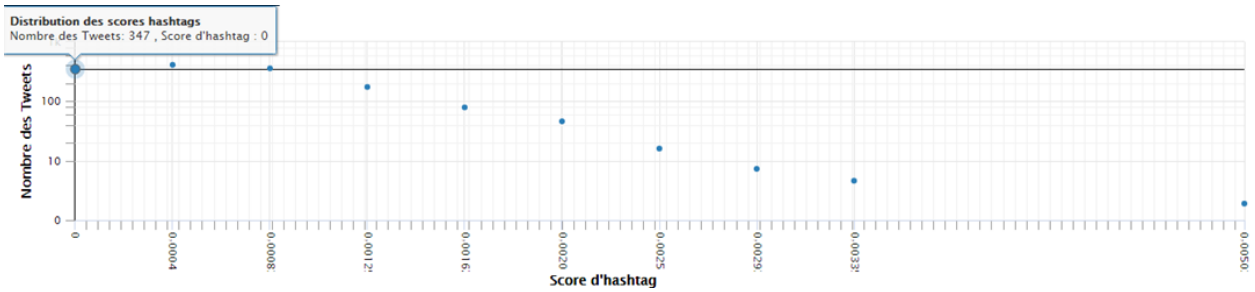

FIGURE 4.15 – Distribution des scores de hashtag des tweets de la plus grande composante connexe du graphe de conversation

– Score d'URL des tweets

Sur la figure 4.16, nous observons un pic au niveau du nombre des tweets qui contiennent exactement une seule URL. Les utilisateurs ont tendance à insérer des URLs dans leurs tweets pour confirmer la fiabilité de l'information ou pour enrichir l'information partagée. Ainsi, la présence d'une URL dans un tweet permet à l'utilisateur de mieux comprendre le contenu du tweet et le tweet peut être considéré comme plus informatif.

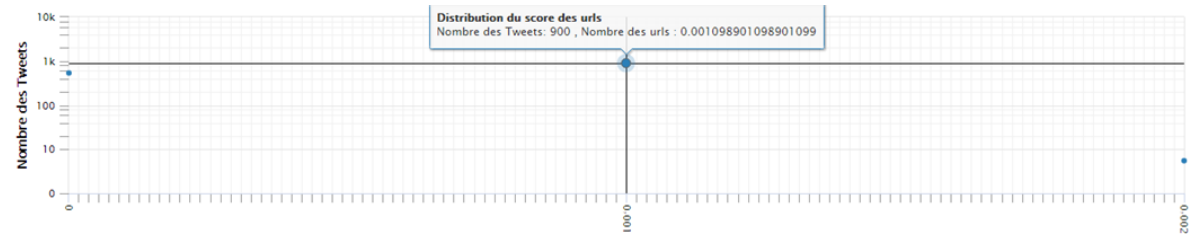

FIGURE 4.16 – Distribution des scores d'URL des tweets de la plus grande composante connexe du graphe de conversation

– Corrélation entre le nombre de retweets et le nombre de favoris

La figure 4.17 présente la corrélation entre le nombre de retweets et le nombre de favoris. Nous remarquons que la corrélation entre les deux scores est forte positive avec une forme non linéaire monotone et un sens non linéaire positif, ce qui justifie l'application de la combinaison linéaire pour le calcul du score d'influence. Cette corrélation nous permet de privilégier le nombre de retweets au nombre de favoris. Ceci permet de confirmer que les utilisateurs qui retweetent sont plus intéressés au tweet que d'autres qui se contentent de le mettre comme favoris.

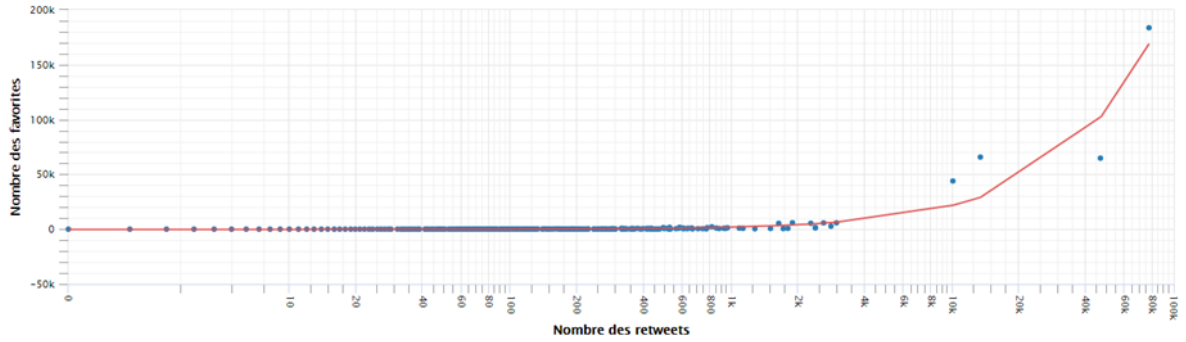

FIGURE 4.17 – Corrélation entre le nombre de retweets et le nombre de favoris de la plus grande composante connexe du graphe de conversation

– **Degré des sommets tweets**

Le degré d'un sommet est le nombre d'arcs incidents à ce sommet $\delta(v_i)$. Le degré d'un sommet est la taille de son voisinage (Greetham and Ward, 2014). La figure 4.18 montre que la distribution des degrés des sommets tweets une loi en forme de loi puissance. En effet, nous retrouvons tous les ingrédients d'un comportement associé à cette loi : effectif élevé des sommets de degré faibles (24 % des sommets ont un degré égal à 2 et 89 % ont un degré inférieur à 10) et effectif faible des sommets en queue de distribution avec un degré élevé.

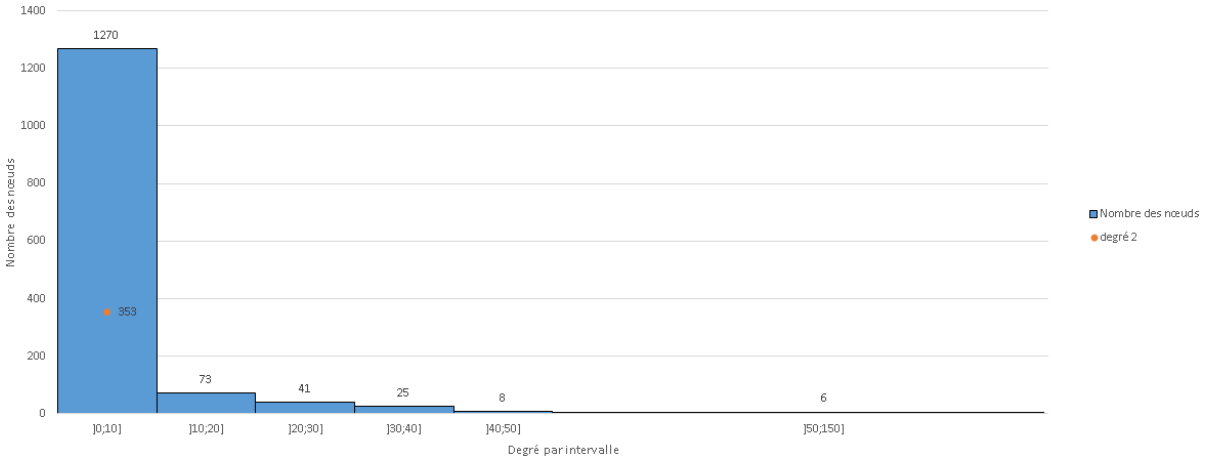

FIGURE 4.18 – Distribution des degrés des sommets tweets de la plus grande composante connexe du graphe de conversation

– **Degré des sommets participants**

La figure 4.19 montre la distribution des degrés des sommets participants. Nous remarquons qu'il y a un nombre très important des sommets participants (1609 sommets) avec un degré égal à 1 (environ 60 % du nombre des sommets participants). Nous constatons, aussi, que le nombre des sommets participants qui ont un degré supérieur à 10 est très faible.

L'analyse de la distribution des degrés nous montre la portée limitée de cet indicateur dans le cadre de la contextualisation. Il peut certes permettre d'identifier aisément les rares sommets -ceux à fort degrés- qui peuvent jouer le rôle de "hub" dans la diffusion

des informations. Mais, pour les autres acteurs impliqués, la distribution n'est finalement pas assez discriminante (beaucoup ont un degré faible).

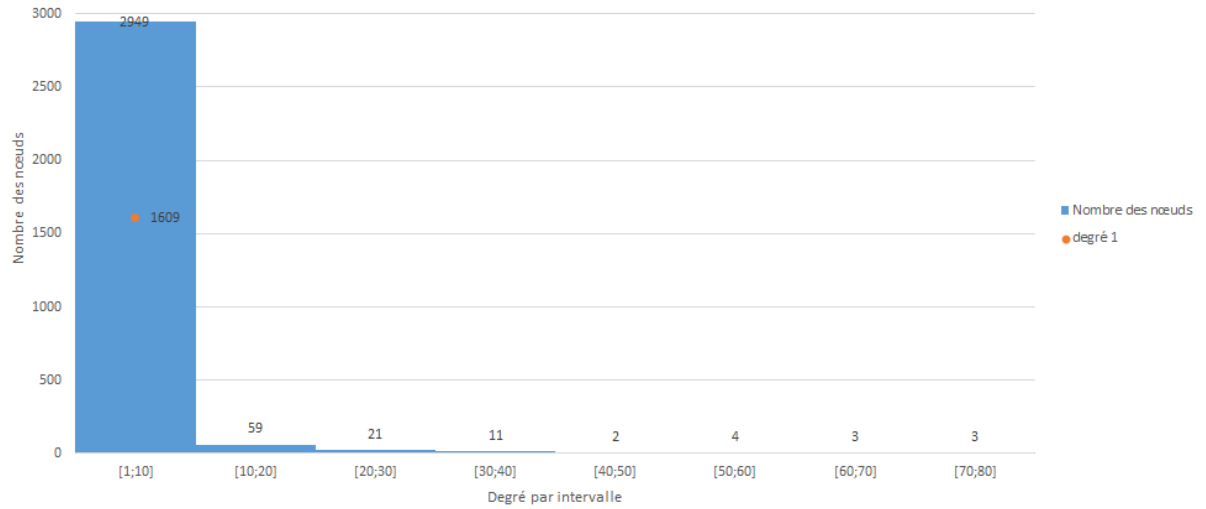

FIGURE 4.19 – Distribution des degrés des sommets participants de la plus grande composante connexe du graphe de conversation

– Degré de centralité des sommets tweets

La figure 4.20 montre la distribution des degrés de centralité des sommets tweets. Nous remarquons que près d'un quart des sommets ont un degré de centralité très faible. Sans surprise, la distribution est positive décroissante et se rapproche de la forme d'une loi sans échelle.

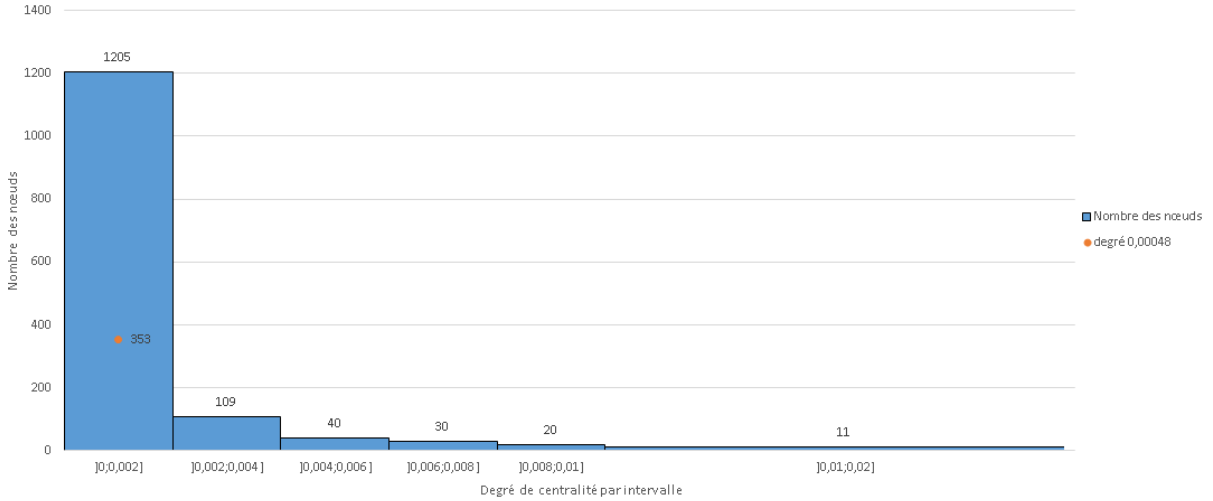

FIGURE 4.20 – Distribution des degrés de centralité des sommets tweets de la plus grande composante connexe du graphe de conversation

– Degré de centralité des sommets participants

La figure 4.21 représente la distribution des degrés de centralité des sommets participants. Nous remarquons que la majorité des sommets ont un degré de centralité très faible (proche de 0).

Sans surprise, le comportement de la distribution de degré de centralité est similaire à celle du degré que nous avons analysée précédemment. L'analyse de sa portée pour la contextualisation est donc identique et nous ne retenons pas dans la suite le degré de centralité parmi les indicateurs de filtrage.

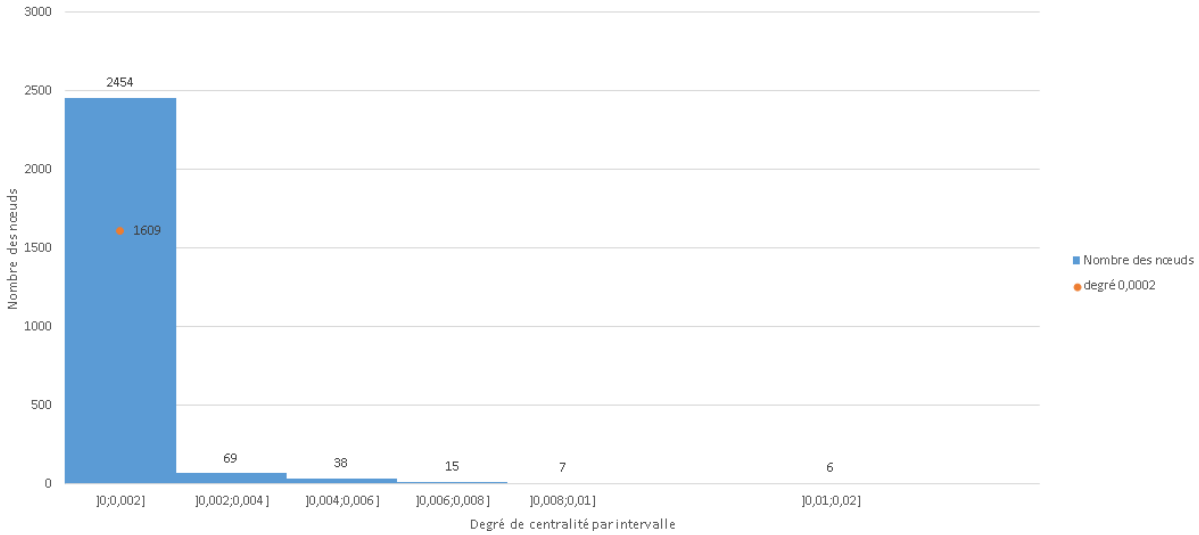

FIGURE 4.21 – Distribution des degrés de centralité des sommets participants de la plus grande composante connexe du graphe de conversation

4.4.2 Résultats expérimentaux observés sur un corpus de graphes de conversations

Dans cette partie, nous discutons les constats majeurs, basés sur l'ensemble de nos expérimentations, ainsi que les indicateurs retenus pour notre système de contextualisation. Dans le tableau 4.4, nous présentons, pour chaque requête, les détails des données analysées dans nos expérimentations.

Requêtes	Type de collecte	Date de la collecte	Nombre de composantes connexes	Taille de la plus grande composante connexe retenue
#IranProtests	mots-clés	décmebre 2017	480 composantes connexes	3919 sommets
#Hannity	mots-clés	novembre 2017	502 composantes connexes	2675 sommets
#NetNeutrality	mots-clés	décmebre 2017	924 composantes connexes	2314 sommets
#francois	mots-clés	août 2017	523 composantes connexes	3772 sommets
#Christmas	mots-clés	dècmebre 2017	566 composantes connexes	3602 sommets
#Trump's Twitter	mots-clés	novembre 2017	518 composantes connexes	3959 sommets
#Balkany	mots-clés	septembre 2017	501 composantes connexes	4329 sommets
#Google	mots-clés	septembre 2017	516 composantes connexes	3992 sommets

#80kmh	mots-clés	février 2018	704 composantes connexes	4778 sommets
#93	mots-clés	septembre 2017	424 composantes connexes	3212 sommets
#eFRAN	mots-clés	janvier 2018	523 composantes connexes	4528 sommets
#scmedias	mots-clés	janvier 2018	748 composantes connexes	3606 sommets
#Gallimard	mots-clés	septembre 2017	401 composantes connexes	3430 sommets
#avocats	mots-clés	octobre 2017	374 composantes connexes	4909 sommets
#WhistleEU	mots-clés	janvier 2018	402 composantes connexes	3303 sommets
#EXO	mots-clés	janvier 2018	453 composantes connexes	5298 sommets
#NowPlaying	mots-clés	octobre 2017	573 composantes connexes	2772 sommets
#BTSARMY	mots-clés	novembre 2017	693 composantes connexes	4520 sommets
#Healthcare	mots-clés	janvier 2018	315 composantes connexes	4944 sommets
#Job	mots-clés	janvier 2018	458 composantes connexes	3552 sommets
#Generation5G	mots-clés	février 2018	478 composantes connexes	3919 sommets
#SDE2018	mots-clés	février 2018	325 composantes connexes	2675 sommets
#SaferInternetDay	mots-clés	février 2018	528 composantes connexes	2314 sommets
#iHeartAwards	mots-clés	janvier 2018	369 composantes connexes	3212 sommets
#BestFanArmy	mots-clés	janvier 2018	423 composantes connexes	3772 sommets
#EXOL	mots-clés	janvier 2018	314 composantes connexes	3654 sommets
#TwitterBestFandom	mots-clés	janvier 2018	476 composantes connexes	2972 sommets
#changethedate	mots-clés	janvier 2018	285 composantes connexes	2782 sommets
#RIB2018	mots-clés	janvier 2018	386 composantes connexes	4256 sommets
#FormPro	mots-clés	janvier 2018	436 composantes connexes	3214 sommets

TABLE 4.4 – Variabilité et détails des données utilisées dans nos expérimentations

4.4.2.1 Propriétés structurelles des conversations

Nous commençons par une analyse de la structure des graphes de conversations (composantes connexes des graphes de conversations) construits à partir de 30 requêtes (sujets). Nous constatons, sur l'ensemble de 30 sujets, la présence d'un grand nombre de composantes connexes indiquant de nombreuses conver-

sations restreintes. Différents patterns apparaissent dans les composantes connexes. La grande majorité d'entre elles se réduisent à des chemins très courts. Elles représentent des conversations de va-et-vient entre souvent deux utilisateurs. D'autres composantes avec une plus grande cardinalité ont des patterns "en étoile" : un seul utilisateur génère un tweet auquel un grand nombre de personnes répondent. Enfin, les plus grandes composantes connexes ont une structure combinatoire beaucoup plus complexe qui se rapproche de celle étudiée dans l'illustration précédente. Ces propriétés concordent avec des analyses présentées précédemment dans la littérature (Tamine et al., 2015; Cogan et al., 2012; Kumar et al., 2010; Nishi et al., 2016). Les auteurs ont également mis en évidence la forte présente d'échanges très restreints se réduisant sur le graphe à des chemins de longueur faible. Cette forte présence est évaluée à 60% sur l'ensemble des graphes par (Cogan et al., 2012; Tamine et al., 2015).

4.4.2.2 Degré, taille et degré de centralité

Les distributions observées des sommets ressemblent à une loi puissance. Cette loi, déjà mise en évidence à la fin du XIX ème siècle par l'économiste italien Vilfredo Pareto, a été souvent retrouvée dans de nombreux réseaux qu'il s'agisse de réseaux physiques (réseaux de connexions corticales, Internet, réseaux de câblages électroniques, etc.) ou certains réseaux sociaux (Dorogovtsev and Mendes, 2003). Elle caractérise des réseaux avec une très forte diversité où notamment un grand nombre de sommets ont un degré faible et quelques sommets peuvent avoir des degrés très élevés. On parle aussi de "loi sans échelle" puisqu'ici la moyenne n'est plus un indicateur caractéristique -contrairement à la loi normale notamment- et que dans le cas théorique de la loi de Pareto la variance peut même être infinie. Dans notre cadre, il n'est pas surprenant de trouver une telle forme pour la distribution des degrés (cf.figure 4.22, figure 4.23). La majorité des sommets (type utilisateurs/ type tweets) ont un faible degré : ils ne communiquent qu'avec un très petit groupe d'utilisateurs, alors que peu d'utilisateurs ont un degré élevé (par exemple, des interactions avec plus de 10 utilisateurs). Cette forme a également été retrouvée par différents auteurs (ex. (Greetham and Ward, 2014)) et est attachée à la variété des conversations, où comme nous l'avons vu précédemment la majorité contient des échanges courts.

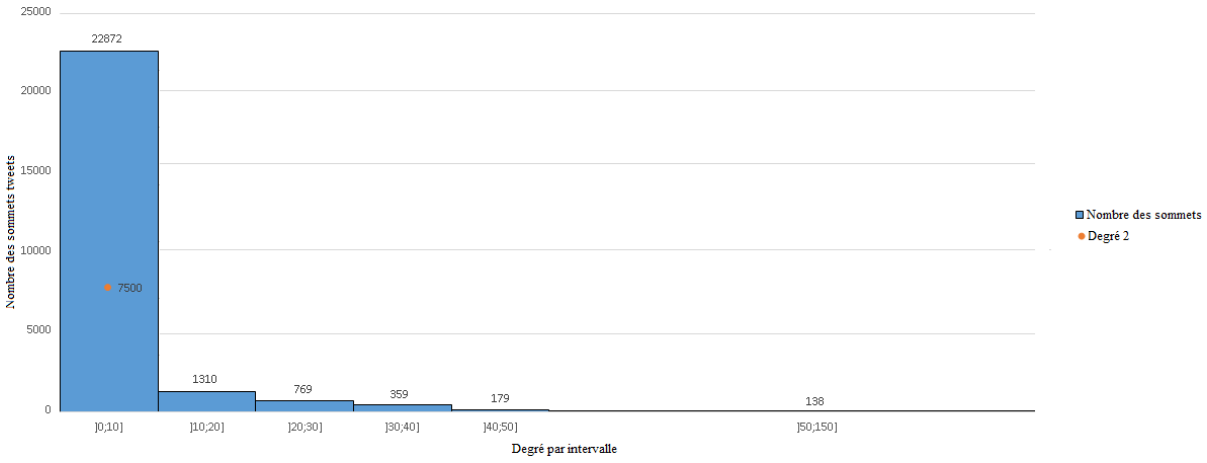

FIGURE 4.22 – Distribution des degrés des sommets tweets des plus grandes composantes connexes du graphe de conversation (30 sujets)

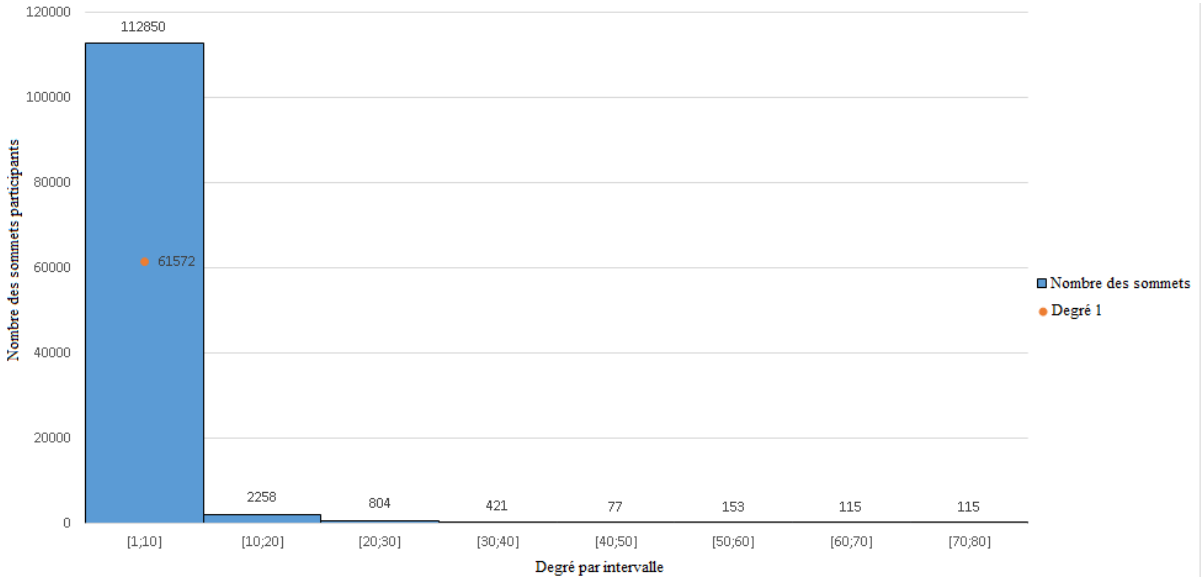

FIGURE 4.23 – Distribution des degrés des sommets participants des plus grandes composantes connexes du graphe de conversation (30 sujets)

Le degré de centralité des sommets, ou l'identification des sommets les plus "centraux" dans le graphe, a souvent été utilisée dans l'analyse des réseaux sociaux (Degenne and Forsé, 2004). Rappelons que le degré de centralité indique l'importance relative d'un sommet dans un graphe en fonction du nombre de ses liens. Dans notre cadre, le degré de centralité vise à identifier les participants fréquemment retweetés, mentionnés auxquels on a fréquemment répondu. Ces participants reçoivent des informations et les diffusent vers un grand nombre de participants. Au contraire, les participants qui ont un petit nombre de relations ne peuvent pas influencer directement un grand nombre de participants du graphe. Notre analyse montre que la quasi-totalité des sommets dans les graphes ont un degré de centralité très faible et que l'on retrouve à nouveau sans surprise une distribution des degrés de centralité qui suit les caractéristiques d'une loi puissance (cf.figure 4.24).

