

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE  
Ministère de l'enseignement supérieur et de la recherche scientifique

Université des Sciences et des Technologies Houari Boumedienne



Faculté d'électronique et d'informatique  
Département Informatique

Mémoire de projet de fin d'études de Master

Master Ingénierie des Logiciels

Thème :

---

Recherche d'influenceurs dans le réseau social Twitter

---

Proposé et dirigé par : K. BOUKHALFA ET R. DJIROUNE

Présenté par : FERROUKHI KAHINA ET MADANY ZAKIA

Soutenu le : 22/06/2015

Devant le jury composé de :

*Président :* R. GUEBAILI

*Membre :* L. BERKANI

*Promoteur :* K. BOUKHALFA ET R. DJIROUNE

Numéro : 048/2015  
Promotion : 2014-2015

## Résumé

L'objectif de notre travail est la détection des influenceurs sur le réseau social Twitter. Notre travail consiste à étudier la notion d'influence dans un réseau social, en effet, l'influence n'est pas une chose facile à définir car cette dernière est liée au facteur temps.

Dans un premier temps, nous avons procédé à l'extraction des données du réseau social et cela en utilisant l'API du réseau. La récupération des données se fait principalement par le biais de mots-clés selon le domaine de recherche choisi. Les données récupérées sont principalement des publications (ou tweets). Une fois les tweets récupérés, ces derniers vont subir une opération d'élagage pour ne garder que les tweets pertinents avant d'être étiquetés.

Ainsi à partir des publications récupérées, nous procédons à l'extraction des utilisateurs posteurs et retweeters des tweets. Une fois les utilisateurs récupérés, ces derniers vont être filtrés avant d'être classés en catégorie.

Suite à cela, un score d'influence est calculé pour chaque utilisateur et cela suivant deux critères l'impact et la popularité. Le score d'influence est calculé en utilisant la fonction de score « Z-Score ».

Notre outil de détection d'influenceurs sur Twitter a été développé en utilisant le langage Python, ainsi qu'une base de données NoSQL orientée graphe : Neo4j.

# Sommaire

<b>Introduction générale</b>	<b>8</b>
Problématique . . . . .	8
Objectif du mémoire . . . . .	9
Structure du mémoire . . . . .	9
 <b>I État de l’art</b>	 <b>10</b>
<b>I Médias Sociaux</b>	<b>11</b>
I.1 Introduction . . . . .	11
I.2 Définition des médias sociaux . . . . .	11
I.3 Les types des médias sociaux . . . . .	12
I.3.1 Les médias de discussion . . . . .	12
I.3.2 Les médias de publication . . . . .	12
I.3.3 Les réseaux sociaux numériques de contact . . . . .	12
I.3.4 Les réseaux sociaux numériques de contenu . . . . .	13
I.4 Réseaux sociaux . . . . .	13
I.5 Classification des réseaux sociaux . . . . .	13
I.5.1 Les réseaux sociaux offline . . . . .	14
I.5.2 Les réseaux sociaux online . . . . .	14
I.6 Évolution des réseaux sociaux . . . . .	14
I.7 La structure sociale . . . . .	15
I.8 Les types de relations sociales . . . . .	15
I.8.1 Les relations symétriques . . . . .	15
I.8.2 Les relations asymétriques . . . . .	16
I.9 Caractéristiques des réseaux sociaux . . . . .	17
I.9.1 Définition de la technologie NoSQL . . . . .	17
I.9.2 Types de bases de données NoSQL . . . . .	17
I.10 Conclusion . . . . .	19

<b>II L'influence dans les réseaux sociaux</b>	<b>20</b>
II.1 Introduction . . . . .	20
II.2 La diffusion de l'information . . . . .	20
II.3 Les influenceurs et les leaders d'opinion . . . . .	21
II.4 Définition de l'influence . . . . .	21
II.5 L'influence sur Internet . . . . .	21
II.5.1 Quels-sont les profils influenceurs ? . . . . .	22
II.6 Le marketing sur Internet . . . . .	23
II.6.1 Définition du marketing viral . . . . .	23
II.6.2 Pourquoi choisir le marketing viral ? . . . . .	24
II.7 Limites et inconvénients du marketing viral . . . . .	24
II.8 Conclusion . . . . .	25
<b>III Détection des influenceurs dans les réseaux sociaux</b>	<b>26</b>
III.1 Introduction . . . . .	26
III.2 La recherche d'information . . . . .	26
III.2.1 La recherche d'information sociale (RIS) . . . . .	27
III.3 Le Réseau social : Twitter . . . . .	28
III.3.1 Fonctionnalités de Twitter . . . . .	29
III.3.2 Structure de données d'un utilisateur Twitter . . . . .	29
III.4 La recherche d'information sociale sur Twitter . . . . .	31
III.5 L'influence sur Twitter . . . . .	32
III.6 Les mesures d'influence sur Twitter . . . . .	32
III.7 L'estimation de l'influence sur Twitter . . . . .	33
III.8 Calcul du score d'influence . . . . .	34
III.8.1 La fonction de score Z-Score . . . . .	34
III.8.2 Paramètres statistiques utilisée pour le Z-Score . . . . .	34
III.8.3 Les limites du Z-Score . . . . .	35
III.9 Conclusion . . . . .	35
<b>II Notre contribution</b>	<b>36</b>
<b>I Conception de l'approche</b>	<b>37</b>
I.1 Introduction . . . . .	37
I.2 Méta-modèle d'un réseau social . . . . .	38
I.3 Description de l'approche . . . . .	39
I.3.1 Étape 1 : Préparation du corpus . . . . .	40
I.3.2 Étape 2 : Détection des l'influenceurs . . . . .	45
I.4 Conclusion . . . . .	49

<b>II</b>	<b>Modélisation de l'approche</b>	<b>50</b>
II.1	Introduction . . . . .	50
II.2	Diagramme de cas d'utilisation . . . . .	50
II.3	Diagramme de séquence . . . . .	51
II.3.1	Digramme de séquence d'authentification . . . . .	52
II.3.2	Diagramme de séquence de préparation du corpus . . . . .	52
II.3.3	Diagramme de séquence de consultation des entités . . . . .	53
II.3.4	Diagramme de séquence de détection des influenceurs . . . . .	54
II.3.5	Diagramme de séquence d'incrémentation du corpus . . . . .	54
II.3.6	Diagrammes d'activité . . . . .	55
II.3.7	Diagramme de classes . . . . .	56
II.4	Conclusion . . . . .	57
<b>III</b>	<b>Implémentation de l'outil</b>	<b>58</b>
III.1	Introduction . . . . .	58
III.2	Outils de développement . . . . .	58
III.2.1	Langage de programmation . . . . .	59
III.2.2	Environnements de développement . . . . .	59
III.2.3	Les API utilisés . . . . .	60
III.2.4	Base de données orientée graphes Neo4j . . . . .	62
III.2.5	Stockage et modèle physique de données . . . . .	63
III.3	Réalisation de notre outil de détection des influenceurs . . . . .	67
III.3.1	La partie construction du corpus . . . . .	67
III.3.2	La partie détection des influenceurs . . . . .	68
III.4	Présentation de la base de données graphe . . . . .	69
III.5	Conclusion . . . . .	70
	<b>Conclusion générale</b>	<b>71</b>
	Perspectives . . . . .	72