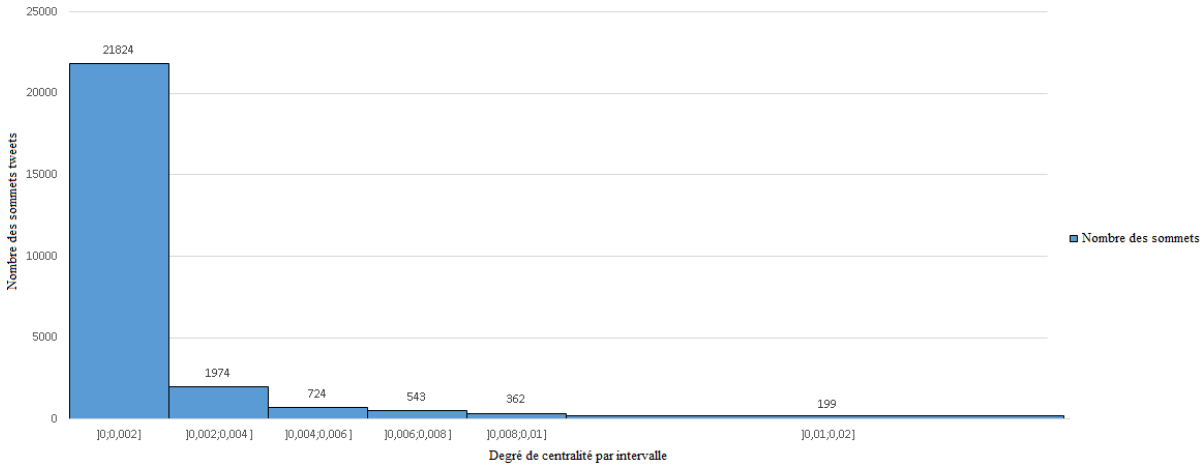

FIGURE 4.24 – Distribution des degrés de centrailité des sommets tweets des plus grandes composantes connexes du graphe de conversation (30 sujets)

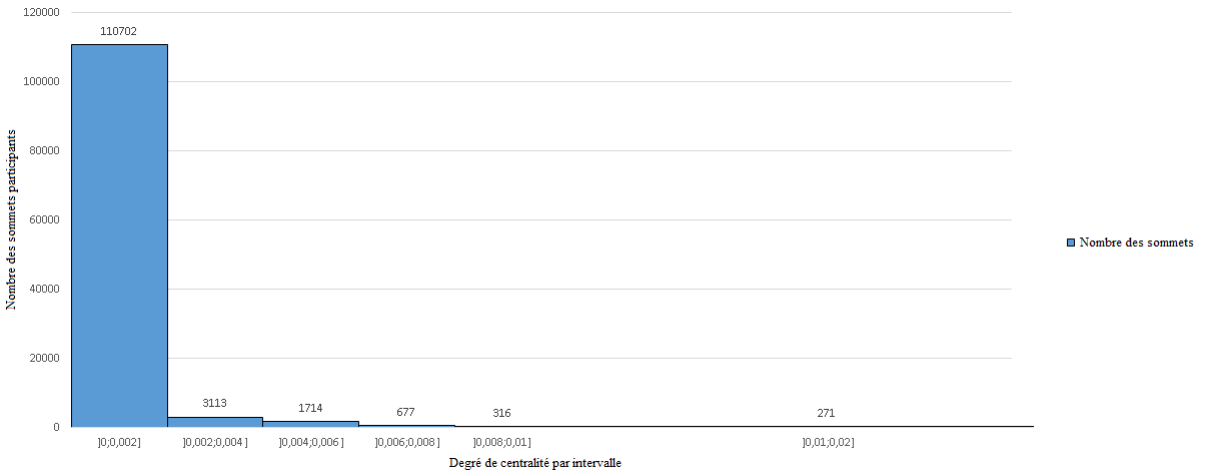

FIGURE 4.25 – Distribution des degrés de centraillité des sommets participants des plus grandes composantes connexes du graphe de conversation (30 sujets)

4.4.2.3 Métadonnées sociales : hashtags # et URLs

Comme nous l'avons déjà indiqué, le hashtag est considéré comme un marqueur de sujet, une indication du contexte du tweet ou de l'idée centrale exprimée dans le tweet. L'adoption du même hashtag par d'autres utilisateurs indique le partage d'un contenu similaire ou exprime une idée connexe. Les hashtags sont souvent utilisés pour élargir le public potentiel d'un tweet, de sorte qu'un tweet contenant un hashtag devrait s'attendre à un public plus large que la normale. Une conversation sur Twitter peut émerger à mesure qu'un nombre croissant d'utilisateurs adopte le même hashtag ou des hashtags similaires autour d'un sujet. Notre analyse a confirmé que la majorité des tweets contiennent des hashtags ; cela permet d'élargir la taille de l'audience de ces tweets et de les rendre plus informatifs (cf.figure 4.26). Les travaux de (Alvarez-Melis and Saveski, 2016; Mehrotra et al., 2013) ont montré la grande utilité des hashtags dans le cadre de la recherche d'information : ils ont notamment prouvé expérimentalement que leur utilisation apporte une amélioration au niveau des résultats de recherche.

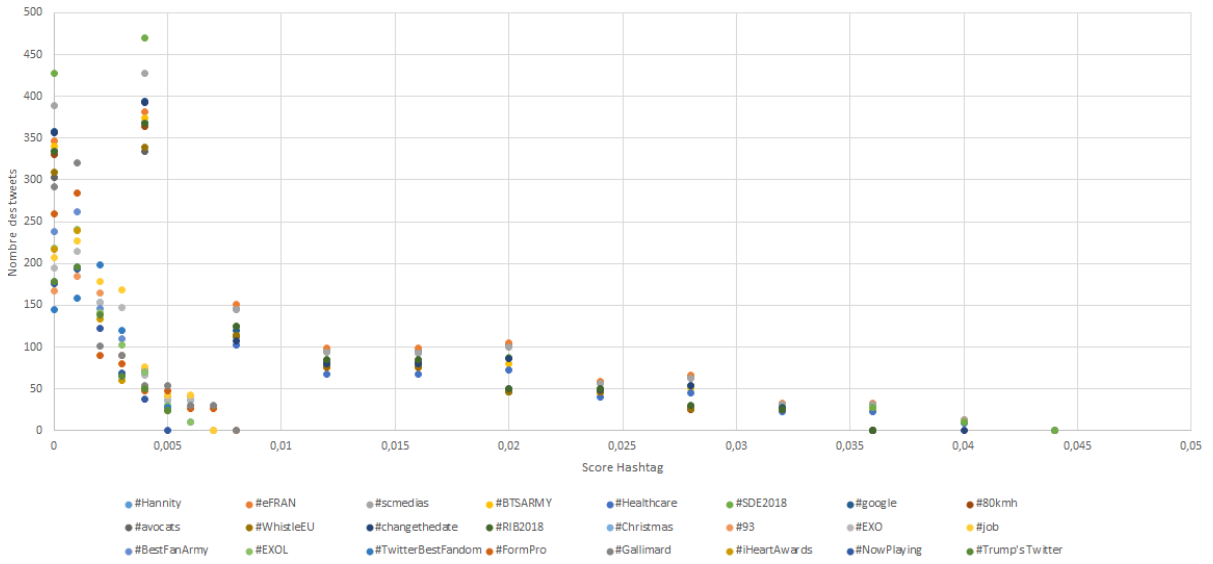

FIGURE 4.26 – Distribution des scores d'hashtag des tweets des plus grandes composantes connexes du graphe de conversation (30 sujets)

Nous avons aussi étudié la présence des URLs dans les tweets. Le résultat montre que la majorité des tweets contiennent une URL dans leur texte (cf.figure 4.27). La grande utilisation des URLs est due à

la limite de la taille des messages imposée par Twitter. Les URLs dans le tweet permettent d'augmenter sa visibilité et leur présence suggère donc qu'un tweet avec des URLs est plus informatif. De plus, les URLs permettent non seulement de fournir des informations supplémentaires pour les internautes, mais présentent également un vocabulaire complémentaire qui pourra être utilisé pour faciliter la compréhension de contenu de tweet. Ces différents éléments ont montré leur intérêt en recherche d'information (Amer et al., 2015) où le contenu court du tweet est un frein pour les algorithmes traditionnels. Nous les considérons donc également comme importants pour la contextualisation.

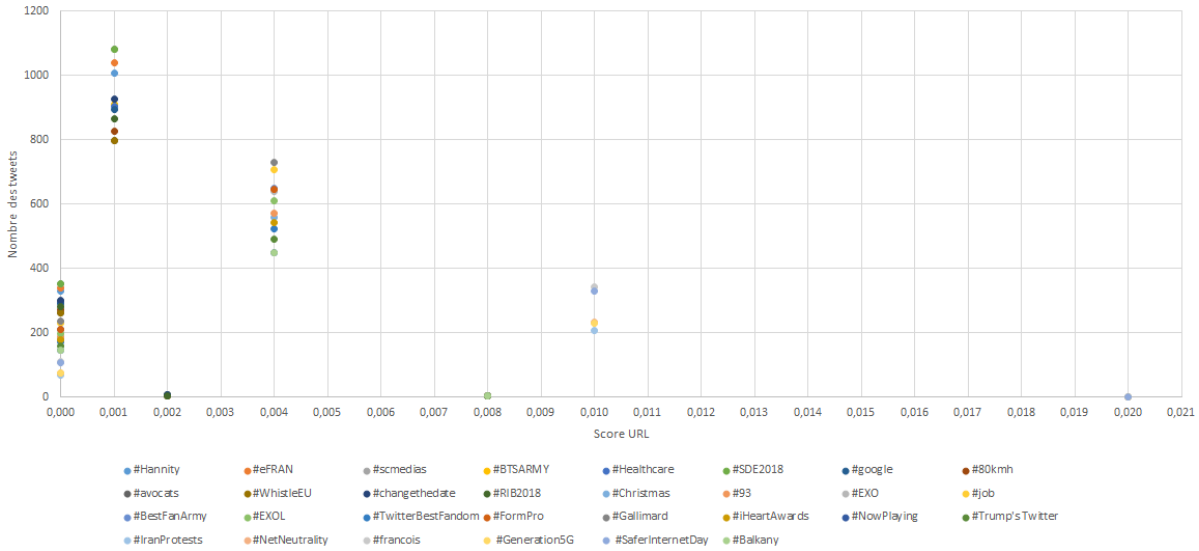

FIGURE 4.27 – Distribution des scores d'URLs des tweets des plus grandes composantes connexes du graphe de conversation (30 sujets)

4.4.2.4 Score d'influence des participants

Le calcul opérationnel du score d'influence des participants (formule 3.5 de la section 3.5.4) dépend d'un paramètre (β). Afin de s'appuyer sur une estimation de leur valeur pour l'extraction d'informations pertinentes pour la contextualisation, nous avons effectué plusieurs essais. Il ressort des expérimentations qu'une valeur adaptée pour β se situe autour de 0.7. Ceci dénote le fait que le nombre de retweets est très important par rapport au nombre de favoris. Ce qui confirme l'hypothèse que les utilisateurs qui retweetent sont plus intéressés au tweet que d'autres qui se contentent de le mettre comme favoris.
La figure 4.28 présente un histogramme illustrant la distribution de la moyenne des scores d'influence des participants des plus grandes composantes connexes du graphe de conversation. Les points de couleur sur la figure représentent les différents sujets- à partir desquels les requêtes ont été effectuées pour la construction des graphes de conversations- et les scores d'influence moyens associés. Les écarts-types sont ici très faibles ; ce qui indique une invariance comportementale parmi les discussions associées aux différents sujets traités. Comme nous l'avions déjà remarqué dans notre analyse illustrative, la figure 4.28 confirme que la majorité des participants sont non influents et que leurs messages n'ont pas attiré l'attention d'autres participants. L'analyse nous conduit à choisir un seuil égal à 10 pour filtrer le score d'influence pertinent pour la contextualisation.

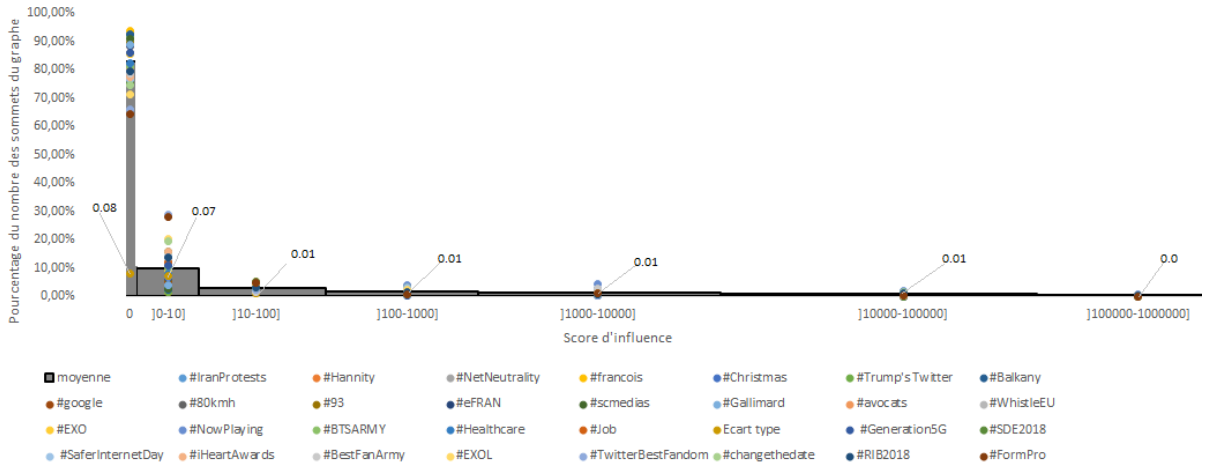

FIGURE 4.28 – Distribution de la moyenne des scores d'influence des participants des plus grandes composantes connexes du graphe de conversation (30 sujets)

4.4.2.5 Score de popularité des participants

Dans le cadre de notre travail, nous adoptons le critère de popularité pour estimer la qualité des participants. En d'autres termes, intuitivement, plus le public de l'auteur est grand, plus il est probable que le tweet obtienne une large audience. En se basant sur les résultats de nos expérimentations, présentés dans la figure 4.29, nous constatons que la majorité des participants dans le graphe de conversation sont populaires. Ce résultat suggère que, plus l'utilisateur est populaire, plus ses tweets sont intéressants, probablement en raison de la diversité des éléments d'opinions et d'informations consommés par l'utilisateur. Ainsi, l'analyse nous conduit à choisir un seuil égal à 1000 followers afin de filtrer les participants populaires pour la contextualisation.

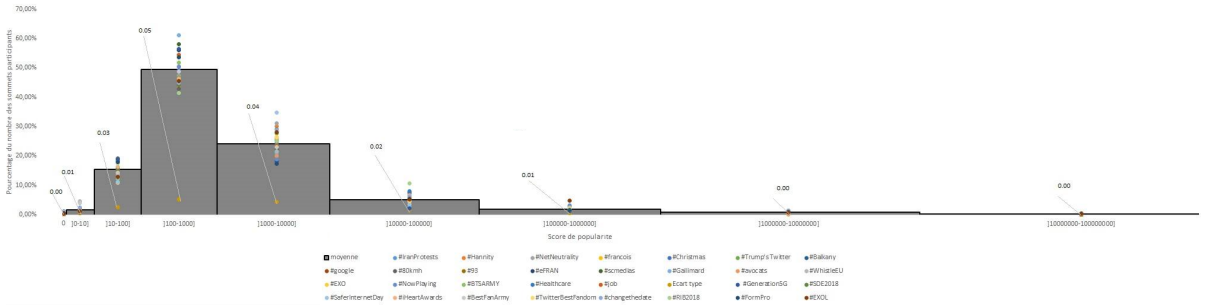

FIGURE 4.29 – Distribution de la moyenne des scores de popularité des participants des plus grandes composantes connexes du graphe de conversation (30 sujets)

4.4.2.6 Score d'expertise des participants

(Amer et al., 2015) ont utilisé ce critère dans leur travail pour estimer l'expertise des participants qui peut être un facteur de pertinence, car si ces participants sont experts alors, à priori, la conversation pourrait être intéressante et de qualité. Le même objectif est derrière l'utilisation de ce score dans notre travail.

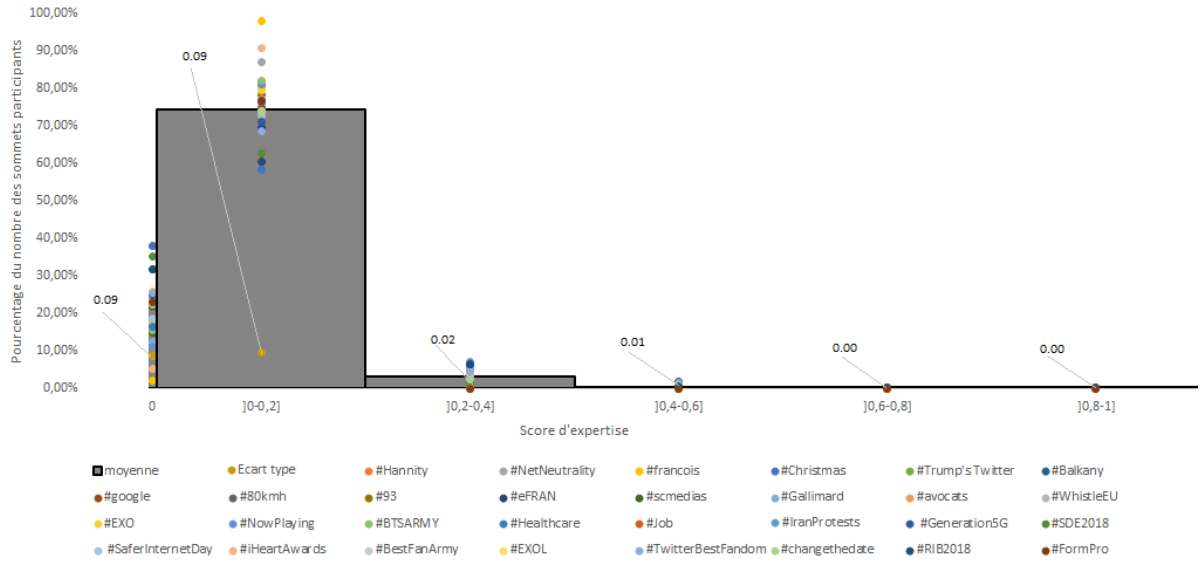

FIGURE 4.30 – Distribution de la moyenne des scores d'expertise des participants des plus grandes composantes connexes du graphe de conversation (30 sujets)

La figure 4.30 présente un histogramme illustrant la distribution de la moyenne des scores d'expertise des participants des plus grandes composantes connexes du graphe de conversation. Comme nous l'avions déjà remarqué dans notre analyse illustrative, la figure 4.30 confirme que le nombre majeur des participants ont un score d'expertise faible. L'analyse nous conduit à choisir un seuil égal à 0,2 pour le filtrage des participants experts.

4.4.2.7 Score d'activité des participants

La figure 4.31 confirme également les résultats mis en évidence sur le graphe de conversation illustratif : de très nombreux participants ont un score d'activité nul et le score de popularité à 0.2, déjà précédemment observé, semble un seuil pertinent pour le filtrage des participants actifs pour la contextualisation.

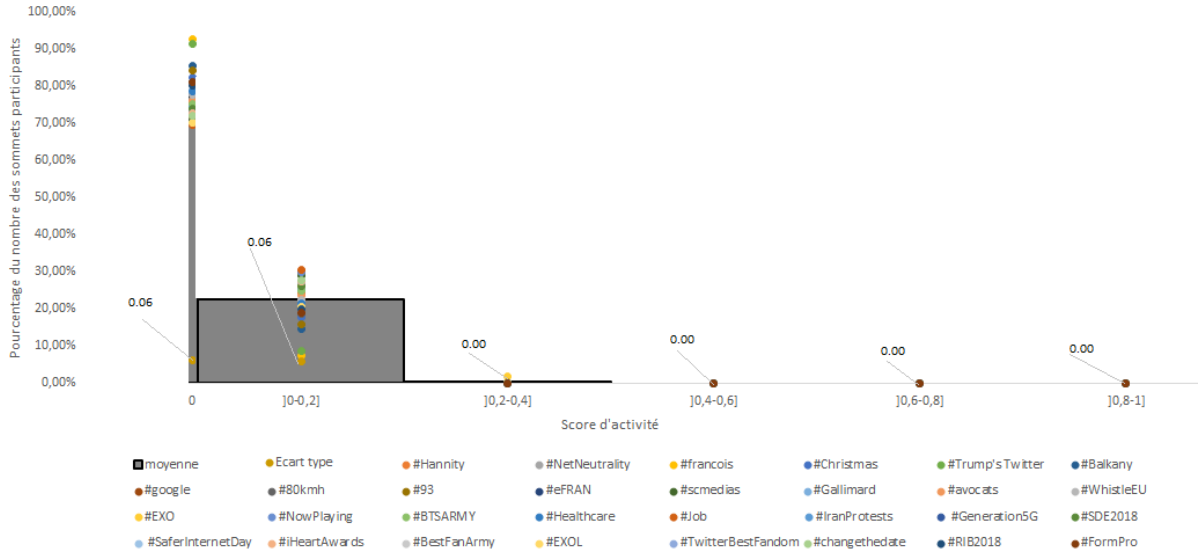

FIGURE 4.31 – Distribution de la moyenne des scores d'activité des participants des plus grandes composantes connexes du graphe de conversation (30 sujets)

4.4.2.8 Score d'influence des tweets

Le calcul du score d'influence des tweets dépend d'un paramètre (paramètre α dans la formule 3.3 section 3.5.3). Dans une démarche similaire à celle menée pour le calcul du score d'influence des participants nous avons effectué plusieurs tests pour estimer une valeur pertinente. Les résultats présentés sur la figure 4.32 -où nous utilisons les mêmes conventions que pour la figure 4.28 montrent que la grande majorité des tweets présents dans les graphes de conversations peuvent être considérés comme influents. Le seuil de score à 10 pour le filtrage permet d'éliminer les tweets non influents et nous conservons cette valeur pour la contextualisation.

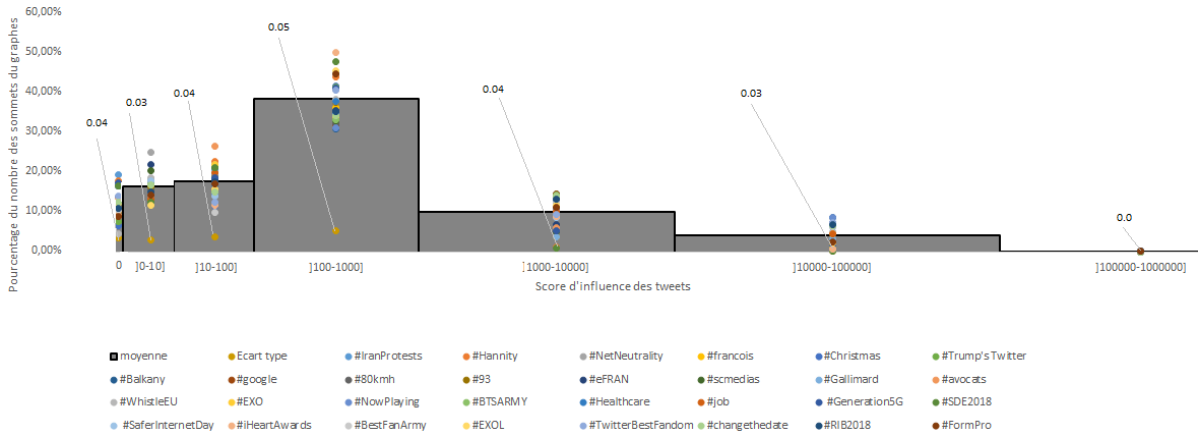

FIGURE 4.32 – Distribution de la moyenne des scores d'influence des tweets des plus grandes composantes connexes du graphe de conversation (30 sujets)

4.5 Conclusion

Dans ce chapitre, nous nous sommes concentrés sur les analyses expérimentales afin de pouvoir identifier les indices d'analyse pertinents pour extraire des graphes de conversations les informations utiles pour la contextualisation. Des corrélations positives entre des paires d'indices nous ont permis de réduire le nombre d'indices retenus et certaines distributions très hétérogènes ont confirmé la présence de variations très importantes dans les rôles respectifs joués par les différentes composantes du graphe, qu'il s'agisse des tweets ou des participants. Nous avons également analysé les distributions des indices retenus pour définir les seuils de valeur à choisi pour chacun d'eux pour le filtrage des informations. L'intégration opérationnelle de ces informations dans la contextualisation est présentée au chapitre suivant.

5

Contextualisation basée sur les conversations

Sommaire

5.1 Introduction 62
5.2 Approche proposée : processus de contextualisation 63
5.2.1 Construction des requêtes 63
5.2.2 Collecte des tweets et construction des graphes de conversations 64
5.2.3 Indicateurs d'analyse 64
5.2.4 Génération du contexte 66
5.3 Contextualisation des tweets : notre système proposé 68
5.3.1 Architecture générale du système 69
5.3.2 Contextualisation des tweets : Exemples de scénarios d'utilisation 71
5.4 Conclusion 77

5.1 Introduction

Dans ce chapitre, nous exploitons les propriétés des graphes de conversations mises en évidence dans le chapitre précédent pour proposer une nouvelle approche de contextualisation d'un tweet ambigu. La ligne directrice de la représentation finale du contexte fournie s'inspire partiellement de celle proposée dans le travail de (Abel et al., 2011) qui propose de fournir des liens connexes liés au tweet mais nous l'avons complétée et adaptée à notre problématique. L'approche développée par Abel et al. est basée sur la mise en correspondance des tweets avec des articles d'actualité, suivie d'un enrichissement sémantique basé sur le contenu de l'article. Cependant, dans notre contextualisation, nous ne supposons aucun lien entre les tweets et les articles d'actualité. Ici le tweet est associé à un graphe de conversation dans lequel nous extrayons des informations sur les noeuds et les relations pour les restituer à l'utilisateur via une interface interactive.
Le chapitre est organisé en deux parties. La première décrit les différentes étapes du processus mis en oeuvre pour la contextualisation. La deuxième présente le système développé pour l'opérationnalisation de ce processus.

5.2 Approche proposée : processus de contextualisation

Considérons un tweet *t*. Le processus de contextualisation de ce tweet est structuré selon quatre phases successives :

1. Construction des requêtes ;
2. Collecte des tweets et construction des graphes de conversations ;
3. Calcul des indicateurs d'analyse ;
4. Génération du contexte.

Dans la suite nous précisons ces différentes phases et argumentons nos choix.

5.2.1 Construction des requêtes

La première phase que nous effectuons consiste à appliquer un ensemble de pré-traitements aux tweets. Il s'agit de formater le contenu de ces derniers en vue de l'étape de recherche d'information. La question de recherche dans cette étape est la suivante :
Quelle est l'unité d'information la plus appropriée pour répondre aux besoins en informations (le contexte) ?
Pour résoudre ce problème, nous procédons sur deux niveaux :

1. Réduire le niveau d'ambigüité d'un tweet

   Nous avons décidé de traiter uniquement les tweets provenant des comptes d'information (par exemple @CNN), afin d'éviter les tweets purement personnels qui ne peuvent être contextualisables.

2. Choix des unités d'informations

   Le symbole #, appelé hashtag (ou mot-dièse en français), est un caractère utilisé pour préfixer les mots-clés dans les tweets. Les réseaux sociaux l'utilisent largement pour indiquer des mots, des groupes ou des sujets pertinents. Les hashtags sont des éléments d'informations très importants qui peuvent être directement fournis par les auteurs des tweets. Citons par exemple (Deveaud and Boudin, 2013b) qui ont exploité la prolifération des hastags pour améliorer la récupération des articles Wikipédia pertinents en RI. D'autres approches ont proposé de regrouper les tweets en fonction des hashtags qu'ils contiennent (Efron, 2010). De plus, les hashtags peuvent fournir un contexte implicite. Plusieurs études ont mis en évidence l'effet de ces éléments. Par exemple, dans une analyse sur la sémantique des hashtags, (Huang et al., 2010) ont montré que le "hashtagging" sur Twitter est plus communément utilisé pour se joindre aux discussions publiques que pour organiser le contenu à récupérer.

   Dans notre approche, nous considérons les hashtags, qui sont souvent associés à des noms, des événements ou des lieux, comme les mots-clés du tweet, car ils sont normalement des noms, des événements ou des lieux. De plus, nous complétons cette information par les mentions. Certaines statistiques indiquent que plus de 50 % des tweets contiennent une @mention qui fait référence à un utilisateur pour porter le tweet à son attention (Carter et al., 2013).

Ainsi, pour un tweet initial *t* nous considérons l'ensemble des hashtags et des mentions comme les unités d'informations importantes que nous utilisons comme requête pour construire les graphes de conversations. Plus précisément, la chaîne de pré-traitement est présentée par la figure 5.1 : pour chaque tweet *t*, nous appliquons un processus de formatage dans lequel nous supprimons tous les retweets (RT) et les stopwords (sur la base du stoplist de inquery standard). En sortie, le tweet *t* est décrit par l'ensemble des hashtags et des mentions qu'il contient (cf.figure 5.2).

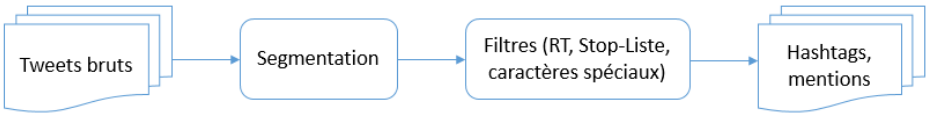

FIGURE 5.1 – Chaîne de pré-traitements des tweets bruts et construction de requête

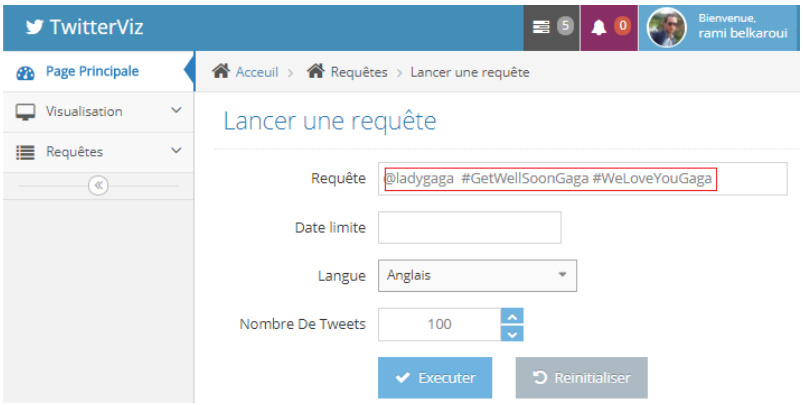

FIGURE 5.2 – Exemple de construction d'une requête à partir d'un tweet initial

5.2.2 Collecte des tweets et construction des graphes de conversations

Le contenu de la requête construite à l'étape précédente permet de collecter les tweets puis de construire le graphe de conversation. L'algorithme de construction du graphe est inspiré de celui proposé par (Cogan et al., 2012). Cependant dans leur cadre, le graphe est restreint à la prise en compte de la relation de réponse entre les tweets. Comme nous avons étendu la définition d'un graphe de conversation afin de combiner les informations sur les tweets et les participants, nous avons dû adapter l'algorithme de construction du graphe. Notre approche enchaîne deux étapes principales :

À partir d'un corpus D de tweets, la première étape consiste à partir du champ "$in-reply-to-status-id$", à déterminer l'ensemble des tweets $tweet_{0_1}$, $tweet_{0_2}$, $tweet_{0_K}$ qui sont les initiateurs des conversations ; chaque tweet $tweet_{0k}$ est «racine» d'un graphe de conversation G_{C_K}. Le même processus se fait automatiquement sur les sommets participants en exploitant le champ "$screen-name$" pour identifier les auteurs des tweets. Chaque participant (auteur) est relié à son tweet par la relation réponse. Cette étape se fait d'une manière récursive.
À chaque fois qu'un tweet racine $tweet_{0k}$ est identifié, la deuxième étape s'enclenche pour construire le graphe de conversation G_{C_i} de façon descendante à partir de chaque tweet racine $tweet_{0_k}$. Un balayage du corpus D permet d'identifier les tweets qui sont une réponse au $tweet_{0_k}$ (niveau 1) en vérifiant le champ "$in-reply-to-status-id$" de chaque tweet. Ainsi, les relations "mentionner", "retweeter" et "favori" sont identifiées en exploitant les champs suivants : "$user-mentions$", "$retweeted-status$" et "$favorited$". Ce même processus se répète à chaque fois pour identifier les autres niveaux de graphe en balayant l'ensemble restant des tweets. À l'issue de cette phase, nous obtenons un graphe de conversation (cf. figure5.3).

5.2.3 Indicateurs d'analyse

À partir des analyses effectuées dans le chapitre 4, nous avons retenu un sous-ensemble d'indicateurs qui permettent de filtrer les entités pertinentes dans les graphes de conversation (figure 5.4). Ces indices sont structurés en trois familles : les indicateurs portant sur les sommets participants, les indicateurs portant sur les sommets tweets, et les indicateurs structurels généraux portant sur l'ensemble des sommets du graphe.

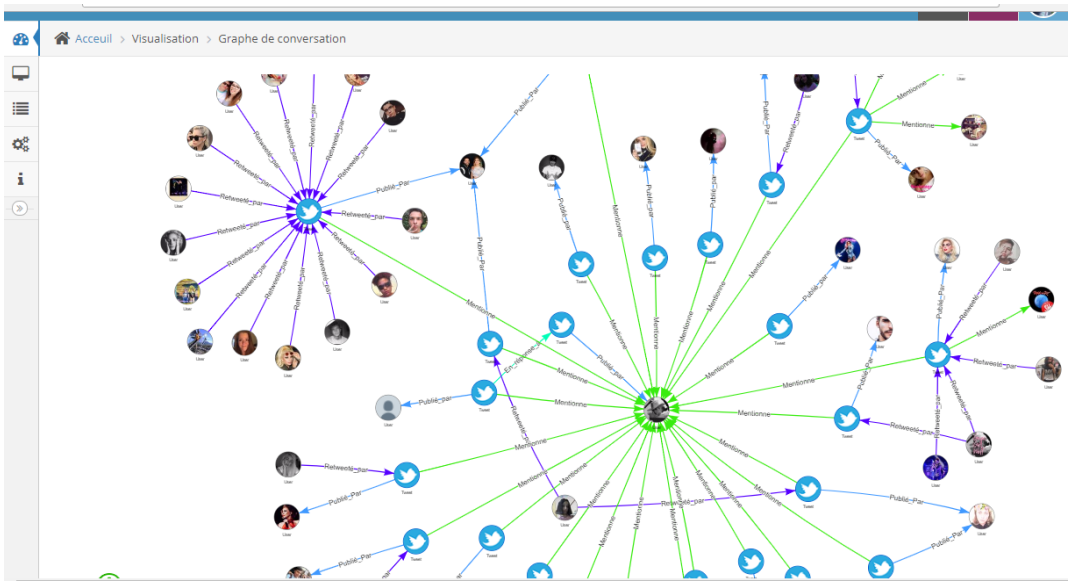

FIGURE 5.3 – Exemple d'un graphe de conversation

– **Qualité des participants des conversations**. Nous estimons la qualité des participants à travers leur popularité, leur influence, leur activité et leur expertise.

– **Qualité des tweets des conversations**. La qualité d'un tweet dans le graphe de conversation tient compte des méta-données qui sont les URLs, les hashtags, les mentions qu'il contient ainsi que son score d'influence qui traduit le nombre de fois que ce tweet a été retweeté ou/et mis en favoris. Nous retenons pour l'estimation de la qualité : le score hashtag des tweets de la conversation, le score URL, le score de mention, le score d'influence et l'horodatage.

– **Structure générale de la conversation**. Seule la plus grande composante connexe est conservée. Les autres indicateurs permettent d'affiner l'analyse du graphe comme nous le montrons dans les scénarios de la section 5.3.2.

FIGURE 5.4 – Liste des indicateurs d'analyse

5.2.4 Génération du contexte

Comme nous l'avons souligné dans le chapitre 2, il existe de nombreuses définitions de contexte. Nous proposons ici de nous appuyer sur la définition suivante : le contexte d'un document représente toute connaissance ou informations pertinentes caractérisant le document et ayant une liaison jugée importante avec ce dernier. Ici, le document textuel est le tweet, et le contexte d'un tweet peut être défini par les informations multidimensionnelles liées au tweet et récupérées à partir des conversations sur Twitter (interactions, termes connexes, lien hypertexte, entités, personnes, etc.). La pertinence des informations est mesurée par les indices présentés précédemment. Cette définition ne reprend pas toutes les ouvertures proposées dans la littérature mais elle est opérationnelle et nous semble réunir les différentes composantes nécessaires pour notre objectif.

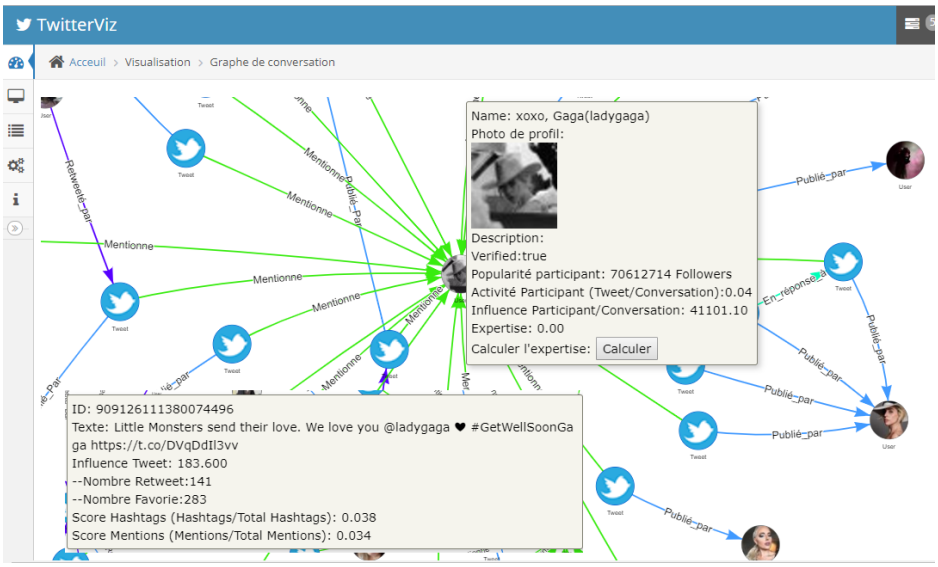

FIGURE 5.5 – Calcul des indicateurs pour chaque sommet

Cependant, pour sa mise en oeuvre effective, la question fondamentale qui se pose alors est la sélection des dimensions du contexte à prendre en compte. Pour y répondre, nous nous appuyons sur les travaux de Tamine-Lechami (Tamine-Lechani, 2008) eux-mêmes inspirés par le principe de la poly-représentation développée par Ingwersen (Ingwersen, 1994). Le principe de la poly-représentation considère que les documents peuvent être classés dans un espace composé par un ensemble de variables multiples (structure du document, genre du document, l'auteur du document, style du contenu, les hyperliens, etc.). Cette classification vise à augmenter la probabilité que ces documents soient utilisés dans certaines situations de recherche/contexte de recherche. Parmi les dimensions proposées par Tamine-Lechani, nous en avons retenu trois auxquelles nous avons ajouté une dimension temporelle (figure 5.6) :

- La représentation du document. Cette dimension concerne les éléments textuels tels que la longueur du texte et les méta-données sociales (hashtags, URLs et les mentions).
- La source du document. L'auteur représente la source du tweet. Cette dimension concerne les caractéristiques de la source tels que l'expertise de l'auteur, le genre de l'auteur (centres d'intérêt et ses connaissances), sa popularité, son activité, son degré d'influence, etc.
- La qualité de l'information. Elle concerne sa pertinence et sa visibilité (degré d'influence, nombre de retweets, nombre de favoris).

- La fraîcheur de l'information. Elle permet de définir le contexte temporel et concerne le temps de publication par rapport à un sujet ou à un besoin d'information.

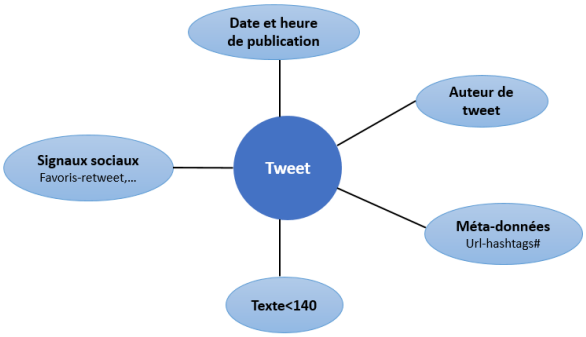

FIGURE 5.6 – Modélisation du tweet

Ces différentes informations sont mobilisées pour la représentation finale proposée à l'utilisateur. Dans ce qui suit, nous présentons les informations qui sont affichées sur les sommets pertinents du graphe en précisant l'association de chaque information à sa dimension.

- Le profil Twitter (y compris sa photo) : l'utilisateur peut consulter le profil de l'auteur du tweet (via un lien direct vers le profil) pour avoir plus de détails sur l'auteur.

- La description du profil : cette description donne une information sur la biographie du participant. Dans la bibliographie, les utilisateurs décrivent généralement leurs activités, leurs domaines d'expertise et leurs centres d'intérêt.

- La popularité (nombre de followers) : elle caractérise l'importance associée à la lecture des tweets d'un participant populaire.

- L'expertise : elle caractérise l'importance associée à la lecture des tweets d'un participant expert sur le sujet.

- L'influence : un tweet publié par un participant influent est plus susceptible d'attirer l'attention des autres participants.

  Cet ensemble d'informations correspond à la dimension source du document.

Pour les sommets tweets, nous affichons les informations suivantes :

- Le texte : pour prendre évidemment connaissance du contenu textuel.

- Les hashtags : le lecteur du tweet peut ne pas être familier avec les hashtags inclus dans le texte du tweet. Pour l'assister dans sa compréhensions, nous recueillons leurs descriptions en utilisant le dictionnaire de hashtags [1]. Ce dictionnaire fournit des définitions pour les hashtags utilisés lors des discussions sur Twitter (cf. figure5.7). Lorsque cette information n'est pas disponible, des alternatives sont proposées pour rechercher le hashtag sur les moteurs de recherche.

  $\Longrightarrow$ Cette information correspond à la dimension de la représentation du document.

[1] https ://tagdef.com/fr/

- L'influence du tweet : un tweet influent (il a eu un nombre important de retweets et de favoris) est susceptible d'être un tweet important.

  $\Longrightarrow$ Cette information correspond à la sous-dimension de la qualité de l'information.

- l'URL : l'affichage de l'URL mentionnée dans le contenu textuel du tweet permet à l'utilisateur de compléter ses informations.

  $\Longrightarrow$ Cette information correspond à la dimension de la représentation du document.

- Horodatage de publication : l'affichage de l'horodatage de publication du tweet permet à l'utilisateur d'avoir une idée sur la fraîcheur de l'information par rapport au sujet de discussion.

  $\Longrightarrow$ Cette information correspond à la dimension de la fraîcheur de l'information.

FIGURE 5.7 – Exemple du dictionnaire des hashtags

5.3 Contextualisation des tweets : notre système proposé

Dans le cadre de notre travail, le contexte des tweets ambigus est présenté sous forme d'indicateurs et de graphes de conversations via un support visuel interactif. Nous proposons un système de visualisation qui est, à notre connaissance, le premier système, libre et extensible, pour la contextualisation des tweets en utilisant les conversations sur Twitter. Ce système implémente des algorithmes de construction et d'analyse des conversations, ainsi que des fonctionnalités avancées de collecte et de préparation des données sur Twitter. Notre système est doté d'une interface graphique facilitant l'accès à ses fonctionnalités (cf. figure 5.8). L'intérêt de la visualisation est de faciliter le traitement des informations complexes sur Twitter et l'essor aujourd'hui de la visualisation de l'information s'ancre dans une longue tradition de valorisation des données.

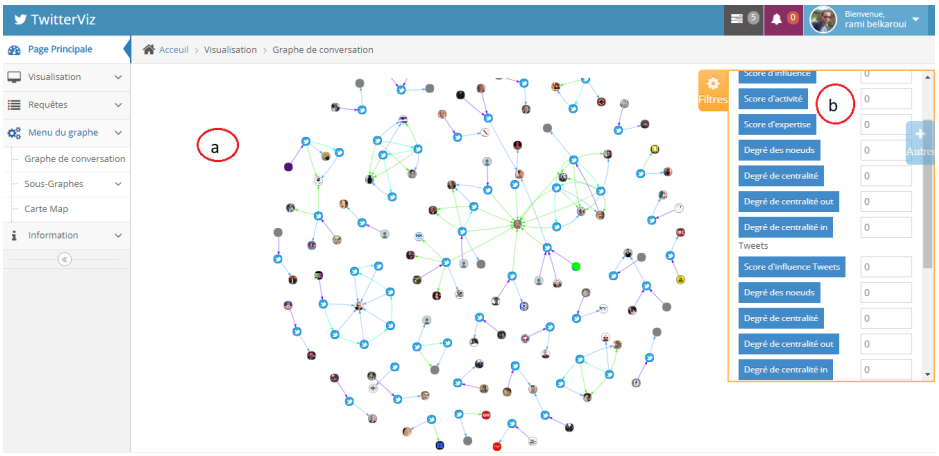

FIGURE 5.8 – Système de contextualisation et de visualisation des tweets proposé

5.3.1 Architecture générale du système

Notre système intègre trois services qui implémentent l'ensemble de la chaîne de traitement, de la collecte des tweets à l'analyse des graphes de conversations :

- Le service de manipulation (i.e. import, collecte et préparation) des données (cf. figure 5.9) ;
- Le service de construction et de visualisation des graphes de conversations (cf. figure 5.8.a) ;
- Le service d'analyse des conversations (cf. figure 5.8.b).