# Liste des tableaux

III.1	Le nombre d'utilisateurs actifs sur les réseaux sociaux [6]. . . . .	27
III.2	Exemples des données exploitables dans les réseaux sociaux [BADACHE, 2012]. . . . .	28
III.3	Fonctionnalités de Twitter [PEPIN, ]. . . . .	30
III.4	Les opérateurs de recherche sur Twitter. . . . .	32
I.1	Calcul du ratio(outdegree/indegree) . . . . .	47
II.1	Description des cas d'utilisation . . . . .	52

# Table des figures

I.1	Timeline du développement des réseaux sociaux . . . . .	14
I.2	Représentation des relations symétriques. . . . .	16
I.3	Représentation des relations asymétriques . . . . .	16
I.4	Représentation des bases de données NoSQL. . . . .	19
II.1	La diffusion de l'information [E. and P., 1955]. . . . .	21
II.2	La pyramide de l'influence [Augure, ]. . . . .	22
III.1	Icône d'un compte certifié sur Twitter . . . . .	30
III.2	Relations entre utilisateurs du réseau Twitter. . . . .	31
I.1	Méta-modèle du réseau social Twitter. . . . .	39
I.2	Schématisation de notre approche . . . . .	40
I.3	Préparation du Corpus . . . . .	41
I.4	Sauvegarde des entités dans la base de données . . . . .	45
I.5	Étape de détection des influenceurs . . . . .	46
II.1	Diagramme de cas d'utilisation. . . . .	51
II.2	Diagramme de séquence authentification. . . . .	53
II.3	Diagramme de séquence de préparation du corpus . . . . .	53
II.4	Diagramme de séquence de consultation des entités. . . . .	54
II.5	Diagramme de séquence détection des influenceurs. . . . .	54
II.6	Diagramme de séquence d'incrémentation du corpus. . . . .	55
II.7	Diagramme d'activité. . . . .	56
II.8	Diagramme de classe de notre outil. . . . .	57
III.1	Les étapes d'authentification sur Twitter API. . . . .	61
III.2	Le modèle physique de données. . . . .	64
III.3	Nouveau utilisateur posteur, nouveau utilisateur retweeteur. . . . .	65
III.4	Nouveau tweet, nouveau utilisateur posteur, nouveau retweeteur. . . . .	65
III.5	Nouveau Follower de l'utilisateur retweeteur. . . . .	66
III.6	Nouveau Follower de l'utilisateur posteur. . . . .	66

---

III.7 Follower du posteur existe, Follower du retweeter n'existe pas. . . . .	66
III.8 Architecture globale de notre application. . . . .	67
III.9 Interface d'authentification. . . . .	68
III.10 Interface de recherche et récupération des données du réseau social. . . . .	68
III.11 Réglage paramètres d'influence. . . . .	69
III.12 Affichage des résultats de calcul de score. . . . .	69
III.13 Aperçu de la base de données Neo4j. . . . .	70



# Introduction générale

## Problématique

De tout temps, l'entreprise est vue comme un système économique et social qui a pour objectif de fournir aux clients des services et/ou produits susceptibles de les intéresser. Cependant, avec le grand nombre d'entreprises qui existent aujourd'hui, être toujours compétitif est devenu de plus en plus difficile.

En effet, dans un marché concurrentiel, il est primordial pour une entreprise de se distinguer des concurrents et ceci en ayant recours aux nouvelles stratégies de communication et de marketing. De ce fait, les entreprises sont en permanence à la recherche d'une meilleure efficacité économique et de produits novateurs capables de maintenir et d'augmenter leurs parts de marché.

Il y a quelques années, les campagnes marketing ciblaient essentiellement les médias traditionnels à savoir : la télévision, la presse et la radio. Cependant, depuis l'apparition d'Internet les campagnes marketing ont vu un nouveau jour avec notamment les techniques de mailings lists<sup>1</sup>. Aujourd'hui les médias sociaux, les médias traditionnels et le marketing sont en pleine remise en question.

Les médias sociaux sont de nouvelles technologies Web accessibles gratuitement et qui permettent de rassembler un grand nombre de personnes dans le but de communiquer en ligne, partager et commenter du contenu en temps réel.

De nos jours, toute organisation ou entreprise vit désormais sous la surveillance des médias, les médias sociaux et plus précisément les réseaux sociaux qui aident énormément les professionnels du marketing à promouvoir un produit ou une marque. Pour cela, il suffit qu'un simple texte se propage sur ces réseaux pour atteindre les clients et les encourager à acheter tel ou tel produit. C'est cette forme de publicité subtile qui est la plus apte à influencer un potentiel client.[19]

L'influence d'un individu réside dans sa capacité de persuader et inciter d'autres individus à adopter certains comportements. Cependant certains individus interagissent plus que d'autres et/ou disposent d'un statut différent (célébrité, expertise, connaissance personnelle, etc.) qui leur permet d'exercer un pouvoir d'influence potentiel sur leur entourage. Ces derniers sont considérés comme des « leaders d'opinion » agissant comme des relais (ou connecteurs) pour la diffusion de l'information, leur impact dans les réseaux sociaux prend une place de plus en plus importante puisqu'ils ont une réelle force de conviction.

Par ailleurs, au niveau des réseaux sociaux, leur périmètre d'influence est exponentiel. Celui-ci n'est plus limité à une sphère privée, ni géographique. D'où l'intérêt pour les services marketing d'identifier ces individus qui sont de réels influenceurs.

Dans cette perspective notre étude vise à aider les entreprises à sélectionner les meilleures mesures pour identifier les potentiels influenceurs et activer les bons leviers d'influence, afin de livrer une offre qui correspond parfaitement aux attentes des clients.

---

1. mailing list, ou liste de diffusion en français, est une utilisation spécifique du courrier électronique qui permet de diffuser des informations aux utilisateurs qui y sont inscrits.