Le système traite deux types de données : les données décrivant l'ensemble de messages publiés par les utilisateurs, et les données décrivant la structure des graphes de conversations sur Twitter. La figure 5.10 décrit l'architecture logicielle du système, i.e. les entrées/sorties des différents services et la manière dont ils communiquent entre eux. Notons que le service de manipulation des données établit la liaison entre les données brutes d'entrée et les données manipulées par les services de construction et de visualisation des graphes. Nous précisons les fonctionnalités des différents services ci-dessous.

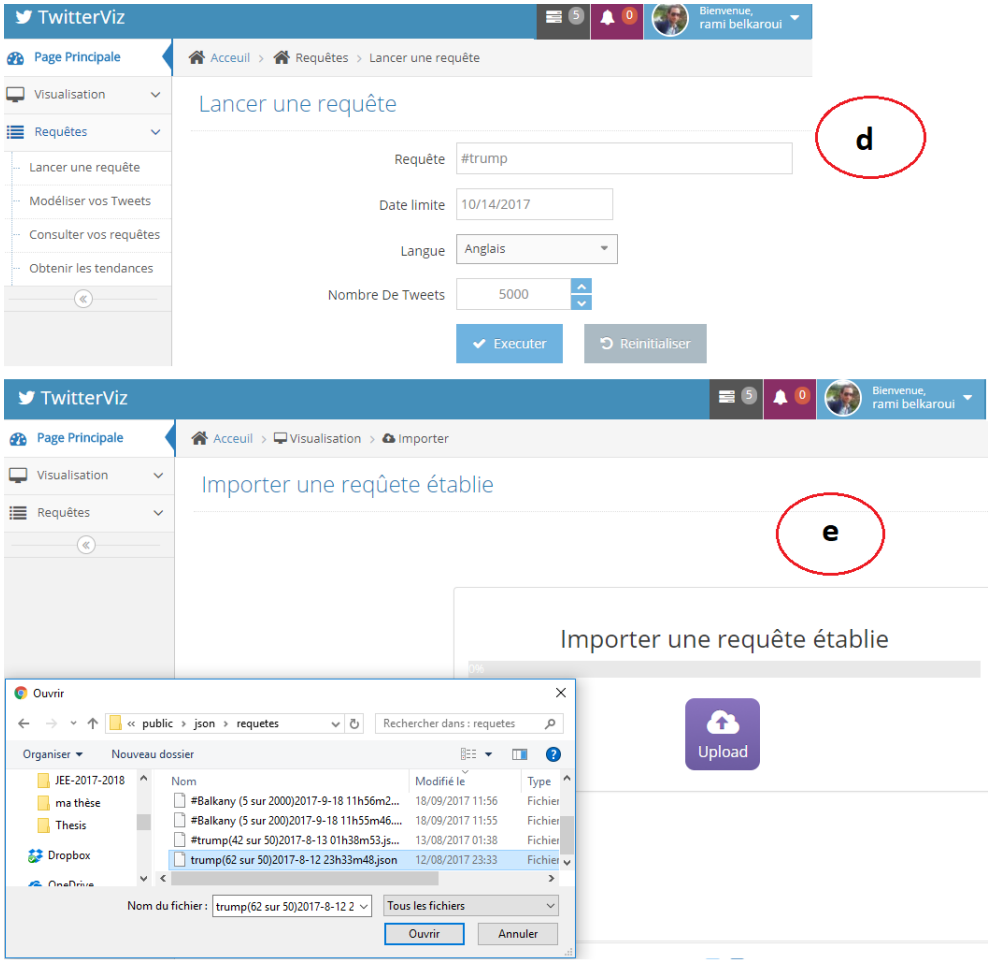

FIGURE 5.9 – Interface du service de manipulation des données : Collecte, import

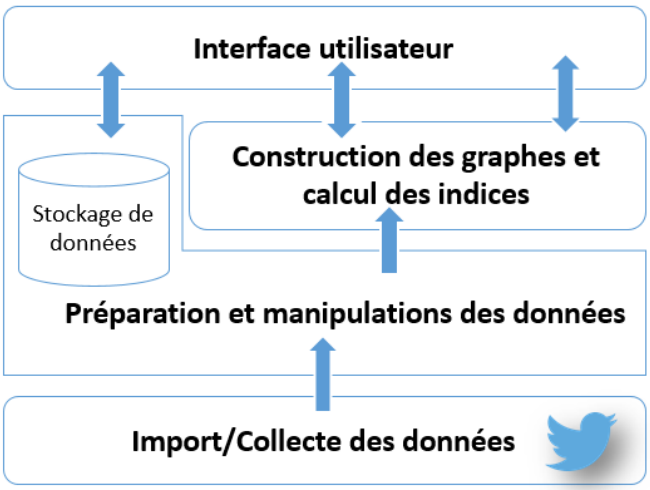

FIGURE 5.10 – Architecture du système proposé

5.3.1.1 Service de manipulation des données

Ce service gère un ensemble de messages publiés sur Twitter obtenus avec une requête de recherche. Il permet soit la collecte des données en temps réel, soit l'import d'un corpus de données déjà collecté. Il exige une authentification préalable via un compte Twitter.

Collecte des données. Elle est réalisée par la saisie d'une requête (cf. figure 5.9.d). Le système fait appel à l'API Twitter pour collecter les tweets qui correspondent à la requête. Lors de la collecte, il est possible de filtrer les données à collecter. Le filtrage est possible par le choix d'une période de temps (date de publication), la langue du tweets et le nombre de tweets souhaité.

Import des données. L'import des données est réalisé à partir d'un fichier JSON contenant les tweets déjà collectés (cf. figure 5.9.e). Lors de l'import, le système prend en charge les données et déclenche la construction des conversations et le calcul des indicateurs.

5.3.1.2 Le service de construction et de visualisation des graphes

Ce service permet d'enclencher l'algorithme de construction des conversations sur le corpus de données préparé et éventuellement filtré avec le service de manipulation des données, puis de visualiser les résultats (cf. figure 5.11).

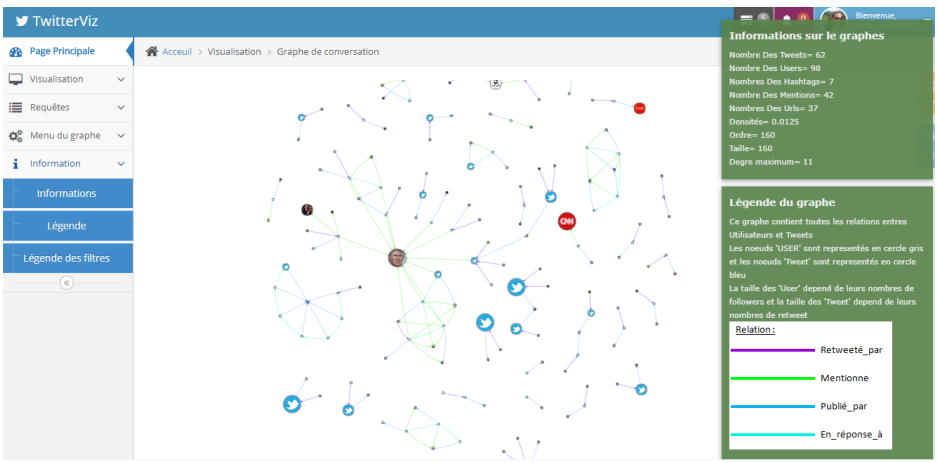

FIGURE 5.11 – L'interface principale de visualisation des graphes de conversations

5.3.1.3 Le service d'analyse des conversations

Ce service permet de calculer des indicateurs d'analyse des conversations (cf. figure 5.12) pour aider l'utilisateur à comprendre le contexte.

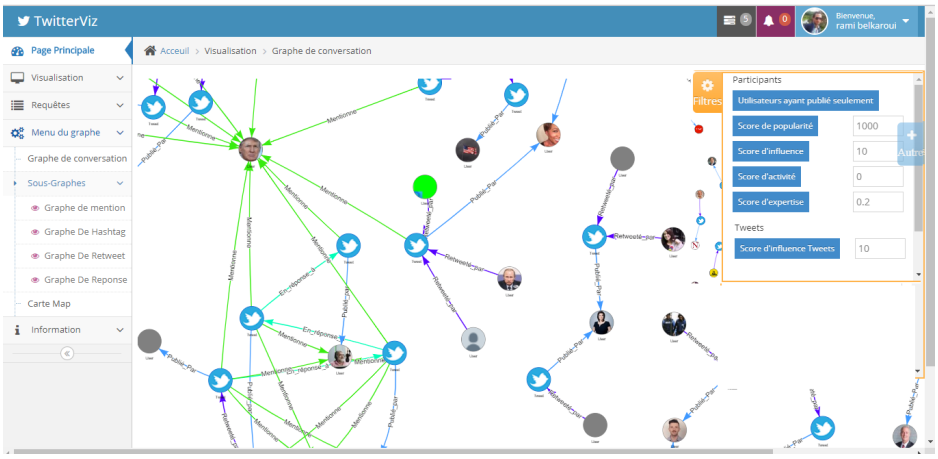

FIGURE 5.12 – L'interface principale du service d'analyse des conversations

5.3.2 Contextualisation des tweets : Exemples de scénarios d'utilisation

Dans cette partie, nous proposons deux scénarios d'utilisation illustrant les capacités de notre système. Le premier scénario décrit comment un non-expert peut utiliser le système pour contextualiser un tweet donné. Le deuxième scénario montre comment un chercheur peut profiter d'autres fonctionnalités avancées proposées par notre système. Mais préalablement nous tentons un état de l'art synthétique de la notion de "tweet ambigu" qui est le point d'entrée de notre système.

5.3.2.1 Tweet ambigu

De façon générale, la question de la définition de l'ambiguïté d'une expression est une question interdisciplinaire que l'on trouve abordée évidemment par les linguistes mais aussi en outre par les philosophes. Si ces domaines s'éloignent de nos connaissances en informatique, il semble néanmoins d'après nos lectures qu'il n'y ait pas de consensus. Citons à titre illustratif quelques définitions proposées :

- "An expression is ambiguous if the expression has more than one meaning" (Gillon, 1990).
- "A sentence is ambiguous if the sentence can be true in very different states of affairs" (Kempson, 1977).
- "L'ambiguïté représente une caractéristique des signes ou des structures syntaxiques de la langue (d'une langue particulière) qui menace potentiellement le succès des échanges communicatifs" (Grice, 1957).
- "L'ambiguïté est une propriété intrinsèque, inaliénable, de tout message centré sur lui-même, (...)" (JaKobson, 1960).
- L'ambiguïté est : "(a) un cas de non biunivocité entre formes et sens, (b) qui donne lieu à un choix nécessaire et impossible, et (c) qui constitue cas d'univocité dédoublée" (Fuchs, 1996).

La définition de Jakobson présente l'ambiguïté comme une principale propriété inhérente de l'émetteur : l'émetteur ne prend pas en charge ce que le récepteur va comprendre de son message. La définition de Fuchs peut être considérée comme la plus opérationnelle. Elle signifie que l'ambiguïté est présente "lorsqu'une expression possède plusieurs sens et qu'ils sont mutuellement exclusifs". Fuchs (1994), indique que la relation qui opère entre la forme et le sens d'une expression est indépendante des locuteurs et des situations car elle se situe au niveau du système de la langue. Une fois intégrée à un discours, le contexte rend la situation d'ambiguïté moins fréquente.

Pour contourner la difficulté de la question d'une définition universelle, différents types d'ambiguïté ont été identifiés. (Gillon, 1990) en a identifié trois : "L'ambiguïté se manifeste sous trois formes, l'ambiguïté lexicale polysémique, l'ambiguïté lexicale homonymique, et l'ambiguïté structurale non lexicale". (Nicolas, 2006) les a complétés en ajoutant l'ambiguïté sémantique (une forme est sémantiquement ambiguë si on peut lui faire correspondre au moins deux sens distincts), l'ambiguïté structurale (une forme manifeste une ambiguïté structurale si on peut lui faire correspondre au moins deux structures étiquetées distinctes) et l'ambiguïté contextuelle créée lorsqu'une forme est traitée en dehors de son contexte.

Cette dernière typologie pourrait permettre d'intégrer le cas des "tweets ambigus" qui n'ont à notre connaissance pas été définis précisément dans la littérature. On retrouve évidemment néanmoins dans les travaux des références à l'ambiguïté mais sans définition explicite. Citons à nouveau à titre illustratif quelques propositions :

- "De par sa taille, un Tweet est naturellement ambigu et souvent sous-spécifié, ce qui peut rendre la compréhension compliquée pour une personne ne possédant pas le contexte approprié" (Deveaud and Boudin, 2014).
- "considering the limited length of tweets, the bursty segments with higher user frequencies may not be semantically meaningful and are often ambiguous" (Li et al., 2012).
- "A tweet is an authentic use of Natural Language (...), this restriction increases the possible ambiguity of a tweet making it difficult" (Weerasooriya et al., 2016).

Ainsi, dans les travaux de la littérature, les auteurs considèrent généralement que les spécificités des tweets (taille, vocabulaire, syntaxe, etc.) rendent ces derniers ambigus. Notons d'ailleurs que nous n'avons pas retrouvé de définition précise de la notion de "tweet d'ambigu" dans les articles issus de la tâche de "Tweet contextualization" du challenge CLEF. Les auteurs semblent se baser sur l'hypothèse selon laquelle les spécificités du tweet concrétisent clairement l'ambiguïté.

De ce bref état des lieux nous pouvons confirmer l'importance de l'ambiguïté comme caractéristique de notre compréhension cognitive et de nos capacités d'interprétation. Et nous retenons comme définition celle proposée par Fuchs en la complétant par la typologie de Nicolas. Sans être vraiment suffisamment opérationnelles pour effectuer une catégorisation automatique sur les tweets, elles permettent de préciser le cadre intuitif qui a souvent suffit aux applications.

5.3.2.2 Utilisation par un non-expert

Dans la suite nous présentons deux scénarios d'utilisation qui se distinguent par le niveau d'expertise des utilisateurs. Mais en préalable il convient de préciser la notion "d'expert". Etymologiquement, le terme expert (du latin expertus, participe passé de experiri, faire l'essai de), adjectif au XIII° siècle, désignait la personne qui par l'expérience et par la pratique avait acquis une grande habileté. Selon le dictionnaire Larousse un expert est une personne "qui connaît très bien quelque chose par la pratique". Une définition similaire est proposée par le Webster "having, involving, or displaying special skill or knowledge derived from training or experience".

Dans la littérature scientifique on retrouve une variété de définitions qui dépendent souvent de plusieurs facteurs : domaine d'application, exigences et caractérisations techniques. Une des définitions que l'on retrouve fréquemment, notamment dans le domaine de l'aide à la décision où la question de l'expertise joue un rôle majeur, est celle proposée par (Shanteau, 1992) : "un expert est une personne reconnue par ses pairs comme celle qui a les connaissances et les compétences pour être au plus haut niveau". Nous conserverons donc cette définition pour nos scénarios d'usage.

Le premier s'adresse à un "non expert" qui pourrait être alors considéré ici comme un usager sans compétences spécifiques en informatique et qui joue le rôle d'exemple paradigmatique. Le second scenario s'adresse à un expert, qui lui est un chercheur dans le domaine de la recherche d'information. Notons que les journalistes en tant qu'utilisateurs et contributeurs majeurs de la "twittosphère" peuvent être concernés par les deux scénarios selon le niveau d'exploration de l'information souhaité.

La représentation proposée au non-expert est de manipulation simple. Elle offre l'accès à un ensemble d'informations sur le texte du tweet et sa source de publication (auteur) sous une forme visuelle interactive qui permet faciliter la compréhension du contexte d'un tweet.

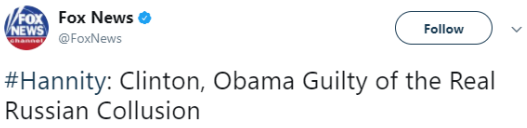

FIGURE 5.13 – Exemple d'un tweet ambigu à contextualiser

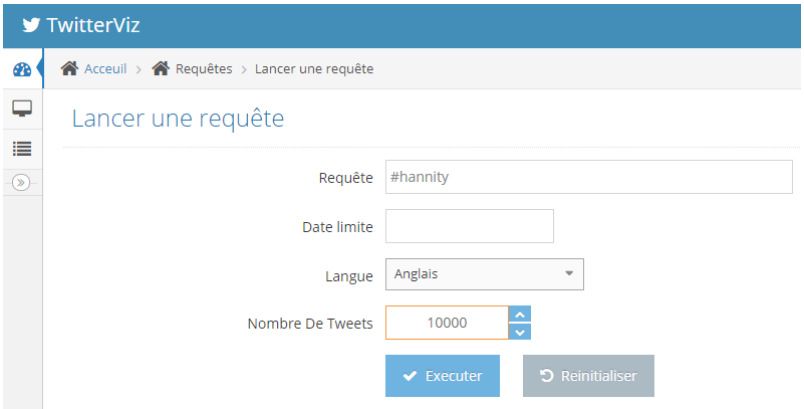

FIGURE 5.14 – Requête de recherche

Pour ce faire, nous utilisons le tweet, présenté par la figure 5.13, pour construire la requête de recherche (cf. figure 5.14) afin de collecter les tweets correspondants et construire les graphes de conversations.

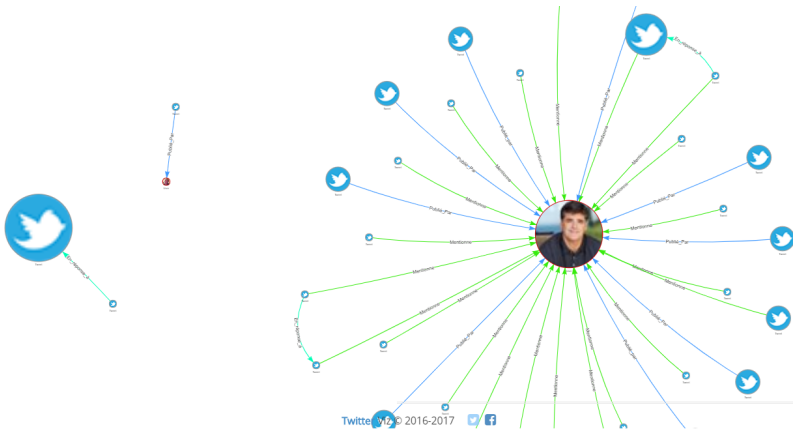

FIGURE 5.15 – Graphe fourni à l'utilisateur pour contextualiser le tweet

Ce tweet est contextualisé par l'ensemble des informations suivantes :
Pour les sommets participants, nous affichons :

- Le profil Twitter (y compris sa photo)
- La description du profil
- La popularité (nombre de followers)

La taille du sommet est en fonction du score de popularité. Pour les sommets tweets, nous affichons les informations suivantes :

- Le texte
- Les hashtags
- L'influence du tweet
- l'URL mentionnée dans le contenu textuel du tweet
- Le temps de publication

La taille du sommet tweet est en fonction de son score d'influence.

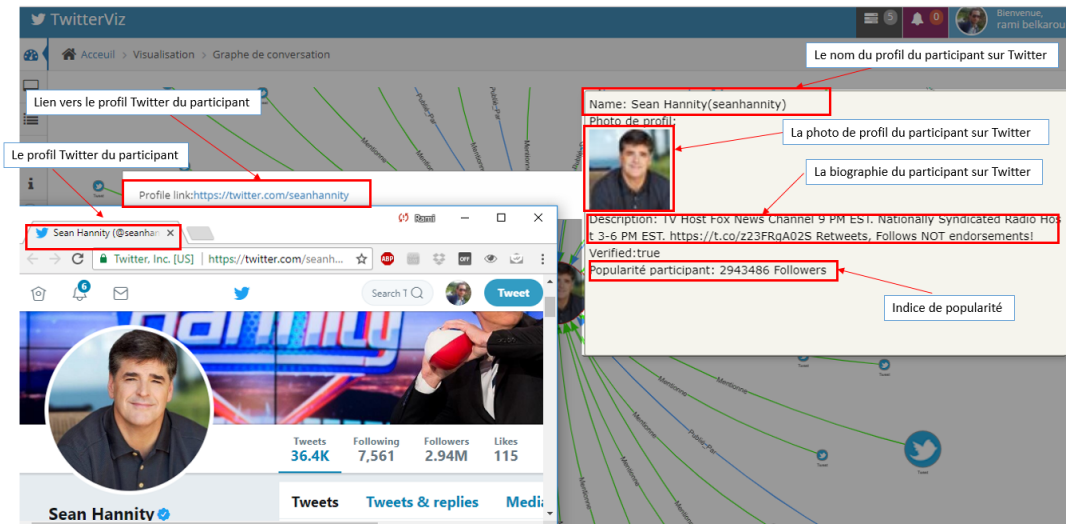

FIGURE 5.16 – Les éléments d'informations affichés sur les noeuds participants

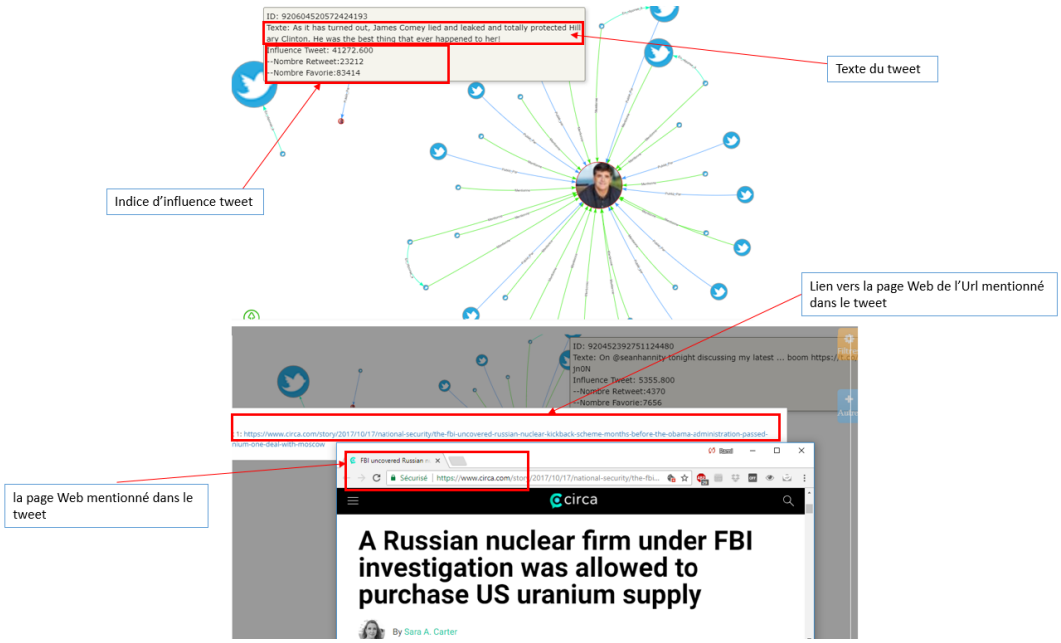

FIGURE 5.17 – Les éléments d'informations affichés sur les noeuds tweets

5.3.2.3 Utilisation par un expert

Le deuxième scénario décrit d'autres fonctionnalités avancées proposées par notre système à un utilisateur expert. Nous considérons un utilisateur expert ayant des connaissances et des expériences liées à la recherche d'informations et aux systèmes de contextualisation. Dans ce scénario, nous traitons le même tweet présenté dans le premier scénario.