## Objectif du mémoire

L'objectif de notre travail est axé sur l'influence dans le réseau social Twitter et plus particulièrement la recherche des influenceurs et cela dans le but de découvrir comment ces derniers aide les entreprises dans leurs campagnes marketing. Dans le cadre de notre travail, nous proposons une approche de recherche d'influenceur en temps réel et cela selon un domaine d'activité. Nous procédons à la recherche et à la récupération des publications selon certains mots-clés. Après cela, nous passons à l'évaluation des utilisateurs qui ont publiés et rediffusés ces publications ainsi que les relations entres eux. Nous calculons alors un score d'influence pour chaque utilisateur. Le score d'influence sera calculé selon deux critères : la popularité et l'impact de l'utilisateur dans le réseau. Ces derniers garantissent une grande visibilité aux entreprises en leur procurant une image de marque durable dans le temps.

## Structure du mémoire

Le présent mémoire est organisé en deux (02) parties : état de l'art et contribution.

La première partie se compose de 3 chapitres. Dans le premier chapitre, nous présentons la nouvelle vague technologique des médias sociaux ainsi que les réseaux sociaux. Dans le second chapitre, nous présenterons les concepts d'influence et de marketing sur Internet. Et dans le troisième chapitre, nous parlerons de la détection d'influence dans les réseaux sociaux et plus particulièrement sur le réseau Twitter.

La deuxième partie présente notre contribution. Nous présenterons dans cette partie notre approche de détection d'influenceurs suivi de sa modélisation. Ensuite pour la mise en oeuvre de notre approche, nous présenterons notre outil en détaillant les choix techniques pour sa réalisation.

Et pour finir, nous clôturons notre mémoire par une conclusion générale de notre travail suivi des perspectives.

Première partie

État de l'art

# Chapitre I

## Médias Sociaux

### I.1 Introduction

La vulgarisation d'Internet et l'apparition des nouvelles technologies Web ont complètement changé les comportements de l'homme, que se soit la simple action de faire des achats, en passant par la gestion des comptes bancaires jusqu'au mode d'interaction sociale.

En effet, grâce à cette nouvelle vague de médias sociaux, les individus peuvent échanger photos et vidéos, partager des reportages, s'exprimer sur des sites et des wikis<sup>1</sup> et participer à des discussions en lignes sur des forums. Ces médias permettent aux particuliers, aux entreprises, aux organisations et aux gouvernements d'interagir avec un grand nombre de personnes.

De nos jours, les médias sociaux ont pris une part importante dans la vie quotidienne de plusieurs individus habitués à vivre dans un environnement social afin de partager des avis du point de vue religieux, social et politique. Ces interactions entre utilisateurs ont fait déclencher des études et des recherches sur l'analyse de l'opinion et du comportement d'un utilisateur face à ces médias sociaux.

Dans ce premier chapitre nous allons définir les médias sociaux avec en vue leurs types, nous détaillerons par la suite les réseaux sociaux avec leurs caractéristiques et les types de relations sociales qui existent.

### I.2 Définition des médias sociaux

Selon [11] « l'expression médias sociaux recouvre les différentes activités qui intègrent la technologie, l'interaction sociale, et la création de contenu [...]. Par le biais de ces moyens de communication sociale, des individus ou des groupes d'individus qui collaborent créent ensemble du contenu Web, l'organisent, l'indexent, le modifient ou font des commentaires, le combinent avec des créations personnelles ».

D'après le Cambridge Advanced Learner's Dictionary, les médias sociaux sont des sites Web et des programmes informatiques qui permettent aux gens de communiquer et de partager l'information sur Internet en utilisant un ordinateur ou un téléphone mobile[1].

Depuis quelques années déjà, le terme « médias sociaux » apparaît fréquemment dans les articles et les blogs. Aujourd'hui, tout est devenu « social » et tout le monde est « média », du simple citoyen à la multinationale. Mais que signifie réellement ceci ?!

---

1. Un wiki est un site web dont les pages sont modifiables par les visiteurs

En examinant attentivement l'expression, on retrouve le mot « média » qui évoque alors les dernières technologies du Web utilisées librement pour créer, indexer, organiser, commenter ou modifier du contenu par les internautes. Et par « social », on entend alors toutes les interactions sociales, réactions, influence entre des individus ou groupe d'individus, liées à un contenu[France, 2010].

On peut apporter une définition relativement simple et claire. Le terme « médias sociaux » désigne un large éventail de services Internet et mobiles qui permet aux utilisateurs de [France, 2010] :

- Participer à des échanges en ligne.
- Diffuser du contenu.
- Rejoindre des communautés virtuelles.

### I.3 Les types des médias sociaux

La diversité des médias sociaux réside dans le fait qu'ils utilisent différentes techniques et plates-formes, telles que les flux RSS<sup>2</sup>, les blogs<sup>3</sup>, les wikis, les plate-formes de partages de contenus (vidéos, photos, texte,..) et les réseaux sociaux.

De ce fait, nous pouvons identifier 4 grandes catégories de médias sociaux :

#### I.3.1 Les médias de discussion

Les outils de discussions sont des outils de conversation en ligne, ils sont utilisés pour discuter à distance à la fois dans le cadre de la vie privée tout comme celui de la vie professionnelle, disponible sur la plupart des plate-formes(ordinateur, mobile, tablette, .. etc). Ces derniers utilisent des techniques tel que la messagerie instantanée et les systèmes de VoIP<sup>4</sup>. Parmi les logiciels existants, nous citons : Skype[14] qui a été racheté par Microsoft en 2011, ainsi que les applications Viber[14] et WhatsApp[14][Dewing, 2013].

#### I.3.2 Les médias de publication

Il existe différents types d'outils de publication, nous citons les plate-formes de blog et les plate-formes de wikis qui utilisent les techniques de flux RSS pour la transmission de nouvelles publications. Le site Wikipédia est l'outil de publication le plus visité avec plus de 20 millions de visiteurs par mois[15].

#### I.3.3 Les réseaux sociaux numériques de contact

On distingue deux types de réseaux sociaux numériques de contact [Dewing, 2013] :

**Les réseaux grand public :** Ils correspondent aux sites comme Facebook<sup>5</sup> où les utilisateurs entretiennent des relations avec d'autres utilisateurs et ceci afin de partager des photos, des vidéos et des expériences personnelles tout en gardant contact avec les amis et les membres de la famille.

---

2. Les flux RSS sont des fichiers dont l'objectif est de stocker une liste de contenus ou de pages web.

3. Un blog est un type de site Web sur lequel un internaute tient une chronique personnelle ou consacrée à un sujet particulier.

4. Voix sur IP ou la téléphonie sur IP

5. <http://www.facebook.com/>

**Les réseaux professionnels :** Sont des sites tels que LinkedIn<sup>6</sup> ou Viadeo<sup>7</sup> qui offrent aux utilisateurs la possibilité de créer un réseau de contact professionnel et d'avoir une réputation. Les entreprises utilisent aussi ce genre de réseaux pour rentrer en contact avec les futurs employés, qui peuvent déposer leurs CV en ligne et choisir les contacts qui sont dans le même secteur d'activité.

### I.3.4 Les réseaux sociaux numériques de contenu

On parle de réseaux sociaux numériques de contenu car ces derniers offrent la possibilité aux utilisateurs de partager et de consulter des contenus vidéos (Youtube, Dailymotion, Vimeo,...), musiques (Deezer, Soundcloud,...), photos (Flickr, Instagram, Pinterest,...).

Ces catégories se chevauchent dans une certaine mesure. Par exemple, Facebook est un réseau social numérique de contact avec des outils de discussions intégrés [Dewing, 2013].

## I.4 Réseaux sociaux

En 1954, l'anthropologue américain John A. Barnes a défini pour la première fois l'appellation « réseau social », selon lui, un réseau social est l'ensemble des interactions sociales qui unissent un groupe d'individus. Chaque individu peut faire valoir plusieurs réseaux sociaux : amical, familial, professionnel ou d'intérêt spécifique (sportif, culturel...). Ceux-ci peuvent se regrouper ou s'imbriquer [F. Filliettaz, 2011].

D'autres auteurs, définissent un réseau social comme un ensemble d'acteurs qui partagent des relations entre eux. Les acteurs peuvent être de simples individus mais aussi des organisations et des institutions [Hanneman and Riddle, 2005].

Une autre définition décrit un réseau social comme une structure sociale dont les composants sont des identités sociales telles que des individus ou des organisations représentées par des noeuds. Ces identités sont liées entre elles ou connectées à travers une ou plusieurs relations différentes d'interdépendance, représentées par des arrêts/arcs, créés lors des interactions sociales comme l'amitié ou pour un intérêt commercial, religieux, politique [G. ERETEO, 2009].

Selon une étude publiée sur le Journal of Computer-Mediated Communication, les réseaux sociaux sont définis comment étant l'ensemble des services du Web qui permettent aux individus de construire un profil public ou semi-public dans une communauté virtuelle, articulé d'une liste d'autres utilisateurs avec lesquels ils partagent une connexion, la nature de cette connexion varie d'une communauté à une autre.(relation d'amitié sur Facebook, relation d'abonnements/abonnés sur Twitter,...)[Ellison, 2008].

Ainsi, nous remarquons à travers ces définitions que le concept de réseau social change d'un milieu à un autre, il existe donc deux types de réseaux sociaux. Les réseaux sociaux offline, dits traditionnels, et les réseaux sociaux online ou numériques, que nous présenterons dans ce qui suit.

## I.5 Classification des réseaux sociaux

Les réseaux sociaux ont été médiatisés, ces dernières années, en raison de l'essor de leur forme numérique sur le Web. Néanmoins la structure sociale que représente un réseau social existe sous une forme élaborée,

---

6. <https://www.linkedin.com/>

7. <http://dz.viadeo.com/fr/>

et cela depuis que l'homme interagit avec ses semblables.

### I.5.1 Les réseaux sociaux offline

Selon Pierre Mercklé, sociologue, un réseau social est un ensemble de relations entre un groupe d'acteurs, lui même organisé ou non. Ces relations peuvent être de nature différente et les acteurs sont principalement des individus, mais pas nécessairement[Mercklé, 2004].

### I.5.2 Les réseaux sociaux online

Un réseau social numérique ou online, est un ensemble de personnes réunies par un lien social, qui utilisent Internet pour leurs interactions sociales. Ce n'est que vers la fin des années 1990 que les réseaux sociaux sont apparus sur Internet, réunissant des personnes via des services d'échanges personnalisés (les plates-formes sociales).

## I.6 Évolution des réseaux sociaux

L'homme est fait pour vivre en communauté, il partage et communique avec son entourage. Avec l'apparition d'Internet et des réseaux sociaux, la communication avec des personnes à l'autre bout du globe est devenue beaucoup plus accessible, ainsi les internautes communiquent, collaborent, et partagent leur points de vues religieux, sociaux, politiques,etc.

En effet, depuis l'apparition des réseaux sociaux, la communication et les interactions sociales ont connu une nouvelle dimension, la communication virtuelle est devenue courante de nos jours. La figure I.1 représente un axe temporel du développement des réseaux sociaux.

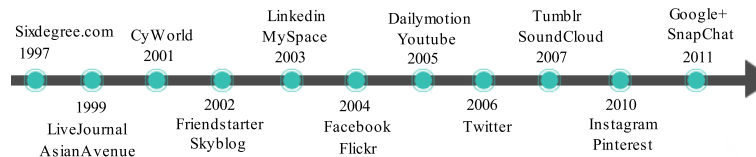


FIGURE I.1 – Timeline du développement des réseaux sociaux

Parmi les premiers réseaux sociaux apparus nous citons : Sixdegrees.com en 1997 qui permet aux utilisateurs la création de profils et la gestion des listes d'amis. Le service a échoué en 2001. Entre les années 1997 et 2001 plusieurs réseaux sociaux ont vu le jour tel que Classmates, qui aide ces utilisateurs à retrouver leurs anciens camarades d'école et d'université, mais aussi AsianAvenue et LiveJournal.

Les réseaux comme Friendster, Myspace et Facebook sont apparus entre 2002 et 2004. A son premier trimestre, Friendster avait dépassé les trois millions d'utilisateurs alors que Myspace était développé par une compagnie de marketing. En octobre 2005 Myspace est le quatrième site le plus consulté au monde derrière Yahoo!, AOL et MSN et devant eBay et Facebook, qui rappelons le à l'origine a été développé pour les étudiants de l'université de Harvard[Ellison, 2008].

Aujourd'hui Facebook compte, selon les résultats de 2014, environ 1,4 milliard d'utilisateurs actifs[7]. L'utilisation des réseaux sociaux est de plus en plus répandue puisque ces derniers offrent plus d'interactivité, de souplesse et d'accessibilité aux utilisateurs.



Ainsi au fil du développement des réseaux sociaux, leur modélisation a connue une évolution, cette dernière propose une schématisation flexible en accord avec la structure sociale.