- Consultation des sujets tendance

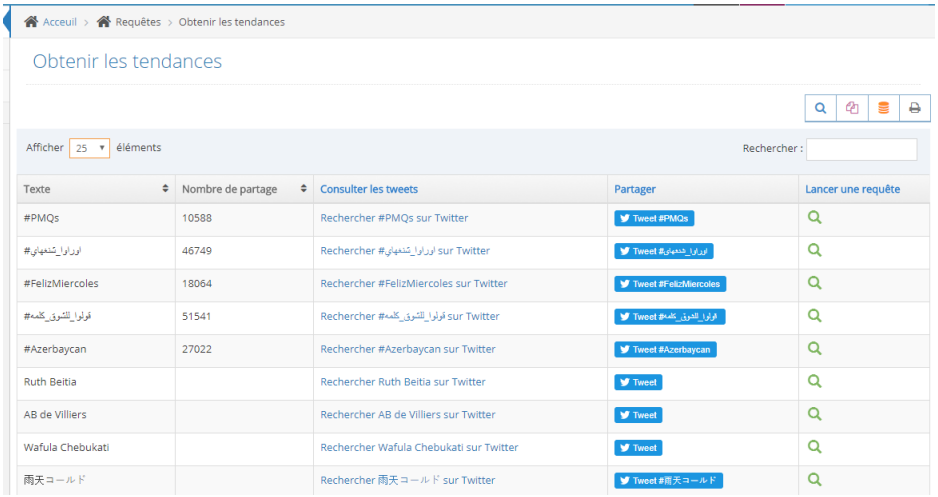

FIGURE 5.18 – Fonctionnalités de consultation des sujets tendance

Lors du lancement de la requête, cette fonctionnalité permet de consulter les sujets tendance et d'identifier les meilleures requêtes (cf. figure 5.18).

- Consultation des listes des tweets et des participants

  Les listes des tweets et des participants peuvent être téléchargeables (cf. figure 5.19). De plus, le graphe peut être exporté sur l'extension du logiciel de visualisation Gephi [2] qui

[2] https ://gephi.org/

est l'un des plus populaires en fouille visuelle de données.

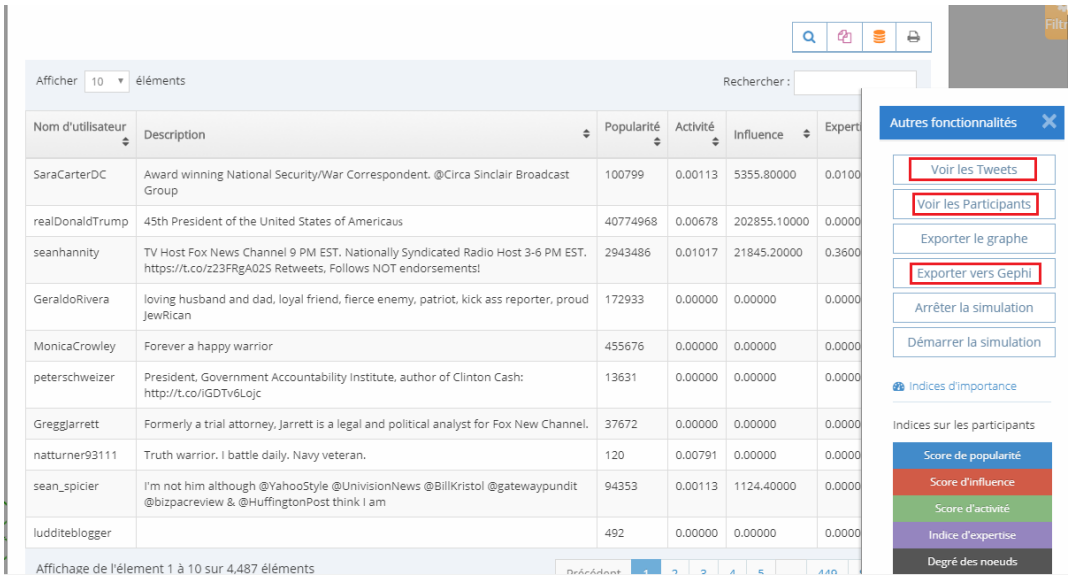

FIGURE 5.19 – Fonctionnalités de consultation des listes de tweets/participants et l'export du graphe vers le logiciel de visualisation Gephi

- Fonctionnalité du graphe de hashtags

  Le graphe de hashtags (cf. figure 5.20) représente l'ensemble des hashtags présents dans les tweets de la conversation liés au hashtag principal utilisé dans la requête. Pour chaque hashtag, nous présentons sa définition en utilisant un moteur de définition des hashtags. Ce graphe permet à l'utilisateur de découvrir un ensemble de sujets liés au sujet de sa requête.

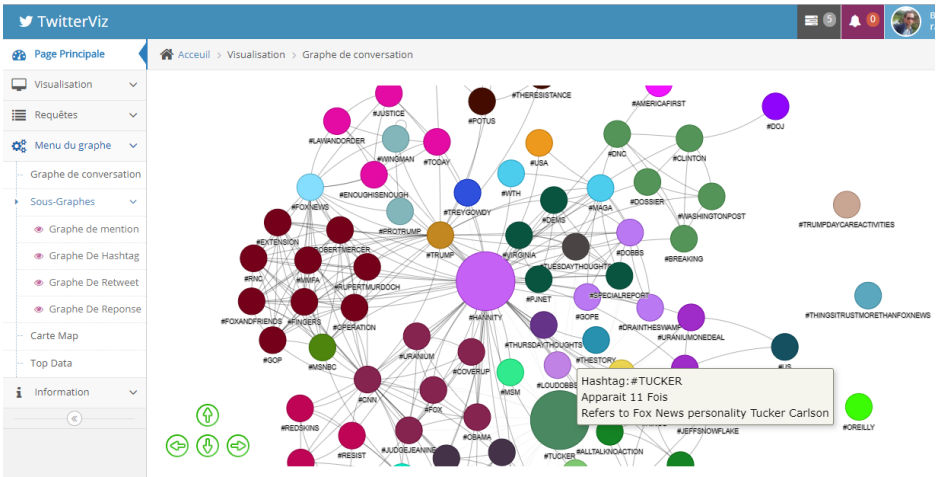

FIGURE 5.20 – Un graphe de hashtags

- Fonctionnalité du graphe de mentions

  Le graphe de mentions (cf. figure 5.21) permet de visualiser les profils des utilisateurs impliqués dans la conversation par la relation mention. Par exemple, si les mentions restituées sont relatives à un ensemble de politiciens, il est probable que le sujet de la conversation porte sur un thème politique.

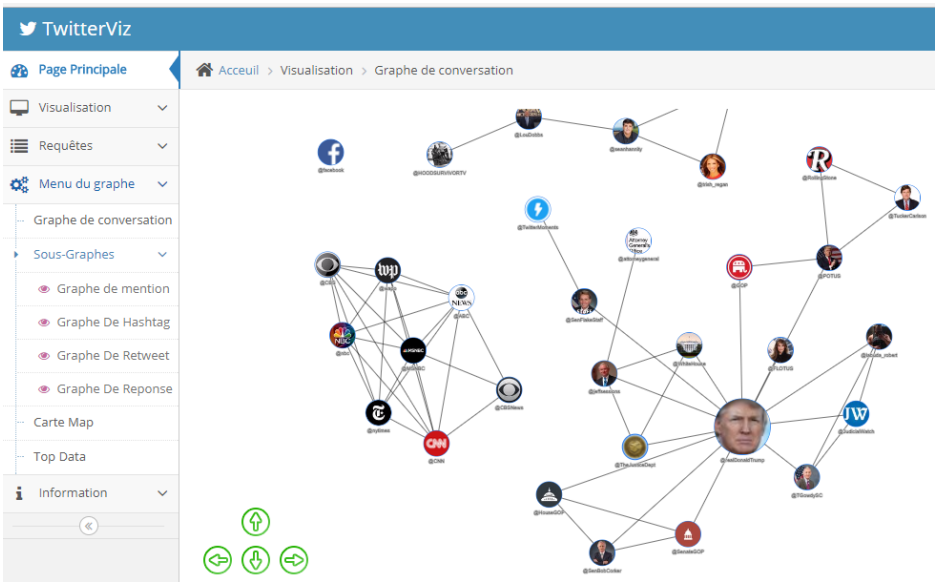

FIGURE 5.21 – Un graphe de mention

- Fonctionnalité des N-meilleures données et des métadonnées

  Cette fonctionnalité fournit les participants influents, les URLs les plus utilisées, les hashtags les plus utilisés, les participants qui ont eu plus de réponses, les participants les plus mentionnés (cf. figure 5.22).

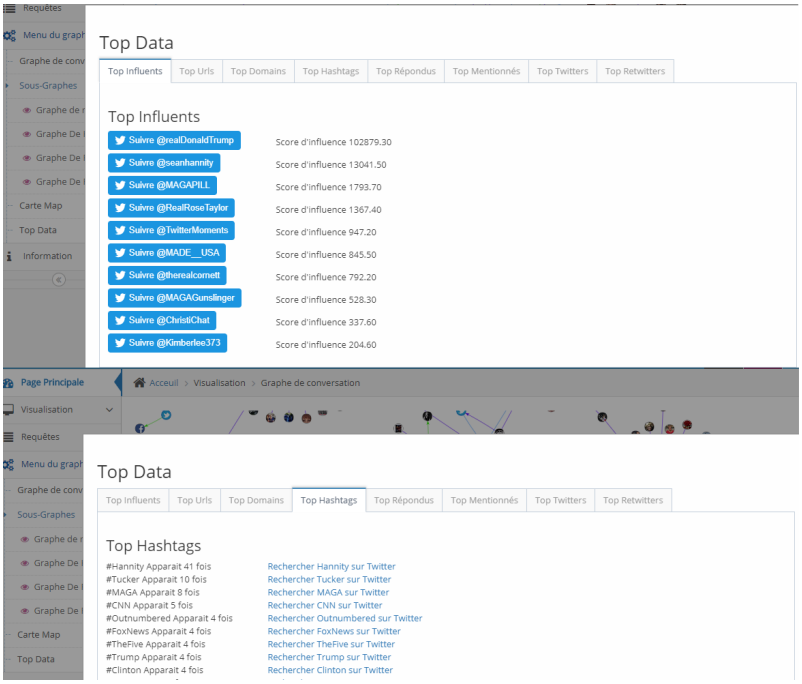

FIGURE 5.22 – Fonctionnalité des N-meilleures données et des métadonnées

5.4 Conclusion

Dans ce chapitre, nous avons proposé un couplage original entre les conversations sur Twitter et la contextualisation des tweets. Ce couplage a été opérationnalisé via une interface graphique adaptée. Des scenarios d'usage ont permis à la fois d'illustrer les différentes fonctionnalités de notre système et de montrer la valeur ajoutée de cette nouvelle contextualisation. Dans le

chapitre suivant, nous dépassons ces scenarios ad hoc pour proposer une évaluation sur un petit panel d'usagers réels.

6

Expérimentations

Sommaire

6.1 Introduction 79

6.2 Contextualisation des tweets : État de l'art sur les données et les résultats 80

6.2.1 Les données INEX/CLEF 80

6.2.2 Évaluation de la contextualisation des tweets 82

6.2.3 Discussion sur les résultats de la campagne INEX/CLEF "Tweet Contextualization" 84

6.3 Évaluations de notre approche : protocole expérimental 85

6.3.1 Étude utilisateurs 85

6.3.2 Les entretiens 87

6.4 Résultats 87

6.4.1 Résultats issus du questionnaire 88

6.4.2 Résultats extraits des entretiens 90

6.5 Conclusion 91

6.1 Introduction

Pour les systèmes centrés utilisateur où la part de l'exploration de données joue un rôle central, la question de la validation reste largement débattue dans la communauté. Les tâches ne peuvent se réduire à des tâches très précises (ex : pointer une cible) comme il est courant de le faire dans les expérimentations "classiques" menées dans le domaine de l'Interface Homme-Machine. En explorant un espace informationnel, l'utilisateur découvre des connaissances dont il n'avait pas nécessairement pu verbaliser les caractéristiques recherchées a priori. La mesure de cette valeur ajoutée est donc délicate à définir. Depuis plus d'une décennie, de nombreux articles de préconisation ont été écrits (Plaisant, 2004; Heer and Agrawala, 2006; Munzner, 2008; Carpendale, 2008; Bostock et al., 2011; Lam et al., 2012) et tout en soulignant les écueils à éviter, tous confirment la difficulté de la tâche. Cependant, en dépit de ces difficultés, la restitution visuelle des résultats issus de processus de recherche d'information et de fouille de données n'a

cessé de se développer et a montré son intérêt dans de nombreuses applications qui sont présentées notamment dans des conférences dédiées (e.g. IEEE Information Visualization, IEEE Visual Analytics Science and Technology, International Conference on Information Visualisation). Son essor est associé à celui du domaine de la "Visual Analytics" défini par Thomas et Cook dans leur papier séminal en 2005 comme "la science du raisonnement analytique facilité par les interfaces visuelles interactives" (Thomas, 2005).

L'analyse des réseaux sociaux a été un des domaines majeurs de développement de cette approche. Pour Twitter en particulier, la plupart des approches mettent l'accent sur le contenu et plus particulièrement sur l'analyse des flux de messages courts (Šilić and Bašić, 2010). Ces approches peuvent se regrouper en trois grandes familles qui s'appuient sur des techniques différentes : les nuages de mots (Morstatter et al., 2013; Lee et al., 2013), le clustering (Gansner et al., 2013), et les séries temporelles (Lu and Yang, 2012; Dou et al., 2012). Pour l'analyse structurelle, les travaux s'appuient sur la littérature très abondante de la visualisation de graphes (Brandes et al., 2013).

Dans cette thèse, l'aspect novateur n'a pas été centré sur la visualisation et pour la restitution des composantes des graphes de conversations qui constituent le contexte nous nous sommes appuyés sur des outils existants, en particulier sur l'outil d'analyse "R[1]", l'outil de gestion et de visualisation des graphes "Neo4j[2]", le framework "angularJS[3]" et d'autres bibliothèques de développement (Johnpaul and Mathew, 2017; Sweet, 2017; McCarthy et al., 2017). L'évaluation porte donc sur l'apport de notre nouvelle contextualisation pour la compréhension des tweets. Avant de définir notre cadre expérimental basé sur un questionnaire complété par des entretiens semi-directifs, il nous a semblé utile de nous référer aux démarches d'évaluation mises en place lors du challenge INEX "Tweet contextualization" qui s'étend sur les six dernières années. Bien que la définition du contexte soit très différente de celle que nous avons adoptée, ces travaux nous permettent d'identifier les facteurs importants à évaluer et certains écueils à éviter.

Le reste du chapitre est donc composé de trois parties principales. La première partie est une synthèse des protocoles d'évaluation développés dans le challenge INEX. La deuxième partie présente notre protocole d'évaluation et la troisième présente les résultats obtenus qui semblent confirmer un apport de notre proposition pour la compréhension des tweets ambigus.

6.2 Contextualisation des tweets : État de l'art sur les données et les résultats

Dans cette partie, nous décrivons les données utilisées pour évaluer la contextualisation des tweets telles qu'elles ont été définies dans les campagnes d'évaluation INEX "Tweet Contextualization" 2011, 2012, 2013, 2014, 2016 et 2017. Ensuite, nous présentons les deux mesures d'évaluation adoptées dans le cadre de la tâche "Tweet Contextualization" ainsi que les résultats obtenus lors des campagnes d'évaluation INEX/CLEF.

6.2.1 Les données INEX/CLEF

En 2011, la contextualisation des tweets a été évaluée dans le cadre de la tâche de "Question Answering" qui a été renommée en 2012 "Tweet Contextualization". En 2011, l'ensemble des données en entrée (requêtes) comprenait 132 tweets, chaque tweet étant constitué de l'id, du titre et du texte. Le texte de chaque tweet était composé par le titre d'un article du "New York Times" publié en juillet 2011, avec l'URL de cet article (cf. figure 6.1). Pour chaque tweet, les

[1]https ://www.r-project.org/
[2]https ://neo4j.com/
[3]https ://angularjs.org/

participants devaient fournir un résumé de 500 mots qui permettait de répondre à la question : de quoi parle ce tweet ? Le résumé devait être composé uniquement d'extraits de Wikipédia.

```
<topic id="2011003">
        <title>Obama to Back Repeal of Law Restricting Marriage</title>
        <txt> WASHINGTON - President Obama will endorse a bill to repeal
  the law that limits the legal definition of marriage to a union between a man
   and a woman, the White House said Tuesday, taking another step in support of
   gay rights.
        </txt>
</topic>
```

FIGURE 6.1 – Exemple des données tweets fournies par INEX en 2011

À partir de 2012, la tâche est devenue plus réaliste et les sujets évalués ont été recensés chaque année par des tweets collectés manuellement par les organisateurs. Ces tweets ont été sélectionnés et vérifiés afin de s'assurer que :

- Ils contiennent du "contenu informatif" (en particulier, pas de messages purement personnels) ; seuls les comptes non personnels ont été pris en compte.
- La collection des documents de Wikipédia comporte du contenu associé, de sorte qu'une contextualisation soit possible.

En 2012, le corpus textuel a été présenté par une mise à jour de la dernière version de Wikipédia. Ainsi, l'ensemble des requêtes (données en entrée) a été radicalement modifié. Il s'agit d'environ 1000 tweets réels écrits en anglais, collectés à partir de comptes informatifs tels que @CNN, @TennisTweets, @PeopleMag, @science, etc. Les tweets sont disponibles dans un format JSON, comme illustré dans la figure 6.2. Cependant, la tâche reste la même : fournir un résumé de 500 mots.


```
{"created_at":"Fri, 03 Feb 2012 09:10:20 +0000",
"from_user":"XXX",
"from_user_id":XXX,
"from_user_id_str":"XXX",
"from_user_name":"XXX",
"geo":null,
"id":XXX,
"id_str":"XXX",
"iso_language_code":"en",
"metadata":"result_type":"recent",
"profile_image_url":"http://XXX",
"profile_image_url_https":"https://XXX",
"source":"<a href=''http://XXX",
"text":"blahblahblah",
"to_user":null,
"to_user_id":null,
"to_user_id_str":null,
"to_user_name":null}
```


FIGURE 6.2 – Exemple des données tweets fournies par INEX 2012

En 2013, les tweets étaient sur des sujets plus divers que ceux de 2012, afin que les participants puissent mieux mesurer l'impact du traitement par hashtag sur leurs approches par

exemple. En particulier, il y avait 598 tweets en anglais à contextualiser à partir d'un "dump" de Wikipédia.

En 2014, les organisateurs de CLEF RepLab 2013 ont collecté 240 tweets en anglais. En 2014, les participants devraient fournir un contexte aux tweets du point de vue des entités connexes (related). Les tweets avaient au moins 80 caractères et ne contenaient pas d'URL. Un tweet a les types d'annotation suivants : la catégorie, un nom d'entité et une étiquette de sujet faite manuellement (voir un exemple dans la figure 6.3). Le nom de l'entité devait être utilisé comme un point d'entrée dans Wikipédia ou DBpedia. Le contexte des résumés générés devait être pleinement lié à cette entité. Comme les années précédentes, il devait être un résumé extrait d'un "dump" de Wikipédia.

ID	Catégorie	Entité	Sujet	Contenu
213051315880869888	automotive	Fiat	sales	Seeing a lot of #Fiat cars downtown these days. #Traffic

FIGURE 6.3 – Exemple de format de tweet INEX 2014

À partir de l'année 2016, la tâche à été renommée "CLEF Cultural Microblog Contextualization" pour devenir plus spécifique et focalisée sur la contextualisation des tweets culturels liés aux événements. Ce choix est basé sur le fait que les événements génèrent un grand nombre de tweets. Cependant, la tâche reste la même : fournir un résumé de 500 mots, composé de passages du corpus Wikipédia. Ainsi, les données tweets gardent le même format fourni en 2014. Ces tweets ont plus de 80 caractères, ils ne contiennent pas d'URL et sont écrits dans plus de 20 langues différentes. Exactement les mêmes données sont fournies en 2017 ; un ensemble de 1.100 tweets en 20 langues différentes.

6.2.2 Évaluation de la contextualisation des tweets

Dans cette partie, nous rappelons la définition des mesures proposées dans le cadre de INEX/CLEF pour évaluer la contextualisation des tweets. La contextualisation est évaluée par deux dimensions : le contenu informatif (informativité) et la lisibilité des résumés produits. La mesure de l'informativité est introduite dans (SanJuan et al., 2012) afin de mesurer la similarité du contenu textuel d'un résumé avec un résumé de référence créé par des humains. Cependant, la lisibilité vise à mesurer à quel point le résumé est clair et facile à comprendre. Ces deux dimensions ne sont pas indépendantes et ne peuvent pas être entièrement automatisées car l'évaluation de la lisibilité nécessite une intervention humaine. Dans ce qui suit, nous présentons les détails de chaque mesure.

6.2.2.1 Mesure de l'informativité

Dans une tâche de contextualisation, l'informativité ne peut pas être évaluée à l'aide des mesures RI standards puisqu'il ne s'agit pas ici de trouver tous les passages pertinents mais de compléter une information (le tweet) existante pour améliorer sa compréhension. Selon (Louis and Nenkova, 2009; Saggion et al., 2010) la mesure d'informativité est basée sur le chevauchement lexical entre un ensemble de passages pertinents (référence textuelle) et les résumés générés automatiquement par les participants. La référence textuelle est construite comme suit. Les passages des participants sont filtrés par les organisateurs qui ont supprimé manuellement les passages non informatifs, et les phrases identiques ont été fusionnées ensemble. Plus précisément, "l'informativité" est définie par la formule suivante :

$$Dis(T,S) = \sum_{t \in T} (P-1) \times \left(1 - \frac{min(log(P), log(Q))}{max(log(P), log(Q))}\right) \tag{6.1}$$

avec $P = \frac{f_T(t)}{f_T} + 1$ and $Q = \frac{f_S(t)}{f_S} + 1$.
$\frac{f_T(t)}{f_T} \in (0,1]$ et $\frac{f_S(t)}{f_S} \in (0,1]$, P>1, Q>1 et $max(log(P), log(Q)) > 1$.

où : T est l'ensemble des termes dans le groupe des passages pertinents, $f_T(t)$ est la fréquence d'un terme t dans le groupe, f_T est le nombre total de termes dans le groupe, $f_S(t)$ est la fréquence d'un terme t dans un résumé, f_S est le nombre total de termes dans un résumé. Le logarithme permet de traiter des mots très fréquents. Les valeurs les plus faibles de $Dis(S;T)$ correspondent à la plus grande informativité.

6.2.2.2 Mesure de lisibilité

La deuxième mesure proposée par INEX/CLEF est la lisibilité des résumés. Les mêmes sujets/tweets sont utilisés pour évaluer la lisibilité des résumés (SanJuan et al., 2012). L'évaluation de la lisibilité est effectuée manuellement. Pour chaque passage, dans chaque résumé, les évaluateurs devaient indiquer si le passage contenait l'un des inconvénients suivants :

- Syntaxe : le passage inclut des problèmes syntaxiques (par exemple, une mauvaise segmentation).
- Anaphore : le passage contient des références non résolues (impossibilité pour le lecteur de rattacher un pronom anaphorique à son antécédent).
- Redondance : le passage contient des informations redondantes (c'est-à-dire des informations qui sont déjà mentionnées).
- À écarter : le passage est incohérent et incompréhensible (n'a aucun sens dans le contexte donné).