## I.7 La structure sociale

Un réseau social est généralement représenté par un graphe. Les individus sont représentés sous forme de noeuds et les interactions entre individus sous forme d'arc/arrêt entre les noeuds. La structure sociale peut être définie par les unités qui constituent le graphe. Ces entités sont les individus, les communautés, ainsi que les relations entre eux.

**Les acteurs sociaux :** L'unité la plus importante au niveau des réseaux sociaux est l'acteur ou le déclencheur d'évènement et de partage. L'acteur peut être un ou plusieurs groupes et/ou communautés formés par les individus sociaux.

**Les liens sociaux :** L'utilisateur à tendance à avoir des relations d'amitié ou de partage d'avis avec les autres acteurs ces relations sont les liens sociaux qui peuvent être de plusieurs types. Ces derniers sont présentés dans la partie suivante.

## I.8 Les types de relations sociales

Il existe plusieurs types de relations sociales, dans ce qui suit, nous allons définir deux types de relations sociales : les relations symétriques et les relations asymétriques, ainsi que leurs représentations graphiques. En effet, la représentation visuelle du réseau se fait généralement à l'aide de graphes et permet de comprendre et de mettre en évidence les relations qui existent entre les membres du réseau.

### I.8.1 Les relations symétriques

Beaucoup de réseaux sociaux gèrent les relations symétriques qui traduisent la même considération des relations entre les utilisateurs. Un réseau social qui comprend ce type de relation permet aux utilisateurs de maintenir une liste d'amis et de créer ainsi des relations d'amitié. Facebook est un exemple typique de réseau social qui utilise ce type de relations.

Une relation d'amitié est de la forme suivante : A envoie une demande d'ajout à B, la relation d'amitié est instanciée seulement une fois que B accepte la demande de A.

Ce modèle est plus dédié à créer et à maintenir des relations personnelles avec des personnes de confiance et avec lesquels on est censé partager du contenu personnel.

### Représentation des relations symétriques

Les relations symétriques peuvent être représentées par un graphe non orienté  $G = (V, E)$ , où  $V$  représente les noeuds comme les utilisateurs et un ensemble d'arrêtes  $E \subset V \times V$  qui représentent la relations entre les noeuds. La figure I.2 illustre ce type de relation.

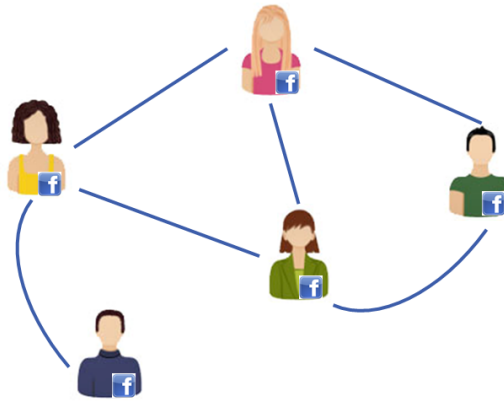


FIGURE I.2 – Représentation des relations symétriques.

### I.8.2 Les relations asymétriques

De nombreux réseaux sociaux gèrent les relations sociales qui relient deux utilisateurs avec deux angles différents en fonction de l'utilisateur. Ces relations utilisent le concept de "abonnés/abonnements" et sont généralement au coeur des plates-formes de microblogging. Par exemple Twitter et Instagram utilise ce type de relations.

Une relation asymétrique permet à un utilisateur de créer et de maintenir une liste de personnes en lui permettant de se souscrire à leurs flux d'informations. Ce modèle de réseau social est plus dédié à la diffusion de l'information que pour le partage mutuel de contenu.

On considère les relations entre deux utilisateurs A et B sur Twitter. A est souscrit au flux généré par B (A follow B) alors que B n'est pas obligé de le faire (A follow B, mais B n'est pas obligé de follower A)

#### Représentation des relations asymétriques

Les relations asymétriques peuvent être représentées par un graphe orienté  $G = (V, A)$ , où  $V$  représente les noeuds comme les utilisateurs et un ensemble d'arcs dirigés  $A \subset V \times V$  pour les relations entres les noeuds. La figure I.3 illustre ce type de relation.

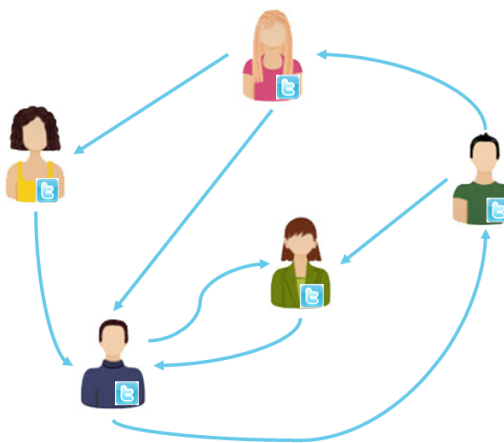


FIGURE I.3 – Représentation des relations asymétriques

## I.9 Caractéristiques des réseaux sociaux

Nous vivons depuis, de nos jours, une véritable évolution des usages d'Internet, ce qui a introduit un nouveau mode de communication et d'interaction sociale.

Il y a quelques années déjà, le contenu des sites Web était complètement statique, mis à jour uniquement par le webmaster, les internautes avaient le droit de lire le contenu seulement. Nous parlons donc d'un flux d'information unidirectionnel.

Aujourd'hui, les internautes ne sont plus seulement récepteurs, mais sont devenus relais et émetteurs. Ces consommateurs, autrefois simple cible marketing, peuvent désormais donner leurs avis, faire et défaire le succès d'un produit, recommander ou non une marque. Il s'agit donc d'un flux d'information bidirectionnel.

De ce fait, les réseaux sociaux se veulent de plus en plus interactifs. Ils sont devenus plus que de simples plates-formes de discussions, ils offrent d'autres outils de divertissement tels que les applications et les jeux. C'est cette interactivité et cet échange qui forme des communautés ou des groupes d'individus regroupés selon leurs intérêts communs.

Cependant, toutes ces interactions et ces échanges créent une énorme masse de données, ce qui rend nécessaire l'utilisation de nouveaux moyens technologiques pour faire persister ce grand volume de données ainsi que pour les interroger et les traiter. Pour cela, les bases de données NoSQL<sup>8</sup> sont une solution à ce problème.