Pour l'évaluation de la lisibilité, trois métriques sont utilisées :

- Métrique tolérante (fondée sur la pertinence) : un passage est considéré comme valide si la case "À écarter" n'a pas été cochée.
- Métrique intermédiaire : un passage est considéré comme valide si les cases "À écarter" et "Syntaxe" n'ont pas été cochées.
- Métrique stricte : un passage est considéré comme valide si aucune case n'a été cochée.

En 2011, la lisibilité des résumés est estimée comme le nombre de mots (jusqu'à 500) dans des passages valides (SanJuan et al., 2012). Depuis 2012, le score d'un résumé est le nombre moyen normalisé de mots dans des passages valides (Bellot et al., 2013). L'ordre des phrases n'a pas été jugé par les organisateurs, mais il est très important pour la compréhension du texte (Barzilay et al., 2002).

6.2.3 Discussion sur les résultats de la campagne INEX/CLEF "Tweet Contextualization"

6.2.3.1 Résultats de l'informativité et la lisibilité à INEX/CLEF 2011

Lors de la campagne 2011, 11 équipes avaient participé et soumis 23 runs. Et, 37 303 passages avaient été soumis. Le meilleur système (IRIT) est basé sur un modèle vectoriel qui utilise une pondération spéciale pour les entités nommées lors du calcul de la similarité cosinus entre les phrases et les tweets. Le deuxième système (Montréal) a combiné le modèle de langue avec Cortex (Torres-Moreno and Ramırez, 2010), un système de résumé statistique avancé. Le troisième système (UAM) a combiné le même modèle de langue avec une approche d'optimisation combinatoire pour sélectionner un groupe de phrases à inclure dans le résumé en minimisant une divergence de "Jensen-Shannon" [4] avec les documents récupérés.

Vis-à-vis de la lisibilité, le meilleur système de 2011 (Ermakova and Mothe, 2012b) a mis en oeuvre de la détection d'anaphore, ainsi que la prise en compte de la densité d'information des résumés. Le système classé deuxième, a utilisé un algorithme de vote pour combiner plusieurs approches de construction de résumés (Torres-Moreno and Velázquez-Morales, 2012). Et, les méthodes classiques de résumé automatique basées sur la sélection des phrases ont également montré de bons résultats (Crouch et al., 2012; Deveaud and Boudin, 2012; Ganguly et al., 2012).

6.2.3.2 Résultats de l'informativité et la lisibilité à INEX/CLEF 2012

Après 2011, les tweets sélectionnés en tant que sujets reflètent une grande diversité, ce qui rend la tâche plus complexe. La campagne 2012 est composée de 13 équipes de 10 pays différents et 33 runs ont été soumis. Dans cette année, il y a eu 671 191 passages. Un seul système (Duluth) a surpassé la référence fournie par les organisateurs. Duluth a utilisé un modèle de langue et une technique de résumé basée sur les scores des phrases, en intégrant également des règles simples pour gérer les contenus spécifiques aux tweets. Deux autres systèmes ont eu des scores proches de la référence. Le système de l'équipe de Nantes (classé deuxième) a approfondi la phase de pré-traitement du tweet, combiné avec un modèle de langue spécifique. Le système d'Avignon (classé troisième) a essayé une approche complètement différente basée sur le modèle LDA sur Wikipédia pour extraire des phrases.

6.2.3.3 Résultats de l'informativité et la lisibilité à INEX/CLEF 2013

Les tweets proposés en 2013 étaient aussi nombreux qu'en 2012 mais les organisateurs ont supprimé manuellement les spams. La plupart des tweets sélectionnés nécessitaient cependant un pré-traitement spécifique, comme l'analyse des hashtags.

En 2013, un total de 13 équipes de 9 pays ont soumis 24 runs. Cette année, le système ayant obtenu le meilleur score, en matière d'informativité, a mis en oeuvre un traitement spécifique des hashtags (Deveaud and Boudin, 2012). La meilleure performance de ce système est obtenue en exploitant toutes les fonctionnalités disponibles du tweet, y compris les URLs qui n'étaient pas autorisés par les organisateurs. Cependant, sa deuxième meilleure performance est obtenue sans l'utilisation des URLs. Le deuxième meilleur système participant (IRIT) en informativité et en lisibilité a utilisé des outils de traitement automatique du langage. Ce système a été le premier en informativité en 2011 (Ermakova and Mothe, 2012a). Le troisième meilleur système a été le premier en 2012 (Ganguly et al., 2012), ce qui a confirmé la qualité de ses performances qui résistent à la diversification.

Les meilleurs systèmes intègrent étroitement le pré-traitement de tweet, la recherche de document en utilisant un modèle de langue et une méthode de résumé basée sur les scores des

[4] https ://en.wikipedia.org/wiki/Jensen-Shannon-divergence

phrases. Comme en 2012, presque tous les participants ont utilisé des modèles de langue. L'extraction terminologique et la re-formulation appliquées aux tweets ont également été utilisées en 2013 comme dans les éditions précédentes (Vivaldi and da Cunha, 2013). L'extraction appropriée et l'analyse des tweets et des pages Wikipédia semblent aussi être un point important. Par ailleurs, tous les systèmes qui ont été classés parmi les cinq premiers ont adopté un étiqueteur morpho-syntaxique tel que "TreeTagger" ou "Stanford Core NLP".

Vis-à-vis de la lisibilité, le meilleur système sur les trois années (Ermakova and Mothe, 2012a) a mis en oeuvre une évaluation automatique de lisibilité et de détection d'anaphore, ainsi que la prise en compte de la densité d'information des résumés.

6.2.3.4 Résultats de l'informativité et la lisibilité à INEX/CLEF 2014

En 2014, 4 équipes combinées de six pays ont soumis 12 runs. Le nombre total de passages soumis était de 54 932.Le meilleur système participatif en informativité a utilisé un modèle basé sur les règles (Zingla et al., 2014). La contextualisation étant limitée à un aspect décrit par une entité, les règles d'association peuvent être pertinentes pour extraire des informations qui se focalisent sur cet aspect.

Le meilleur système participatif en lisibilité a utilisé un système de résumé avancé qui introduisait des changements mineurs dans les passages pour améliorer la lisibilité. Changer le contenu des passages n'a pas été autorisé, mais cela tend à montrer que pour traiter la lisibilité, une ré-écriture est nécessaire.

6.2.3.5 Résultats de l'informativité et la lisibilité à INEX/CLEF 2016 et 2017

Dans la campagne de 2016 et 2017, les mesures d'évaluation restent toujours inchangeables : informativité et lisibilité. Cependant, les résultats officiels n'ont pas encore été publiés par les organisateurs (Ermakova et al., 2017; Ermakova et al., 2016). En effet, un constat fait par les organisateurs est que les tweets liés à un événement forment un corpus dense, riche mais très bruyant. Le contenu est souvent imprécis, dupliqué ou non informatif. Ainsi, le traitement d'un corpus multiculturel et multilingue des tweets révèle les limites des approches statistiques et linguistiques. Cela nécessite également des ressources linguistiques pour chaque langue ou pour des événements culturels spécifiques.

6.3 Évaluations de notre approche : protocole expérimental

À notre connaissance, il n'existe pas encore un cadre d'évaluation standardisé de la tâche de contextualisation du tweet autre que ceux définis dans le cadre de INEX/CLEF. Cependant, notre approche ne peut pas rentrer dans ce cadre puisque nous n'utilisons pas ici de résumé textuel. Les mesures "d'informativité" et de "lisibilité" ne peuvent donc pas être mobilisées avec leurs définitions originales. Cependant, les notions qu'elles tentent de mesurer restent évidemment suffisamment générales pour être valides dans le cadre de notre proposition. Notre évaluation combine deux approches : un questionnaire associé à une tâche de contextualisation qui tente d'évaluer l'apport des éléments d'information extraits du graphe de conversation et un entretien semi-directif qui permet de préciser le sentiment général d'appropriation de l'outil et de son apport et de suggérer des voies d'amélioration.

6.3.1 Étude utilisateurs

La tâche d'évaluation a été menée auprès de 30 participants (étudiants, doctorants et membres du laboratoire) confrontés à une analyse de trois sujets différents. Pour limiter les biais éventuels

associés à un sujet donné et élargir les conclusions, chaque sujet a été traité par 10 participants. La figure 6.4 représente un exemple d'un tweet ambigu choisi pour être contextualisé par notre approche. Ce tweet est présenté aux évaluateurs pour une simple lecture et l'évaluation de son contexte approprié suit quatre étapes.

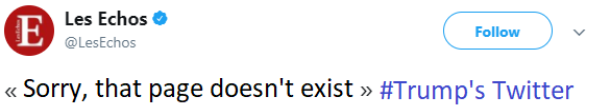

FIGURE 6.4 – Exemple d'un tweet ambigu adopté dans notre évaluation

- **Évaluer l'apport des textes de tweets pour la compréhension du tweet ambigu**

  Nous présentons aux évaluateurs une liste contenant un ensemble de tweets portant sur le même sujet que le tweet ambigu (cf.figure 6.5). Ces derniers sont extraits du graphe de conversation. La question posée est la suivante.

  Question 1 : Après la lecture de l'ensemble des tweets, est-ce que vous avez pu comprendre le contexte du tweet ? Les réponses possibles sont : 0 :toujours ambigu ; 1 :minimalement clair ; 2 :clair ; 3 :très clair et compréhensible. L'objectif est de cerner l'apport de tweets additionnels provenant des conversations dans la compréhension du tweet initial.

ID	Texte	Nom d'utilisateur	Nombre des Retweet	Nombre des Favoris
926223332147687425	Trump's twitter account vanishing and then reappearing moments later is the best evidence yet that the events of th... https://t.co/t3u1SQAecU	Nfinit	5722	18973
922278611977166848	.@realDonaldTrump: Look! It's five former U.S. presidents and #PutinsPuppet. #sTRUMPet #Trump https://t.co/kEmLmnuZdy	BrianPaulStuart	19	36
922961025682542593	#TheFive This GoldStar mother who appreciates President Trump's phone call and was touched by it!! We are with you,... https://t.co/JH0L8s7Tfv	immigrant_legal	641	994
926227896783515648	Donald Trump's Twitter account was deleted, but like herpes, it came back again. https://t.co/Lktnk7Nwbm https://t.co/a0hjc42qOd	TheRoot	1102	2209
926237007952695296	Twitter reacts to disappearance of Trump's Twitter account https://t.co/jkmrzlXUbE https://t.co/glzNLqJGEb	thehill	125	350
926115254467043328	#HouseGOP #TaxPlan #TaxReform #taxes #GOPTaxPlan Before any tax plan is passed we need to see Trump's tax returns... https://t.co/Tp2xMAhtYc	TheTweetwit	301	428
926237481590239237	If Haldeman & Erhlichman were as stupid as Trump's guys, Nixon would have been POTUS for about 2 fewer years https://t.co/wIKUKlevFT	jsaquella	0	1
926233732012441601	Based on Trump EPA's own data, the #CleanPowerPlan could save 4,500 lives every year. They're still repealing it.... https://t.co/MZVa2H3Jd0	350	37	30

FIGURE 6.5 – Ensemble des textes de tweets

- **Évaluer l'apport du graphe de conversation dans la désambiguisation**

  Les extraits du graphe de conversation filtrés selon les indicateurs présentés dans le chapitre précédent sont présentés à chaque participant. Il leur est demandé de prendre leur temps pour naviguer dans le graphe et accéder aux éléments complémentaires de contextualisation. Les différentes possibilités offertes par le système leur sont indiquées : possibilité de consulter le profil de l'auteur du tweet, de consulter la définition d'un hashtag utilisé dans le graphe, etc. Après cette étape d'exploration, deux questions sont posées avec une réponse donnée comme ci-dessus sur une échelle de 0 à 3.

  Question 2 : Est-ce que le contexte du tweet vous semble bien clair ? Les réponses à la question sont données sur une échelle de 0 à 3 allant du "toujours ambigu" à "très clair".

Question 3 : Est-ce que vous trouvez que la visualisation des graphes de conversations est importante et apporte un plus par rapport à la lecture des textes de tweets ? Les réponses à la question sont données sur une échelle de 0 à 3 allant du "pas intéressant" à "très intéressant".

- **Évaluer l'apport des éléments d'information pour la compréhension du tweet**

  Des éléments d'information complémentaires sont ajoutés aux noeuds du graphe via l'interaction (ajout du profil, etc). Il s'agit donc de vérifier l'utilité de cet ajout.

  Question 4 : Est-ce que vous trouvez que l'affichage des éléments d'informations est important pour la compréhension du tweet ? Les réponses à la question sont données sur une échelle de 0 à 3 similaire à celle de la question 3.

- **Évaluer le niveau de compréhension du contexte généré avec le contexte approprié du tweet**

  Dans la dernière phase de l'évaluation, nous présentons aux participants une description du contexte approprié au tweet initial et nous leurs demandons de comparer le contexte qu'ils ont compris, en exploitant les graphes de conversations et les éléments d'information associés par rapport à celui fourni par la description.

  Question5 : Est-ce que vous êtes parvenus à comprendre le contexte du tweet en exploitant les graphes de conversations ? Les réponses possibles à la question sont "parfaitement", "pas vraiment" et "non".

6.3.2 Les entretiens

La deuxième partie de notre évaluation est consacrée aux entretiens semi-structurés qui permettent de comparer, expliquer et élargir les résultats du questionnaire. En IHM on distingue plusieurs types d'interviews (Flick, 2009) et celui que nous avons retenu est l'entretien semi-structuré, basé sur des questions ouvertes pré-établies, destiné à des experts. Cette approche permet de "collecter des informations contextuelles complétant les idées issues de l'application d'autres méthodes" (Bogner et al., 2002).

Nous avons entretenu 6 évaluateurs experts à la fin de la validation de leur évaluation via le questionnaire. Les questions complémentaires ont été les suivantes :

- Question 6 : Quelle est votre expertise en recherche d'information ?
- Question 7 : Que pensez-vous de l'intégration des graphes de conversations pour la représentation du contexte et est-ce que un tel système est utile ?
- Question 8 : Quel est ou quels sont les éléments d'informations jugés importants pour la compréhension du contexte du tweet ?
- Question 9 : Quelles sont les améliorations que vous recommanderiez pour le système proposé ?

Les entretiens ont été effectués en face à face et ont été enregistrés.

6.4 Résultats

Dans cette section, nous présentons les résultats de l'évaluation de notre approche.

6.4.1 Résultats issus du questionnaire

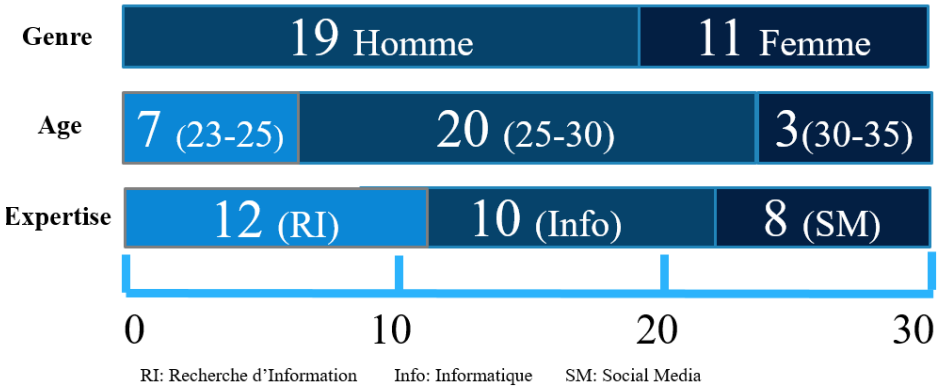

FIGURE 6.6 – Âge, sexe et expertise des participants

Comme l'indique la figure 6.6, les caractéristiques démographiques de notre échantillon sont les suivantes : 36 % des participants sont des femmes, et la tranche d'âge des participants est entre 23 et 35 ans. Plus de 60 % des participants ont une expertise dans le domaine de la recherche d'information ou des réseaux sociaux. Ces caractéristiques montrent un biais dans la constitution de l'échantillon. On retrouve ce biais dans de nombreuses analyses menées en IHM où les participants aux tests sont des membres de laboratoires de sciences du numérique avec une sur-représentation masculine et une pyramide des âges souvent jeune. Cette première évaluation devra être à l'avenir étendue à un échantillon plus large et plus représentatif mais elle nous sert ici de première analyse pour conforter les apports de notre système et intégrer les améliorations indiquées dans une nouvelle version du prototype.

- Question 1 (figure 6.7)

Il semble que la lecture des tweets additionnels apporte peu d'information sur le contexte du tweet et plus de 56.7 % des participants considèrent même que l'apport est nul. Ce résultat confirme notre hypothèse selon laquelle il nous semblait difficile de faciliter la compréhension d'un tweet par la proposition d'autres textes courts pouvant être eux-mêmes ambigus.

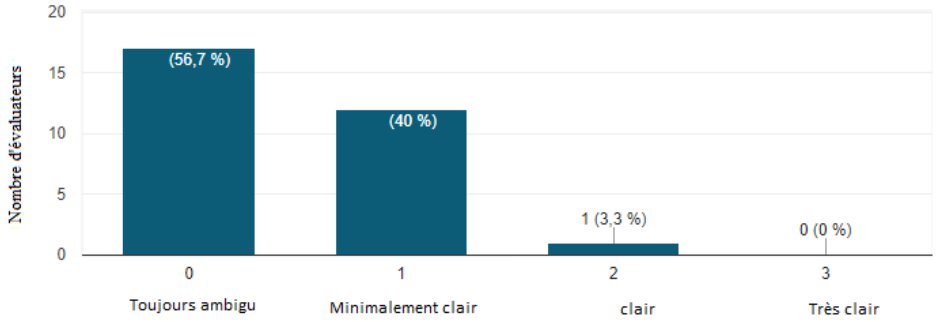

FIGURE 6.7 – Résultats de l'évaluation de l'apport des textes de tweets pour la compréhension du tweet ambigu

- Question 2 (figure 6.8)

La moitié des participants trouvent le contexte plus compréhensible avec l'exploitation des graphes de conversations, et 30 % pensent même que le contexte est clair. Il reste cependant 20% des participants qui ne confirment pas l'apport de la nouvelle contextualisation.

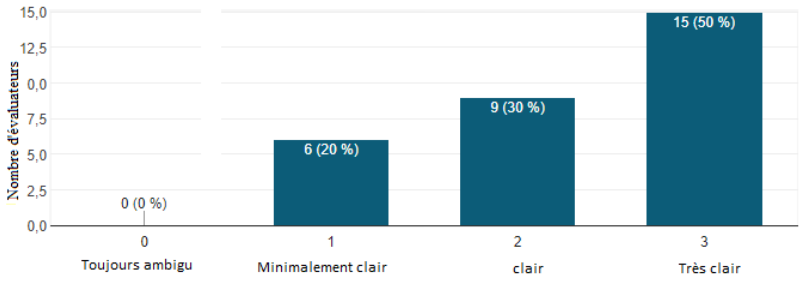

FIGURE 6.8 – Résultats de l'évaluation de l'apport des graphes de conversations pour la compréhension du tweet ambigu

- Question 3 (figure 6.9)

Contrairement à la question 1, les participants apprécient globalement positivement la contextualisation proposée. Plus de la moitié jugent l'apport des graphes comme "très important" et près d'un tiers comme "important".

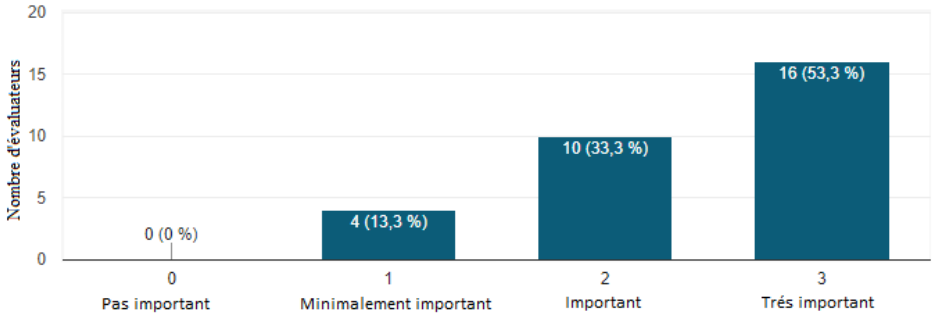

FIGURE 6.9 – Résultats de l'évaluation de l'importance des graphes de conversations par rapport à la lecture des textes de tweets

- Question 4 (figure 6.10)

Les participants sont sensibles aux éléments d'informations affichés sur les graphes de conversations. Plus de la majorité trouvent que ces éléments facilitent la compréhension du contexte et pensent également que la présentation du contexte est moins complexe et facile à comprendre que les informations uniquement textuelles.

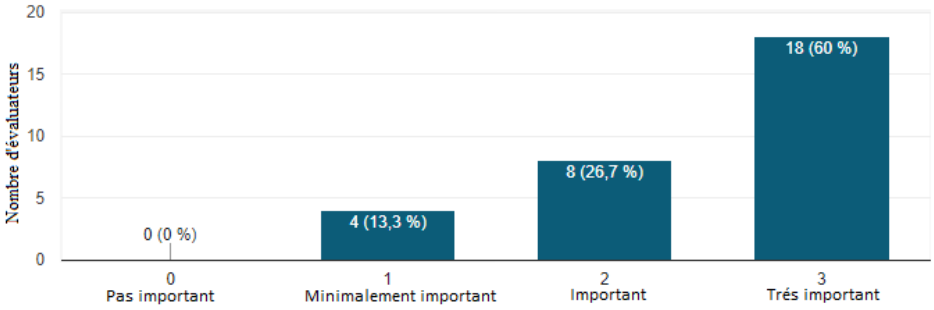

FIGURE 6.10 – Résultats de l'évaluation de l'apport des éléments d'informations pour la compréhension du tweet

- Question 5 (figure 6.11)

La majorité des participants (70 %) ont confirmé qu'ils étaient parvenus à comprendre, d'après eux, la bonne version du contexte approprié au tweet ambigu. Seule une minorité (6,7 %) a déclaré son échec dans la tâche.

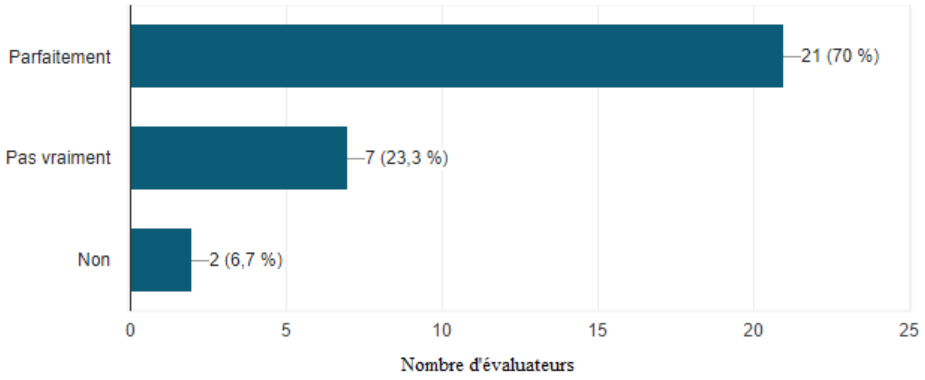

FIGURE 6.11 – Résultats de l'évaluation de la compréhension du contexte par rapport au contexte approprié

6.4.2 Résultats extraits des entretiens

Les six participants interrogés ont tous une expertise dans le domaine de la Recherche d'Information et des domaines associés tels que la recommandation sociale, l'analyse des sentiments sur Twitter, et la détection de communauté dans les réseaux sociaux.

Au cours des entretiens, tous les participants ont confirmé les résultats encourageants qui ressortent des questionnaires. Ils ont notamment insisté sur l'intérêt des graphes pour la représentation du contexte et l'originalité du couplage entre les conversations et la contextualisation. Interrogés sur l'importance des éléments d'informations, certains participants ont souligné l'importance des métadonnées de tweets dans son contexte ; et en particulier, les définitions des hashtags et l'exploitation des URLs des tweets. D'autres participants sont revenus sur l'intérêt des informations sur les auteurs des tweets (influence et popularité) ainsi que sur l'utilité de la consultation des profils.

Dans les points d'amélioration cités, la qualité de la visualisation et de l'interactivité est l'enjeu majeur pour l'avenir. Comme nous l'avons précisé en introduction de ce paragraphe, faute de temps, cet aspect n'a pas été au coeur de nos recherches, mais, dès l'instant où l'on s'appuie sur une démarche centrée sur l'utilisateur opérationnalisée par une interface visuelle,

la qualité de la restitution et les facilités de sa manipulation deviennent des problématiques importantes.

6.5 Conclusion

Dans ce chapitre, nous avons rappelé les différents protocoles d'évaluation mis en place lors du challenge INEX "Tweet contextualization" puis nous avons détaillé notre protocole d'évaluation. Les résultats obtenus sur un échantillon restreint (30 questionnaires et 6 entretiens avec des experts) semblent confirmer la pertinence de notre approche. Ils soulignent l'intérêt des graphes de conversations et des éléments d'information qui les complètent. Ils identifient aussi une voie d'amélioration qui concerne l'interface visuelle de restitution et de manipulation des informations contextuelles. Cette voie ouvre donc sur les perspectives présentées dans le dernier chapitre de ce rapport.

Conclusion

Aujourd'hui, Twitter représente 313 millions d'utilisateurs avec plus de 500 millions de tweets échangés chaque jour. Cependant, malgré cette vitalité, un nombre d'utilisateurs constatent la difficulté de compréhension de certains tweets. La qualité des textes des messages est en effet très variable, et l'interprétation d'un tweet brut sans information complémentaire peut être délicate. La contextualisation est donc une voie de recherche prometteuse pour faciliter l'accès au contenu informationnel de ces messages courts et bruités.

Ces dernières années, plusieurs systèmes ont été proposés pour exploiter le contenu des tweets par une analyse de ses termes pertinents et de ses liens. Ils peuvent servir de base à un enrichissement des contenus. D'autres approches se sont focalisées sur l'ajout d'informations contextuelles pour faciliter leur décryptage. En particulier, le challenge INEX "Tweet Contextualisation" a permis de faire des progrès notables dans cette direction (Bellot et al., 2013; SanJuan et al., 2012). Mais la voie choisie porte sur la construction d'un résumé construit à partir de ressources textuelles, et notamment pour le challenge de Wikipédia.

Dans cette thèse, nous avons choisi une voie complémentaire différente qui n'explore pas des ressources textuelles exogènes mais qui s'appuie sur l'analyse du réseau conversationnel lié au tweet ambigu. Par construction, Twitter offre des opérateurs simples (retweet, reply, etc) qui facilitent les échanges et engendrent ainsi des relations sociales, souvent furtives car liées à un évènement précis en cours, mais dynamiques. Ainsi, un tweet s'intègre dans une conversation d'un nouveau type qui peut se modéliser par un graphe où les sommets sont des individus ou des tweets et les relations modélisent les opérations accessibles. L'ouverture publique des échanges permet un accès à ses conversations qui se passent à des échelles pouvant varier du très local, pour des sujets d'intérêt personnel restreint, au mondial pour des sujets d'actualité.

L'intérêt croissant pour ces conversations a conduit au développement de recherches dans deux directions principales. Stimulés notamment par les perspectives prometteuses de la fouille d'opinions, des efforts conséquents ont été déployés cette dernière décennie pour tenter d'automatiser l'analyse textuelle des contenus des messages et ainsi améliorer la compréhension des sujets abordés. D'autres approches, complémentaires, ont porté leur attention sur l'analyse structurelle des graphes représentant les relations mises en oeuvre dans les conversations. Cette analyse permet donc d'identifier un système de relations dans lequel un tweet s'inscrit. Et, c'est dans cette direction que nous avons orienté nos travaux sur la contextualisation des tweets.

Dans ce cadre, l'apport de notre travail s'est organisé selon deux axes : l'analyse de l'organisation des conversations sur Twitter et la proposition d'une nouvelle contextualisation de tweet qui intègre ces données conversationnelles. Nous avons introduit un modèle de graphe de conversation qui intègre les différentes composantes les plus usuelles sur Twitter. Puis nous avons analysé leur structure via une étude expérimentale portant sur les distributions statistiques observées d'un ensemble d'indicateurs évaluant des caractéristiques combinatoires du graphe. Cette analyse nous a permis d'identifier les indicateurs pertinents pour la contextualisation et de nous guider sur leur paramétrage. Nous avons ainsi proposé un nouveau prototype de contextualisation d'un tweet ambigu qui renvoie à l'utilisateur des informations adaptées sur le graphe de discussion dans lequel le tweet est intégré. Ce prototype a été testé auprès d'un panel d'utilisateurs.

6.6 Perspectives

Les perspectives qu'ouvre ce travail portent sur trois aspects majeurs. La première porte sur les aspects techniques de l'interface afin d'améliorer l'expérience de l'utilisateur. La seconde vise à compléter les informations intégrées dans la contextualisation en exploitant les informations associées notamment aux individus impliqués dans les discussions accessibles sur d'autres réseaux sociaux. La troisième ouvre vers une nouvelle application de la contextualisation tournée vers les systèmes de recommandation. S'ajoute à ces perspectives de développement un nouvel axe applicatif stimulé par l'analyse de données historiques issues des corpus de tweets stockées par la bibliothèque du Congrès.

- **Vers le passage à l'échelle de la restitution visuelle**

  Le système de visualisation interactive développé au cours de cette thèse a été testé par une étude utilisateurs afin de valider les différents modules proposés. Toutefois, des améliorations devront être apportées au prototype en cours pour, d'une part améliorer la restitution visuelle des informations et, d'autre part permettre un déploiement en ligne. La visualisation n'était pas au coeur des travaux menés dans cette thèse et nous n'avons pas optimisé la fluidité des rendus. La volumétrie des données manipulées nécessite le développement des structures de données plus performantes pour la gestion des graphes de discussions et pour l'affichage. Des améliorations identifiées pourront également être apportées aux tracés des graphes afin de pouvoir les compléter, sans perdre en lisibilité, avec des informations sur les tweets et les individus impliqués dans les conversations. Ces améliorations permettront à terme d'envisager un déploiement en ligne du système afin de pouvoir mener des expérimentations à beaucoup plus grande échelle.

- **Vers l'intégration de sources d'informations complémentaires**

  Il nous semble pertinent de chercher à exploiter d'autres informations sociales pour améliorer la contextualisation des tweets. Dans un premier temps, nous envisageons d'approfondir nos recherches afin d'explorer d'autres réseaux sociaux, notamment Facebook. Les informations individuelles générées par des utilisateurs actifs identifiés sur nos graphes de conversations pourraient certainement permettre d'affiner la contextualisation. Les signaux sociaux tels que "J'aime" peuvent également contribuer à mesurer l'influence d'une information. Les travaux récents de Badache and Boughanem, 2015 ont effectivement montré le rôle joué dans ces signaux pour augmenter la qualité des résultats de requêtes (Badache and Boughanem, 2015). Nous envisageons d'explorer cette piste de recherche dans notre cadre ; il s'agira alors de définir un modèle d'agrégation de ces signaux avec ceux associés aux opérateurs d'interaction de Twitter pour compléter la contextualisation des messages courts.

- **Vers la recommandation de contextualisation des tweets**

  Les systèmes de recommandation aident l'utilisateur à filtrer, à travers une quantité écrasante de données, l'information potentiellement pertinente pour ses usages et ses centres d'intérêt. Les systèmes de recommandation sociale basés sur les pratiques des usagers des réseaux sociaux ont connu un essor important ces dernières années. Les exemples incluent la recommandation des tweets (Chen et al., 2012; Diaz-Aviles et al., 2012)), des utilisateurs à suivre (Gupta et al., 2013), des événements (Magnuson et al., 2015), des sujets (Liang et al., 2012), des hashtags (Kywe et al., 2012), des groupes (sous-communautés) à intégrer, etc. Dans ce contexte, nous envisageons de nous appuyer sur l'analyse des graphes de discussions menée ici pour proposer un système de recommandation qui offre la possibilité à un utilisateur de recevoir une contextualisation des tweets perçus comme

ambigus. Il s'agira en particulier de se focaliser sur les tweets publiés par les comptes auxquels un utilisateur est abonné. La mise en oeuvre de cette perspective se heurte néanmoins à un défi délicat qui concerne la définition opérationnelle de la notion d'ambigüité.

- **Vers une ouverture applicative interdisciplinaire : l'analyse des tweets archivés**

  Un corpus de tweets original a été récemment mis sous les feux des projecteurs à cause d'un changement de stratégie. La bibliothèque du Congrès [5] a lancé, depuis 2010, la collecte et l'archivage de tous les tweets publiés, publiquement, sur Twitter. Mais le volume accru des tweets et le doublement de la taille limite des tweets ont mis fin, début 2018, à ce stockage intégral [6] ; la bibliothèque a décidé de modifier sa stratégie de collecte et de passer de l'archivage exhaustif à un archivage sélectif. Cependant, qu'il s'agisse des corpus intégraux ou filtrés ces données offriront aux analystes de l'histoire contemporaine un témoignage de phénomènes sociaux adossés à un support nouveau. Si aujourd'hui une grande partie des travaux sur les données de Twitter portent sur des échelles temporelles courtes avec la détection de nouvelles tendances dans un flux twitter et le suivi d'évènements (Das Sarma et al., 2011; Ritter et al., 2012) ou l'identification de signaux faibles (Pépin et al., 2017), l'analyse des tweets historisés ouvre de nouvelles questions. En particulier, la contextualisation d'un tweet émis il y a plusieurs années sera certainement indispensable pour sa compréhension. Cette contextualisation pourra s'appuyer sur des documents multiples décrivant les évènements de la période d'émission mais son positionnement dans un réseau social devrait être aussi un élément très utile pour son interprétation et la mesure de son influence. Ces perspectives s'intègrent dans le domaine en plein essor des humanités numériques (Wieviorka, 2013; Dacos and Mounier, 2015; Bonaccorsi et al., 2016). La fertilisation croisée des sciences humaines et de l'informatique réinterroge à la fois les pratiques en oeuvre dans les sciences humaines mais aussi les usages de l'informatique et même plus profondément certains de ses concepts.

[5] https ://www.loc.gov/

[6] http ://www.france24.com/fr/20171227-usa-bibliotheque-congres-renonce-a-archiver-tous-tweets

Bibliographie

Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011). Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of the 8th Extended Semantic Web Conference on The Semanic Web : Research and Applications - Volume Part II*, ESWC'11, pages 375–389, Berlin, Heidelberg. Springer-Verlag. 13, 15, 18, 62

Alegre, U., Augusto, J. C., and Clark, T. (2016). Engineering context-aware systems and applications : A survey. *Journal of Systems and Software*, 117 :55–83. 7

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J., and Zhai, C. (2003). Challenges in information retrieval and language modeling : Report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst, september 2002. *SIGIR Forum*, 37(1) :31–47. 8

Alvarez-Melis, D. and Saveski, M. (2016). Topic modeling in twitter : Aggregating tweets by conversations. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 519–522. 24, 36, 39, 42, 57

Amer, N. O., Mulhem, P., and Géry, M. (2015). Recherche de conversations dans les réseaux sociaux : modélisation et expérimentations sur twitter. In *CORIA 2015 - Conférence en Recherche d'Infomations et Applications - 12th French Information Retrieval Conference, Paris, France, March 18-20, 2015.*, pages 55–70. 23, 24, 30, 31, 32, 36, 39, 42, 45, 58, 59

Ansary, K. H., Tran, A. T., and Tran, N. K. (2013). A pipeline tweet contextualization system at INEX 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.* 15

Badache, I. and Boughanem, M. (2015). Document priors based on time-sensitive social signals. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings*, pages 617–622. 93

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 7

Bagdouri, M. (2016). Journalists and twitter : A multidimensional quantitative description of usage patterns. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 22–31. 2

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer : Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA. ACM. 43

Bandyopadhyay, A., Mitra, M., and Majumder, P. (2011). Query expansion for microblog retrieval. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, November 15-18*. 18

Banerjee, S., Ramanathan, K., and Gupta, A. (2007). Clustering short texts using wikipedia. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 787–788, New York, NY, USA. ACM. 18

Barzilay, R., Elhadad, N., and McKeown, K. R. (2002). Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17(1) :35–55. 83

Bazire, M. and Brézillon, P. (2005). Understanding context before using it. In *Proceedings of the 5th International Conference on Modeling and Using Context*, CONTEXT'05, pages 29–40, Berlin, Heidelberg. Springer-Verlag. 7

Belkin, N. (1984). Cognitive models and information transfer. *Social Science Information Studies*, 4(2) :111 – 129. Special Issue Seminar on the Psychological Aspects of Information Searching. 7

Bellot, P., Doucet, A., Geva, S., Gurajada, S., Kamps, J., Kazai, G., Koolen, M., Mishra, A., Moriceau, V., Mothe, J., Preminger, M., SanJuan, E., Schenkel, R., Tannier, X., Theobald, M., Trappett, M., and Wang, Q. (2013). *Overview of INEX 2013*, pages 269–281. Springer Berlin Heidelberg, Berlin, Heidelberg. 3, 83, 92

Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., and Tannier, X. (2016). {INEX} tweet contextualization task : Evaluation, results and lesson learned. *Information Processing and Management*, 52(5) :801 – 819. 13

Bhaskar, P., Banerjee, S., and Bandyopadhyay, S. (2012). A hybrid tweet contextualization system using IR and summarization. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. 19

Bilal, D. (2000). Children's use of the yahooligans ! web search engine : 1. cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for information Science*, 51(7) :646–665. 12

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3) :1–22. 14

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., and Dodds, P. S. (2011). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *CoRR*, abs/1112.1010. 22, 24, 25, 38, 40, 43

Bogner, A., Littig, B., and Menz, W., editors (2002). *Das Experteninterview : Theorie, Methode, Anwendung*. Leske & Budrich, Opladen. 87

Bonaccorsi, J., Carayol, V., and DOMENGET, J.-C. (2016). *Humanités Numériques et Sciences de l'Information et de la communication*. Number 8 in RFSIC. 94

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12) :2301–2309. 79

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet : Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA. IEEE Computer Society. 24, 36, 39, 42

Brandes, U., Freeman, L. C., and Wagner, D. (2013). Social networks. In Tamassia, R., editor, *Handbook on Graph Drawing and Visualization.*, pages 805–839. 80

Brézillon, P. (1999). Context in problem solving : A survey. *Knowl. Eng. Rev.*, 14(1) :47–80. 7

Brézillon, P. (2002). Expliciter le Contexte dans les objets Communicants. In *Les Objets Communicants*, pages 295–303. Hermès. 9

Brézillon, P. and Gonzalez, A. J., editors (2014). *Context in Computing - A Cross-Disciplinary Approach for Modeling the Real World*. Springer. 6

Brézillon, P. and Marquois, E. (2004). Context-based representation of the task/method paradigm. In *FLAIRS Conference*, pages 581–586. 9

Bricon-Souf, N. and Newman, C. R. (2007). Context awareness in health care : A review. *International Journal of Medical Informatics*, 76(1) :2 – 12. 9

Carpendale, S. (2008). Information visualization. chapter Evaluating Information Visualizations, pages 19–45. Springer-Verlag, Berlin, Heidelberg. 79

Carter, S., Weerkamp, W., and Tsagkias, M. (2013). Microblog language identification : Overcoming the limitations of short, unedited and idiomatic text. *Lang. Resour. Eval.*, 47(1) :195–215. 2, 63

Chen, J., Nairn, R., and Chi, E. (2011). Speak little and well : Recommending conversations in online social streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 217–226, New York, NY, USA. ACM. 23, 24, 32, 36, 39

Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., and Yu, Y. (2012). Collaborative personalized tweet recommendation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 661–670, New York, NY, USA. ACM. 93

Cheng, F., Zhang, X., He, B., Luo, T., and Wang, W. (2013). A survey of learning to rank for real-time twitter search. In *Proceedings of the 2012 International Conference on Pervasive Computing and the Networked World*, ICPCA/SWS'12, pages 150–164, Berlin, Heidelberg. Springer-Verlag. 19

Cogan, P., Andrews, M., Bradonjic, M., Kennedy, W. S., Sala, A., and Tucci, G. (2012). Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, HotSocial '12, pages 25–31, New York, NY, USA. ACM. 22, 24, 25, 36, 37, 39, 41, 42, 55, 64

Cool, C. and Spink, A. (2002). Issues of context in information retrieval (ir) : An introduction to the special issue. *Information Processing & Management*, 38(5) :605–611. 9, 10

Crouch, C. J., Crouch, D. B., Chittilla, S., Nagalla, S., Kulkarni, S., and Nawale, S. (2012). The 2012 INEX snippet and tweet contextualization tasks. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. 84

Crowther, J., editor (1995). *Oxford Advanced Learner's Dictionary*. Cornelsen & Oxford, 5th edition. 7

Dacos, M. and Mounier, P. (2015). Humanités numériques : état des lieux et positionnement de la recherche française dans le contexte international. Technical report, Institut français. 94

Damak, F., Pinel-Sauvagnat, K., Boughanem, M., and Cabanac, G. (2013). Effectiveness of state-of-the-art features for microblog search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 914–919, New York, NY, USA. ACM. 19

Danescu-Niculescu-Mizil, C., Gamon, M., and Dumais, S. (2011). Mark my words ! : Linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 745–754, New York, NY, USA. ACM. 22, 24, 37, 39

Dann, S. (2015). Benchmarking micro-blog performance : Twitter content classification framework. In *Maximizing Commerce and Marketing Strategies through Micro-Blogging*, pages 313–332. IGI Global. 3

Das Sarma, A., Jain, A., and Yu, C. (2011). Dynamic relationship and event discovery. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 207–216, New York, NY, USA. ACM. 94

de Tarde, G. (1901). *L'opinion et la foule*. Paris, Alcan. 3

Degenne, A. and Forsé, M. (2004). Les réseaux sociaux. 56

Dervin, B. (1997). Given a context by any other name : Methodological tools for taming the unruly beast. In *Proceedings of an International Conference on Information Seeking in Context*, ISIC '96, pages 13–38, London, UK, UK. Taylor Graham Publishing. 8

Deveaud, R. and Boudin, F. (2012). LIA/LINA at the INEX 2012 tweet contextualization track. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. 84

Deveaud, R. and Boudin, F. (2013a). Contextualisation automatique de tweets à partir de wikipédia. In *CORIA 2013 - Conférence en Recherche d'Infomations et Applications - 10th French Information Retrieval Conference, Neuchâtel, Suisse, April 3-5, 2013.*, pages 125–140. 13, 15

Deveaud, R. and Boudin, F. (2013b). Effective tweet contextualization with hashtags performance prediction and multi-document summarization. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. 18, 19, 63

Deveaud, R. and Boudin, F. (2014). De quoi parle ce Tweet ? Résumer Wikipédia pour contextualiser des microblogs. *Revue I3 - Information Interaction Intelligence*, pages 37–56. 72

Develotte, C., Kern, R., and Lamy, M.-N. (2011). *Décrire la conversation en ligne : le face à face distanciel*. ENS éditions. 3

Dey, A. K. (2001). Understanding and using context. *Personal Ubiquitous Computing*, 5(1) :4–7. 7, 11

Dey, A. K. and Abowd, G. D. (1999). Towards a better understanding of context and context-awareness. In *In HUC99 : Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing*, pages 304–307. Springer-Verlag. 9, 10

Diakopoulos, N. A. and Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA. ACM. 1, 2

Diaz-Aviles, E., Drumond, L., Gantner, Z., Schmidt-Thieme, L., and Nejdl, W. (2012). What is happening right now ... that interests me ? : Online topic discovery and recommendation in twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1592–1596, New York, NY, USA. ACM. 93

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network : Hedonometrics and twitter. *CoRR*, abs/1101.5120. 43

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. (2010). Time is of the essence : Improving recency ranking using twitter data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 331–340, New York, NY, USA. ACM. 1, 2

Dorogovtsev, S. N. and Mendes, J. F. F. (2003). *Evolution of Networks : From Biological Nets to the Internet and WWW (Physics)*. Oxford University Press, Inc., New York, NY, USA. 55

Dou, W., Wang, X., Skau, D., Ribarsky, W., and Zhou, M. X. (2012). Leadline : Interactive visual analysis of text data through event identification and exploration. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 93–102. 80

Dourish, P. (2004). What we talk about when we talk about context. *Personal Ubiquitous Computing*, 8(1) :19–30. 8

Doyle, G. and Frank, M. C. (2015). Audience size and contextual effects on information density in Twitter conversations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 637–648. International World Wide Web Conferences Steering Committee. 22, 24, 37, 40

Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 295–303, Stroudsburg, PA, USA. Association for Computational Linguistics. 30

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1) :61–74. 17

Efron, M. (2010). Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 787–788, New York, NY, USA. ACM. 63

Erkan, G. and Radev, D. R. (2004). Lexrank : Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1) :457–479. 16, 17

Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J., and SanJuan, E. (2016). Cultural micro-blog contextualization 2016 workshop overview : data and pilot tasks. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 1197–1200. 85

Ermakova, L. and Mothe, J. (2012a). IRIT at INEX 2012 : Tweet contextualization. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. 15, 84, 85

Ermakova, L. and Mothe, J. (2012b). *IRIT at INEX : Question Answering Task*, pages 219–226. Springer Berlin Heidelberg, Berlin, Heidelberg. 84

Ermakova, L., Mothe, J., and SanJuan, E. (2017). CLEF 2017 microblog cultural contextualization content analysis task overview. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.* 85

Filatova, E. and Hatzivassiloglou, V. (2004). A formal model for information selection in multi-sentence text extraction. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics. 16

Flick, U. (2009). *An Introduction to Qualitative Research*. SAGE Publications. 87

Frias-Martinez, E., Chen, S. Y., Macredie, R. D., and Liu, X. (2007). The role of human factors in stereotyping behavior and perception of digital library users : A robust clustering approach. *User Modeling and User-Adapted Interaction*, 17(3) :305–337. 12

Fuchs, C. (1996). *Les ambiguïtés du français*. Collection L'essentiel français. Ophrys. 71

Fuhr, N. (2001). Models in information retrieval. In Agosti, M., Crestani, F., and Pasi, G., editors, *Lectures on Information Retrieval*, volume 1980 of *Lecture Notes in Computer Science*, pages 21–50. Springer Berlin Heidelberg. 10

Fung, P. and Ngai, G. (2006). One story, one flow : Hidden markov story models for multilingual multidocument summarization. *ACM Trans. Speech Lang. Process.*, 3(2) :1–16. 16

Galley, M. (2006). A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 364–372, Stroudsburg, PA, USA. Association for Computational Linguistics. 16

Gandon, F. (2017). Tim Berners-Lee, lauréat du prix Turing 2016 pour avoir inventé... le Web. *Bulletin de la société informatique de France*, (11) :129–154. 1

Ganguly, D., Leveling, J., and Jones, G. J. F. (2012). Dcu@inex-2012 : Exploring sentence retrieval for tweet contextualization. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. 15, 84

Gansner, E. R., Hu, Y., and North, S. (2013). Visualizing streaming text data with dynamic graphs and maps. In *Proceedings of the 20th International Conference on Graph Drawing*, GD'12, pages 439–450, Berlin, Heidelberg. Springer-Verlag. 80

Genc, Y., Sakamoto, Y., and Nickerson, J. V. (2011). Discovering context : Classifying tweets through a semantic transform based on wikipedia. In *Proceedings of the 6th International Conference on Foundations of Augmented Cognition : Directing the Future of Adaptive Systems*, FAC'11, pages 484–492, Berlin, Heidelberg. Springer-Verlag. 15, 18

George, S. (2004). Analyse automatique de conversations textuelles synchrones d'apprenants pour la détermination de comportements sociaux. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation (STICEF)*, 10 :p–165. 2

Gillon, B. S. (1990). Ambiguity, generality, and indeterminacy : Tests and definitions. *Synthese*, 85(3) :391–416. 71, 72