### I.9.1 Définition de la technologie NoSQL

Le NoSQL désigne une catégorie de base de données apparues en 2009 qui se différencient du modèle relationnel que l'on trouve dans les bases de données traditionnelles. Les bases de données NoSQL sont fréquemment utilisées pour acquérir et stocker des structures de données dynamiques, variées et hautement scalables. Cette catégorie de produits fait le compromis d'abandonner certaines fonctionnalités classiques des SGBDs relationnels au profit de la simplicité, la performance et une forte scalabilité. La scalabilité est la capacité d'un système à répondre à une demande toujours croissante de la part des utilisateurs en termes de requêtes[HEINRICH, 2012].

Aujourd'hui le terme NoSQL englobe tous les SGBDs qui ne suivent pas la tendance de type relationnel. Cela signifie que le NoSQL n'est pas un seul produit ou même une technologie unique. En effet, il représente de nombreux produits ainsi que plusieurs concepts de stockage et de manipulation de données.

### I.9.2 Types de bases de données NoSQL

Le NoSQL regroupe 4 grandes familles de bases de données qui permettent d'offrir une représentation différentes des données, chacune dispose d'avantages et d'inconvénients en fonction du contexte dans lequel on souhaite l'utiliser. Nous citons [10] :

**Les bases de données clé-valeur :** La représentation en clé-valeur est la plus simple et est très adaptée aux accès rapides d'informations. Les données sont représentées par un couple clé/valeur.

---

8. NoSQL signifie "Not Only SQL", littéralement "pas seulement SQL".

La valeur peut être une simple chaîne de caractères. Parmi les SGBDs orientés clé-valeur, nous citons les plus connus, Riak<sup>9</sup> et Redis<sup>10</sup>.

**Les bases de données orientées colonnes :** La représentation orientée colonnes est celle qui se rapproche le plus des tables dans une base de données relationnelles. Elles permettent d'être beaucoup plus évolutive et flexible puisqu'on peut disposer de colonnes différentes pour chaque ligne. Nous pouvons citer les deux SGBDs orientés colonnes les plus connus, HBase<sup>11</sup> et Cassandra<sup>12</sup>.

**Les bases de données orientées documents :** Une base de données orientée documents est plus adaptée au monde de l'Internet. Sa représentation est très proche de la représentation clé-valeur à l'exception faite que la valeur est représentée sous la forme d'un document. On peut retrouver dans ce document les données organisées de manière hiérarchique comme ce que l'on trouve dans un fichier XML<sup>13</sup> ou JSON<sup>14</sup>. Nous pouvons citer les deux SGBDs orientés documents les plus connus, CouchDB<sup>15</sup> et MongoDB<sup>16</sup>.

**Les bases de données orientées graphes :** La représentation orienté graphe est utilisé pour palier à des problèmes impossibles à résoudre avec des bases de données relationnelles. Le cas d'utilisation typique est bien sur les réseaux sociaux où l'aspect graphe prend tout son sens, mais aussi où des relations complexes entre les acteurs ont besoin d'être décrites. Nous pouvons citer le SGBD orientés graphes le plus connu, Neo4j<sup>17</sup>. La figure I.4 illustre les différentes représentation des bases de données NoSQL.

---

9. <http://basho.com/riak/>

10. <http://redis.io/>

11. <http://hbase.apache.org/>

12. <http://cassandra.apache.org/>

13. Extensible Markup Language

14. JavaScript Object Notation

15. <http://couchdb.apache.org/>

16. <https://www.mongodb.org/>

17. <http://neo4j.com/>








Types de BD NoSQL	Exemples
Clé-valeur	 <b>redis</b>  <b>riak</b>
Colonnes	 <b>HBASE</b>  <b>cassandra</b>
Documents	 <b>CouchDB</b>  <b>mongoDB</b>
Graphes	 <b>Neo4j</b>

FIGURE I.4 – Représentation des bases de données NoSQL.

## I.10 Conclusion

Les médias sociaux et plus particulièrement les réseaux sociaux ont connus une croissance exponentielle ces dernières années, offrant aux utilisateurs la liberté de partager leurs avis et leurs opinions sur tel ou tel produit. Certains de ces utilisateurs ont la capacité d'influencer leurs entourages par de simples messages, ainsi en commentant et en partagent leurs avis, ces derniers laissent derrière eux une véritable mine d'or de données pour les analystes et les entreprises. De ce fait, les entreprises sont à la recherche permanente de ce type d'utilisateur, et cela dans le but d'améliorer leur image de marque dans un marché de plus en plus concurrentiel. Dans le chapitre suivant nous parlerons de l'influence dans les réseaux sociaux.

## Chapitre II

# L'influence dans les réseaux sociaux

### II.1 Introduction

L'une des problématiques des entreprises qui investissent dans les campagnes marketing sur les réseaux sociaux numériques est de déterminer une population cible qui a une capacité de diffusion et de médiatisation identifiable.

Le concept d'influence semble répondre aux caractéristiques souhaitées par les entreprises pour les actions marketing sur Internet. Seulement, il est difficile de détecter les personnes influentes appelés aussi leaders d'opinion.

Dans ce chapitre, nous présenterons la diffusion d'information qui se base sur le concept de leader d'opinion, nous définirons aussi les concepts d'influence et d'influenceur. Nous expliquerons par la suite le marketing viral et les raisons qui poussent les entreprises à utiliser ce genre de stratégie marketing pour promouvoir leurs produits.

### II.2 La diffusion de l'information

Dans les années 1940 et 1950, Paul Lazarsfeld et Elihu Katz ont formulé une théorie avancée sur l'opinion publique, nommée la théorie de la communication à double étage. Cette dernière remet en cause la théorie habituelle du "pouvoir" des médias. L'analyse porte tant sur l'influence des médias dans le cadre politique que dans le cadre marketing et personnel.

Ainsi l'information est diffusée en deux étapes, une petite minorité de "leaders d'opinion" (représentée sous forme d'étoile) agissent comme intermédiaire entre les médias et la majorité de la société (cercle). Ces leaders d'opinion sont les plus exposés aux médias, ce sont donc en grande partie eux qui filtrent, interprètent et transmettent les informations à leurs entourages [E. and P., 1955].

L'influence des médias sur l'ensemble de la population se fait donc en deux temps :

- Le message délivré par les médias est reçu et plus ou moins assimilé par un leader d'opinion.
- Le leader d'opinion fait partager son choix, son opinion à son entourage ainsi il influence leurs choix et leurs comportements.

De ce fait, l'expression « influence personnelle » a été inventée pour désigner ce processus de diffusion. Ainsi la figure II.1 schématise la théorie de la communication à double étage.

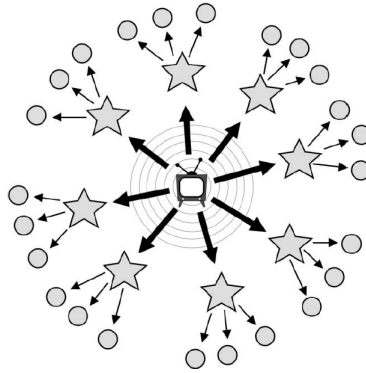


FIGURE II.1 – La diffusion de l'information [E. and P., 1955].

### II.3 Les influenceurs et les leaders d'opinion

Selon la définition retrouvée dans le glossaire en ligne du marketing, le leader d'opinion est un individu qui par sa notoriété, son expertise ou son activité sociale intensive est susceptible d'influencer les opinions ou actions d'un grand nombre d'individu[4].

L'étude [L., 2004] définit le leader d'opinion comme une personne qui exerce une force d'attraction (physique, psychologique et/ou sociale) sur son entourage et qui dispose d'une forte crédibilité dans une catégorie de produit. Ses jugements et comportements influencent les attitudes et les choix de marques de son entourage dans ce domaine. Leur trait le plus caractéristique est l'aura de confiance qui les entoure.

Cependant, un influenceur est un individu qui par son statut ou son exposition médiatique peut influencer les comportements de consommation dans un univers donné. La notion d'influenceur est surtout utilisée sur Internet car ce média est un vecteur d'influence pour de nombreux individus (blogueurs par exemple), on parle alors d'e-influenceur [3].

### II.4 Définition de l'influence

L'influence est un des mécanismes fondamentaux dont se préoccupe la psychologie sociale, elle montre à la fois l'emprise que la société exerce sur l'individu et les modifications qu'elle entraînent au niveau du comportement [Tarquinio, 2006].

Le terme d'influence désigne le processus par lequel une personne fait adopter une conduite ou un point de vue par une autre. L'influence sociale recouvre donc tout ce qui produit un changement de la conduite suite à une relation ou une interaction avec un individu dit influent ou influenceur [2].

Le processus d'influence est notamment à la base du leadership. Un individu a la capacité de persuader les autres individus à coopérer et collaborer à des objectifs sans utiliser de sanction ou de promesse, ce dernier est appelé leader ou leader d'opinion dans certains cas [2].

### II.5 L'influence sur Internet

Quand nous parlons d'influence sur Internet, nous utilisons le terme e-influence.

Ainsi l'e-influence est le pouvoir qu'exerce une entreprise, une personne ou une organisation, d'une manière continue sur d'autres entités, personnes ou groupement social en utilisant les technologies du Web.

Par ailleurs le caractère évolutif d'Internet n'aide pas les spécialistes en marketing car chaque utilisateur est un personnage social à part entière. Bien cerner tout les utilisateurs et engager une discussion avec eux n'est pas une mince affaire. C'est pour cela qu'il est nécessaire aux entreprises de bien déterminer quels sont les influenceurs clés pour leurs marques.

### II.5.1 Quels-sont les profils influenceurs ?

De nombreux services sont disponibles sur Internet pour calculer l'activité d'un utilisateur dans les réseaux sociaux, à leur tête le site et application mobile Klout <sup>1</sup>.

L'outil Klout propose une matrice [9] regroupant les différents types d'influenceurs sur les réseaux sociaux, du spécialiste en passant par l'activiste jusqu'au socialisateur et l'observateur.

Il existe en effet une grande variété de profils et de comportements lorsqu'il s'agit d'influence sur les réseaux sociaux. Nous vous proposons ci-dessous, dans la figure II.2, une vision simplifiée de la pyramide de l'influence. Dans ce qui suit, nous allons détailler chaque palier de la pyramide [Augure, ].

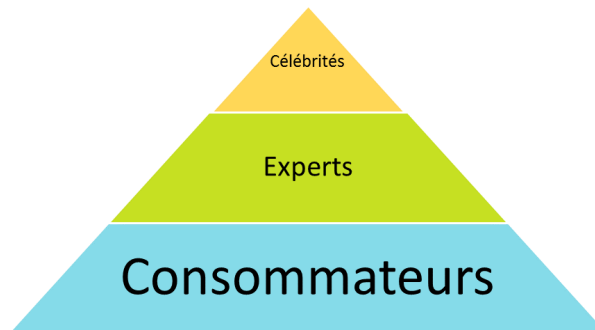


FIGURE II.2 – La pyramide de l'influence [Augure, ].

#### Catégorie : Célébrité

Dans cette catégorie, nous retrouvons les célébrités et les égéries de marque. Ceux sont des utilisateurs ayant un compte certifié avec plus de 40 milles personnes dans leurs cercles d'amis.

Les comptes certifiés ont été introduits récemment dans les réseaux sociaux. Ils sont principalement utilisés pour empêcher l'usurpation d'identité des célébrités et des personnes connues. Ainsi on retrouve un logo spécifique sur chaque profil d'un compte utilisateur certifié [17].

#### Catégorie : Experts

Les utilisateurs de cette catégorie ont énormément d'individus dans leurs réseaux, la plupart sont des journalistes connectés, des blogueurs, ou même des dirigeant d'entreprise ou d'un mouvement associatif.

1. <https://klout.com/>



Ils sont crédibles vis-à-vis de leurs communautés et sur des thématiques bien précises. Cependant, malgré le grand nombre représentant leurs communautés, leurs comptes ne sont pas certifiés par le réseau social.

### Catégorie : Consommateurs

Cette catégorie a pris beaucoup d'importance ces dernières années avec l'avènement des réseaux sociaux. Les consommateurs représentent le grand public. Ils peuvent facilement devenir de véritables ambassadeurs d'une marque ou, au contraire, des détracteurs dont l'impact peut avoir un effet dévastateur.

En 2013, un passager mécontent de la compagnie aérienne British Airways a fait beaucoup parler de lui avec son tweet de mécontentement. Ce dernier a été repris dans de nombreux médias et a même été interviewé par la chaîne CNN [20].

Ainsi l'influence est une notion très délicate à définir et à détecter. Néanmoins la véritable question pour les spécialistes du marketing et les entreprises est de savoir ce qu'ils cherchent réellement en engageant une relation avec un influenceur. Cherchent-ils de la visibilité ? de la crédibilité ? ou à augmenter les ventes de leurs produits ?

## II.6 Le marketing sur Internet

Le marketing sur Internet couramment appelé marketing numérique ou marketing digital désigne l'ensemble des techniques marketing utilisées sur les supports et canaux digitaux [5].

Parmi les techniques utilisées par ce type de marketing, nous citons : l'email marketing<sup>2</sup>, la publicité display<sup>3</sup>, le marketing social et viral, etc.