Göker, A. and Myrhaug, H. I. (2002). User context and personalisation. In *ECCBR Workshop on Case Based Reasoning and Personalisation, Aberdeen*. 8, 9, 11

Gonçalves, B., Perra, N., and Vespignani, A. (2011). Modeling users' activity on twitter networks : Validation of dunbar's number. *PLOS ONE*, 6(8) :1–5. 43

Greenwood, S., Perrin, A., and Duggan, M. (2016). Social media update 2016. *Pew Research Center*, 11. 2

Greetham, D. V. and Ward, J. A. (2014). Conversations on twitter : Structure, pace, balance. In *Proceedings of the 2nd International Workshop on Dynamic Networks and Knowledge Discovery, DyNaK 2014, Nancy, France, September 15, 2014*. 22, 24, 26, 38, 40, 42, 55

Grice, H. P. (1957). Meaning. *The philosophical review*, 66(3) :377–388. 71

Griffin, N. R., Fleck, C. R., Uitvlugt, M. G., Ravizza, S. M., and Fenn, K. M. (2017). The tweeter matters : Factors that affect false memory from twitter. *Computers in Human Behavior*, 77(Supplement C) :63 – 68. 2

Guo, W., Li, H., Ji, H., and Diab, M. Linking tweets to news : A framework to enrich short text data in social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 15

Guo, Y., Qin, B., Liu, T., and Li, S. (2013). Microblog entity linking by leveraging extra posts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 863–868, Seattle, Washington, USA. Association for Computational Linguistics. 15, 18

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., and Zadeh, R. (2013). Wtf : The who to follow service at twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 505–514, New York, NY, USA. ACM. 93

Gupta, S., Nenkova, A., and Jurafsky, D. (2007). Measuring importance and query relevance in topic-focused multi-document summarization. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 193–196, Stroudsburg, PA, USA. Association for Computational Linguistics. 17

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., yen Kan, M., and McKeown, K. R. (2001). Simfinder : A flexible clustering tool for summarization. In *In Proceedings of the NAACL Workshop on Automatic Summarization*, pages 41–49. 17

Heer, J. and Agrawala, M. (2006). Software design patterns for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5) :853–860. 79

Herrera, J. M., Parra, D., and Poblete, B. (2015). Retrieving relevant conversations for q&a on twitter. In *Proceedings of the International Workshop on Social Personalisation & Search, SPS 2015, co-located with the 38th Annual ACM SIGIR Conference (SIGIR 2015), Santiago de Chile, Chile, August 9-13, 2015.*, pages 21–25. 23, 24, 26, 36, 39

Honey, C. and Herring, S. (2009). Beyond microblogging : Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, pages 1–10. 3, 22, 24, 27, 38, 40, 42

Hovy, E. and Lin, C. Y. (1999). Automated Text Summarization in SUMMARIST. In Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*. MIT Press. 16

Hu, Y., Talamadupula, K., Kambhampati, S., et al. (2013). Dude, srsly ? : The surprisingly formal nature of twitter's language. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM*. 2

Huang, J., Thornton, K. M., and Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 173–178, New York, NY, USA. ACM. 63

Hupfer, M. E. and Detlor, B. (2006). Gender and web information seeking : A self-concept orientation model : Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 57(8) :1105–1115. 12

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities : Elements of a cognitive theory for information retrieval interaction. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 101–110, New York, NY, USA. Springer-Verlag New York, Inc. 66

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction : elements of a cognitive ir theory. *Journal of Documentation*, 52 :3–50. 8

INGWERSEN, P. (1996). Cognitive perspectives of information retrieval interaction : Elements of a cognitive ir theory. *Journal of Documentation*, 52(1) :3–50. 13

Ingwersen, P. and Järvelin, K. (2005). Information retrieval in context : Irix. *SIGIR Forum*, 39(2) :31–39. 8

Ingwersen, P. and Järvelin, K. (2005). *The Turn : Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 9, 13

JaKobson, R. (1960). Linguistique et poétique. essais de linguistique générale. 71

Jang, J., Choi, J., Jang, G., and Myaeng, S. (2012). Semantic social networks constructed by topical aspects of conversations : An explorative study. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*. 24, 37, 40, 42

Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T. (1998). Real life information retrieval : A study of user queries on the web. *SIGIR Forum*, 32(1) :5–17. 18

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter : Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM. 3

Johnpaul, C. I. and Mathew, T. (2017). A cypher query based nosql data mining on protein datasets using neo4j graph database. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1–6. 80

Kazai, G. and Milic-Frayling, N. (2008). Trust, authority and popularity in social information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1503–1504, New York, NY, USA. ACM. 13

Kazai, G. and Milic-Frayling, N. (2009). Effects of social approval votes on search performance. In *Information Technology : New Generations, 2009. ITNG '09. Sixth International Conference on*, pages 1554–1559. 13

Kempson, R. (1977). *Semantic Theory*. Cambridge Textbooks in Linguistics. Cambridge University Press. 71

Kim, K.-S. (2008). Effects of emotion control and task on web searching behavior. *Information Processing & Management*, 44(1) :373 – 385. Evaluation of Interactive Information Retrieval Systems. 12

Kim, S., Bak, J., and Oh, A. H. (2012). Do you feel what I feel ? social aspects of emotions in twitter conversations. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*. 22, 24, 37, 40, 42

Kumar, R., Mahdian, M., and McGlohon, M. (2010). Dynamics of conversations. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 553–562, New York, NY, USA. ACM. 22, 24, 38, 40, 41, 55

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). 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, New York, NY, USA. ACM. 18, 41, 43

Kywe, S. M., Hoang, T.-A., Lim, E.-P., and Zhu, F. (2012). On recommending hashtags in twitter networks. In *Proceedings of the 4th International Conference on Social Informatics*, SocInfo'12, pages 337–350, Berlin, Heidelberg. Springer-Verlag. 93

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. (2012). Empirical studies in information visualization : Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9) :1520–1536. 79

Lang, K. (1995). NewsWeeder : learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, pages 331–339. Morgan Kaufmann publishers Inc. : San Mateo, CA, USA. 12

Laniado, D. and Mika, P. (2010). Making sense of twitter. In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part I*, ISWC'10, pages 470–485, Berlin, Heidelberg. Springer-Verlag. 18

Lee, P., Lakshmanan, L. V., and Milios, E. (2013). Keysee : Supporting keyword search on evolving events in social streams. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1478–1481, New York, NY, USA. ACM. 80

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2) :167–195. 14

Leung, C. W.-k., Chan, S. C.-f., and Chung, F.-l. (2006). A collaborative filtering framework based on fuzzy association rules and multiple-level similarity. *Knowl. Inf. Syst.*, 10(3) :357–381. 13

Li, C., Sun, A., and Datta, A. (2012). Twevent : Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 155–164, New York, NY, USA. ACM. 72

Li, F. (2010). Textual analysis of corporate disclosures : A survey of the literature. *Journal of accounting literature*, 29 :143. 2

Liang, H., Xu, Y., Tjondronegoro, D., and Christen, P. (2012). Time-aware topic recommendation based on micro-blogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1657–1661, New York, NY, USA. ACM. 93

Lin, C.-Y. and Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 495–501, Stroudsburg, PA, USA. Association for Computational Linguistics. 17

Louis, A. and Nenkova, A. (2009). Performance confidence estimation for automatic summarization. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 541–548, Stroudsburg, PA, USA. Association for Computational Linguistics. 82

Lu, R. and Yang, Q. (2012). Trend analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, 2(3) :327. 80

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4) :309–317. 7

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2) :159–165. 16

Luzzati, D. (2007). Le dialogue oral spontané : quels objets pour quels corpora. *Revue d'Interaction Homme-Machine*, 8(2). 23

Magnani, M. and Montesi, D. (2010). Toward conversation retrieval. In *Digital Libraries : 6th Italian Research Conference, IRCDL 2010, Padua, Italy, January 28-29, 2010*, volume 91 of *Communications in Computer and Information Science*, pages 173–182. Springer Berlin Heidelberg. 24, 31, 32, 36, 39, 42, 45

Magnani, M., Montesi, D., Nunziante, G., and Rossi, L. (2011). Conversation retrieval from twitter. In Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., and Mudoch, V., editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 780–783. Springer Berlin Heidelberg. 24

Magnani, M., Montesi, D., and Rossi, L. (2012). Conversation retrieval for microblogging sites. *Information Retrieval*, 15(3-4) :354–372. 23, 24, 31, 42

Magnuson, A., Dialani, V., and Mallela, D. (2015). Event recommendation using twitter activity. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 331–332, New York, NY, USA. ACM. 93

Manning, C. D., Schütze, H., et al. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press. 2

Mansourian, Y. (2008). Contextualization of web searching : a grounded theory approach. *The Electronic Library*, 26(2) :202–214. 9

Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge University Press, New York, NY, USA. 8

Mathew, B., A, U., Chakraborty, T., Ganguly, N., and Datta, S. (2016). Mining twitter conversations around e-commerce promotional events. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, CSCW '16 Companion, pages 345–348, New York, NY, USA. ACM. 22, 24, 36, 39

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater : pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8) :1179–1186. 80

McCarthy, J. (1993). Notes on formalizing context. In *Proceedings of the 13th International Joint Conference on Artifical Intelligence - Volume 1*, IJCAI'93, pages 555–560, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 7

McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., and Eskin, E. (1999). Towards multidocument summarization by reformulation : Progress and prospects. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 453–460, Menlo Park, CA, USA. American Association for Artificial Intelligence. 17

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 889–892, New York, NY, USA. ACM. 57

Meij, E., Trieschnigg, D., de Rijke, M., and Kraaij, W. (2010). Conceptual language models for domain-specific retrieval. *Inf. Process. Manage.*, 46(4) :448–469. 2

Meij, E., Weerkamp, W., and de Rijke, M. (2012). Adding semantics to microblog posts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 563–572, New York, NY, USA. ACM. 18

Mendes, P. N., Passant, A., Kapanipathi, P., and Sheth, A. P. (2010). Linked open social signals. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 224–231. 18

Mercier, A. (2013). Twitter l'actualité : usages et réseautage chez les journalistes français. *Recherches en communication*, 39(39) :111–132. 2

Merriam-Webster, I. S. (2006). *Merriam-Webster Online Dictionary*. A Merriam-Webster. Merriam-Webster. 7

Michelson, M. and Macskassy, S. A. (2010). Discovering users' topics of interest on twitter : A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA. ACM. 18

Mihalcea, R. and Tarau, P. (2004). TextRank : Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain. 17

Morchid, M., Dufour, R., and Linarès, G. (2013). Thematic Representation of Short Text Messages with Latent Topics : Application in the Twitter context. In *PACLING 2013*, Tokyo, Japan. 18

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough ? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv :1306.5204*. 80

Munzner, T. (2008). Information visualization. chapter Process and Pitfalls in Writing Information Visualization Research Papers, pages 134–153. Springer-Verlag, Berlin, Heidelberg. 79

Naaman, M., Boase, J., and Lai, C.-H. (2010). Is it really about me ? : Message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 189–192, New York, NY, USA. ACM. 2

Nagmoti, R., Teredesai, A., and De Cock, M. (2010). Ranking approaches for microblog search. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 153–157, Washington, DC, USA. IEEE Computer Society. 30

Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. (2011). Searching microblogs : Coping with sparsity and document quality. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 183–188, New York, NY, USA. ACM. 2

Nicolas, D. (2006). Ambiguïté. In D. Godard, L. R. . F. C. e., editor, *Sémanticlopédie : Dictionnaire de Sémantique*. 72

Nishi, R., Takaguchi, T., Oka, K., Maehara, T., Toyoda, M., Kawarabayashi, K.-i., and Masuda, N. (2016). Reply trees in twitter : data analysis and branching process models. *Social Network Analysis and Mining*, 6(1) :1–13. 22, 24, 37, 39, 41, 55

Olariu, A. (2013). Hierarchical clustering in improving microblog stream summarization. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 424–435. Springer Berlin Heidelberg. 16

Pakray, P., Bhaskar, P., Banerjee, S., Pal, B. C., Bandyopadhyay, S., and Gelbukh, A. F. (2011). A hybrid question answering system based on information retrieval and answer validation. In *IN CLEF 2011 Workshop on QA4MRE*. 19

Panckhurst, R. (1997). La communication «médiatisée» par ordinateur ou la communication «médiée» par ordinateur ? *Terminologies nouvelles*, (17) :56–58. 3

Pépin, L., Kuntz, P., Blanchard, J., Guillet, F., and Suignard, P. (2017). Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets. *Computers & Industrial Engineering*, 112 :450–458. 94

Phelan, O., McCarthy, K., and Smyth, B. (2009). Using twitter to recommend real-time topical news. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 385–388, New York, NY, USA. ACM. 1, 2

Plaisant, C. (2004). The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pages 109–116, New York, NY, USA. ACM. 79

Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. (2011). Do all birds tweet the same ? : Characterizing twitter around the world. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1025–1030, New York, NY, USA. ACM. 2

Pontes, E. L., Torres-Moreno, J., Huet, S., and Linhares, A. C. (2016). Tweet contextualization using continuous space vectors : Automatic summarization of cultural documents. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 1238–1245. 19

Powers, M. and Vera-Zambrano, S. (2017). How journalists use social media in france and the united states : Analyzing technology use across journalistic fields. *New Media & Society*. 2

Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., and Bhatt, S. (2013). What kind of #conversation is twitter ? mining #psycholinguistic cues for emergency coordination. *Comput. Hum. Behav.*, 29(6) :2438–2447. 22, 24, 27, 36, 39, 41

Qazvinian, V. and Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 689–696, Stroudsburg, PA, USA. Association for Computational Linguistics. 17

Ravikumar, S., Balakrishnan, R., and Kambhampati, S. (2012). Ranking tweets considering trust and relevance. In *Proceedings of the Ninth International Workshop on Information Integration on the Web*, IIWeb '12, pages 4 :1–4 :4, New York, NY, USA. ACM. 19

Ritter, A., Cherry, C., and Dolan, B. (2010). Unsupervised modeling of twitter conversations. pages 172–180. 3, 24, 37, 39, 41, 42

Ritter, A., Mausam, Etzioni, O., and Clark, S. (2012). Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1104–1112, New York, NY, USA. ACM. 94

Romero, D. M., Tan, C., and Ugander, J. (2011). Social-topical affiliations : The interplay between structure and popularity. *CoRR*, abs/1112.1115. 22, 24, 27, 36

Russell, M. G., Flora, J., Strohmaier, M., Poschko, J., and Rubens, N. (2011). Semantic Analysis of Energy-Related Conversations in Social Media : A Twitter Case Study. In *International Conference of Persuasive Technology (Persuasive 2011)*, Columbus, OH, USA. 22, 24, 30

Saggion, H. and Lapalme, G. (2002). Generating indicative-informative summaries with sumum. *Comput. Linguist.*, 28(4) :497–526. 15

Saggion, H., Torres-Moreno, J.-M., Cunha, I. d., and SanJuan, E. (2010). Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10, pages 1059–1067, Stroudsburg, PA, USA. Association for Computational Linguistics. 82

Salton, G. (1968). *Automatic Information Organization and Retrieval.* McGraw Hill Text. 7

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA. 7

SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., and Mothe, J. (2012). *Overview of the INEX 2011 Question Answering Track (QA@INEX)*, pages 188–206. Springer Berlin Heidelberg, Berlin, Heidelberg. 3, 82, 83, 92

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). Twitterstand : News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, USA. ACM. 1, 2

Schiffman, B., McKeown, K., Grishman, R., and Allan, J. (2007). Question answering using integrated information retrieval and information extraction. In *In Proceedings of HLT-NAACL*, pages 532–539. 19

Shanteau, J. (1992). Competence in experts : The role of task characteristics. *Organizational Behavior and Human Decision Processes*, 53(2) :252 – 266. Experts and Expert Systems. 73

Siddharthan, A., Nenkova, A., and McKeown, K. (2004). Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics. 17

Šilić, A. and Bašić, B. D. (2010). Visualization of text streams : A survey. In Setchi, R., Jordanov, I., Howlett, R. J., and Jain, L. C., editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 31–43, Berlin, Heidelberg. Springer Berlin Heidelberg. 80

Singhal, A. (2001). Modern Information Retrieval : A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4) :35–42. 7

Smith, M., Barash, V., Getoor, L., and Lauw, H. W. (2008). Leveraging social context for searching social media. In *Proceedings of the 2008 ACM Workshop on Search in Social Media*, SSM '08, pages 91–94, New York, NY, USA. ACM. 13

Smyth, B. and Balfe, E. (2006). Anonymous personalization in collaborative web search. *Information Retrieval*, 9(2) :165–190. 12

Sonnenwald, D. H. (1999). Evolving perspectives of human information behavior : Contexts, situations, social networks and information horizons. In *Exploring the contexts of information behavior : Proceedings of the Second International Conference in Information Needs*. Taylor Graham. 8

Sousa, D., Sarmento, L., and Mendes Rodrigues, E. (2010). Characterization of the twitter @replies network : Are user ties social or topical ? In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC'10, pages 63–70, New York, NY, USA. ACM. 24, 37, 40, 42, 43

Steiner, H. (2009). Reference utility of social networking sites : options and functionality. *Library Hi Tech News*, 26(5/6) :4–6. 3

Sugiyama, K., Hatano, K., and Yoshikawa, M. (2004). Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 675–684, New York, NY, USA. ACM. 8

Sweet, T. M. (2017). A review of statistical rethinking : A bayesian course with examples in r and stan. *Journal of Educational and Behavioral Statistics*, 42(1) :107–110. 80

Tamine, L., Soulier, L., Ben Jabeur, L., Amblard, F., Hanachi, C., Hubert, G., and Roth, C. (2015). Accès à l'information dans les réseaux sociaux : quelles formes de collaboration ? In *MARAMI 2015 – Modèles et Analyses Réseau : Approches Mathématiques et Informatiques*, Actes de MARAMI 2015 – Modèles et Analyses Réseau : Approches Mathématiques et Informatiques, Nîmes, France. 22, 24, 25, 36, 39, 41, 42, 55

Tamine-Lechani, L. (2008). *De la recherche d'information orientée sytème à la recherche d'information orientée contexte : Verrous, contributions et perspectives*. PhD thesis, Université Paul Sabatier-Toulouse III. 66

Tamine-Lechani, L., Boughanem, M., and Daoud, M. (2010). Evaluation of contextual information retrieval effectiveness : overview of issues and research. *Knowledge and Information Systems*, 24(1) :1–34. iv, 9, 12

Teevan, J., Ramage, D., and Morris, M. R. (2011). #twittersearch : A comparison of microblog search and web search. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 35–44, New York, NY, USA. ACM. 2

Temdee, P. and Prasad, R. (2018). Context and its awareness. In *Context-Aware Communication and Computing : Applications for Smart Environment*, pages 15–31. Springer. 7

Thomas, J. (2005). *Illuminating the Path : The Research and Development Agenda for Visual Analytics*. IEEE Computer Society Press. 80

Tombros, A., Ruthven, I., and Jose, J. M. (2005). How users assess web pages for information seeking. *Journal of the American Society for Information Science and Technology*, 56(4) :327–344. 13

Torres-Moreno, J. and Ramırez, J. (2010). Reg : un algorithme glouton appliqué au résumé automatique de texte. In *Proceedings of the 10th Int. Conference on the Statistical Analysis of Textual. Roma, Italia*. 84

Torres-Moreno, J. and Velázquez-Morales, P. (2012). Two statistical summarizers at INEX 2012 tweet contextualization track. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. 84

Traverso, V. (1999). L'analyse des conversations. *128(Paris)*. 3, 23

Van Oers, B. (1998). From context to contextualizing. *Learning and instruction*, 8(6) :473–488. 9

Vivaldi, J. and da Cunha, I. (2013). Tweet contextualization : a strategy based on document retrieval using query enrichment and automatic summarization. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. 18, 19, 85

Weerasooriya, T., Perera, N., and Liyanage, S. R. (2016). A method to extract essential keywords from a tweet using nlp tools. In *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 29–34. 72

Wen, X., Lin, Y.-R., Trattner, C., and Parra, D. (2014). Twitter in academic conferences : Usage, networking and participation over time. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 285–290, New York, NY, USA. ACM. 2, 3

Wieviorka, M. (2013). *L'impératif numérique ou La nouvelle ère des sciences humaines et sociales ?* Col·lecció Debats. CNRS. 94

Winograd, T. (2001). Architectures for context. *Hum.-Comput. Interact.*, 16(2) :401–419. 9

Wu, J. (2013). *Saving God's Face : A Chinese Contextualization of Salvation Through Honor and Shame*. William Carey International University Press. 9

Xie, H. I. (2008). Users' evaluation of digital libraries (dls) : Their uses, their criteria, and their assessment. *Inf. Process. Manage.*, 44(3) :1346–1373. 13

Yahia, S. A., Benedikt, M., Lakshmanan, L. V. S., and Stoyanovich, J. (2008). Efficient network aware search in collaborative tagging sites. *Proc. VLDB Endow.*, 1(1) :710–721. 13

Yardi, S. and Boyd, D. (2010). Tweeting from the town square : Measuring geographic local networks. The AAAI Press. 1, 2

Zhao, L., Zeng, Y., and Zhong, N. (2011). A weighted multi-factor algorithm for microblog search. In *Proceedings of the 7th International Conference on Active Media Technology*, AMT'11, pages 153–161, Berlin, Heidelberg. Springer-Verlag. 30

Zingla, M. A., Ettaleb, M., Latiri, C. C., and Slimani, Y. (2014). INEX2014 : tweet contextualization using association rules between terms. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 574–584. 85

Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K., and Tolmie, P. (2015). Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 347–353. 24, 25, 36, 39

l'université nantes angers le mans

PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Rami BELKAROUI

Vers une contextualisation des tweets basée sur une analyse des graphes de conversations

Towards a tweets contextualization based on conversation graphs analysis

Résumé

Même avec le récent passage à 280 caractères, les messages de Twitter considérés dans leur singularité, sans information additionnelle exogène, peuvent confronter leurs lecteurs à des difficultés d'interprétation. L'ajout d'une contextualisation à ces messages s'avère donc une voie de recherche prometteuse pour faciliter l'accès à leur contenu informationnel. Dans la dernière décennie, la majorité des travaux se sont concentrés sur la construction de résumés à partir de sources d'information complémentaires telles que Wikipédia. Nous avons choisi dans cette thèse une voie complémentaire différente qui s'appuie sur l'analyse des conversations sur Twitter afin d'extraire des informations utiles à la contextualisation d'un tweet. Ces informations ont été intégrées dans un prototype qui, pour un tweet donné, propose une visualisation d'un sous-graphe du graphe de conversation associé au tweet. Ce sous-graphe extrait automatiquement à partir de l'analyse des distributions des indicateurs structurels, permet de mettre en évidence notamment des individus qui jouent un rôle majeur dans la conversation et des tweets qui ont contribué à la dynamique des échanges. Ce prototype a été testé sur un panel d'utilisateurs, pour valider son apport et ouvrir des perspectives d'amélioration.

Mots clés

Contextualisation, Tweets, Conversations, Contexte, Twitter.

Abstract

Even with the recent switch to 280 characters, Twitter messages considered in their singularity, without any additional exogenous information, can confront their readers with difficulties of interpretation. The integration of contextualization on these messages is therefore a promising avenue of research to facilitate access to their information content. In the last decade, most works have focused on building summaries from complementary sources of information such as Wikipedia. In this thesis, we choose a different complementary path that relies on the analysis of conversations on Twitter in order to extract useful information for the contextualization of a tweet. These information were integrated in a prototype which, for a given tweet, offers a visualization of a subgraph of the conversation graph associated with the tweet. This subgraph, automatically extracted from the analysis of structural indicators distributions, allows to highlight particular individuals who play a major role in the conversation and tweets that have contributed to the dynamics of exchanges. This prototype was tested on a panel of users to validate its efficiency and open up prospects for improvement.

Key Words

Contextualization, Tweets, Conversations, Context, Twitter.