En effet, depuis quelques années déjà, Internet est devenu le premier média en termes de décision d'achat devant la télévision et la radio [UNG, 2010]. Les consommateurs sont désormais des médias à part entière et il devient indispensable aux marques et aux entreprises de trouver un moyen pour interagir et entrer en contact avec ces derniers. Afin de mettre en valeur leurs nouveaux produits/offres, les marques ont à leur disposition une nouvelle forme de marketing appelée le marketing viral.

### II.6.1 Définition du marketing viral

Le marketing viral consiste à jouer sur la viralité d'un message c'est-à-dire sa capacité à se diffuser rapidement dans l'espace et dans le temps, d'un destinataire à un autre. Tout comme un virus, un message viral a comme origine une source<sup>4</sup>, le premier « contaminé », celui qui développe le virus, et des contaminés<sup>5</sup>[Jonathan, ].

Le marketing viral se diffuse principalement sur Internet et utilise les outils de communication interpersonnels pour faire passer le message de la marque à une large audience, rapidement et à moindre coût. En effet, le marketing viral est une stratégie qui permet aux entreprises de faire des économies financières. Utiliser le consommateur comme média permet de s'affranchir des nombreux coûts que la publicité engendre [Jonathan, ].

---

2. L'email marketing regroupe l'ensemble des utilisations de l'e-mail faites à des fins marketing.

3. Le terme de publicité display désigne les formes de publicité digitale utilisant des éléments graphiques ou vidéos

4. La source du message est généralement l'annonceur, la marque, ...etc.

5. Les contaminés sont les consommateurs, les auditeurs, ..etc.

Dans ce qui suit, nous citerons les principales raisons de l'utilisation du marketing viral pour promouvoir une marque ou un produit.

### II.6.2 Pourquoi choisir le marketing viral ?

Le marketing viral est de plus en plus utilisé ces derniers temps, c'est devenu un véritable phénomène, surtout depuis l'apparition des réseaux sociaux. Il pourrait se comparer au bouche à oreille traditionnel, c'est le consommateur qui vend au consommateur [16]. Voici les principales raisons de son utilisation dans les campagnes marketing [18].

1. **La simplicité** : effectivement, une campagne de marketing viral est simple à mettre en oeuvre. Pas besoin d'être expert, un simple texte sur les réseaux sociaux, en quelques clics et la campagne est mise en place.
2. **Un faible coût** : en effet, le marketing viral est peu coûteux, l'annonceur ne paye rien, vu que la diffusion est principalement faite sur Internet.
3. **Amélioration de la visibilité** : avec une campagne de marketing viral, on atteint rapidement un grand nombre de personnes. Par définition, la viralité est le fait que le message se propage très rapidement à beaucoup de personnes.
4. **Amélioration de la notoriété** : le marketing viral est une sorte de bouche à oreille moderne qui permet de faire connaître et d'améliorer l'image de la marque.
5. **La croissance exponentielle** : avec le marketing viral on atteint rapidement les personnes intéressées par notre produit, tout en touchant beaucoup de monde. La taille de la cible grandit ne se limite plus au clients de base mais elle est multipliée par X à chaque transfert du message. Dans ce cas, l'action ne se limite plus à la fidélisation mais à l'obtention de nouveaux clients.
6. **Communiquer différemment** : le marketing viral est un type de communication plus proche du consommateur ainsi suite à une campagne de marketing viral, la marque peut savoir ce que pensent les receveurs du message, de ce fait, les consommateurs participent aussi à cette campagne en donnant leurs avis et leurs choix d'achats.
7. **Un retour sur investissement positif** : une raison qui n'est pas négligeable, le marketing viral peut vraiment augmenter le chiffre d'affaire de l'entreprise.

### II.7 Limites et inconvénients du marketing viral

Le marketing viral est un outil de communication certes puissant, mais à utiliser avec précaution. En effet, ce dernier peut avoir des inconvénients quand il est mal utilisé, parmi les échecs d'une campagne marketing viral, nous citons [Fatmi, 2010] :

1. **La lassitude** : à force de voir une publicité revenir constamment, le consommateur se lasse et ceci peut avoir pour effet de ternir l'image de la marque. En effet, il va partager ses mauvaises impressions et son dérangement avec son entourage.
2. **Détournement/modification** : toucher beaucoup de monde a aussi ses inconvénients. Le consommateur peut, de manière volontaire ou non, modifier le contenu du message. Ceci pourrait avoir un

effet positif, mais dans la plupart des situation, le détournement se fait de manière négative et a un effet désastreux pour l'image de la marque.

3. **Généralisation :** le marketing viral est de plus en plus utilisé, et ne surprend plus le consommateur. En effet, le consommateur connaît pertinemment les objectifs publicitaires ce qui le pousse à passer par-dessus et à l'ignorer ce type de message. Ce qui engendre une perte d'efficacité de ce système marketing.

## II.8 Conclusion

Dans ce chapitre nous avons expliqué les concepts d'influence, d'influenceur et de leader d'opinion. Nous avons fait aussi un petit tour d'horizon sur le marketing viral et l'importance de son utilisation pour les entreprises. Cependant, il est indispensable de trouver les bons relais et les influenceurs qui permettent de relayer et de propager l'information auprès d'autres utilisateurs. Ainsi nous présenterons, dans le prochain chapitre, la détection de l'influence dans les réseaux sociaux.

## Chapitre III

# Détection des influenceurs dans les réseaux sociaux

### III.1 Introduction

Nous décrivons dans le présent chapitre les principaux concepts traitant la thématique de notre projet à savoir la détection de l'influence sur les réseaux sociaux qui traite aussi la recherche d'information sociale. Nous commençons par présenter le principe de la recherche d'information en enchaînant par son existence dans les réseaux sociaux. Par la suite nous parlerons du fond de notre problématique, qui est « la détection d'influenceurs dans les réseaux sociaux ». Pour cela, notre étude se basera sur le réseau social Twitter, que nous étudierons avec ces différentes fonctionnalités et ces structures de données, et ceci après avoir passé en revue le concept de recherche d'information sociale.

### III.2 La recherche d'information

Le développement d'Internet et la généralisation de l'informatique ont conduit à la production d'un grand volume d'information. En effet, la quantité d'information disponible, particulièrement à travers le Web, se mesure en millions de téra et pétaoctets. Par conséquent, il est de plus en plus difficile de localiser précisément ce que l'on recherche dans cette masse d'information [Arezki, ].

La recherche d'information est le domaine par excellence qui s'intéresse à répondre à ce type d'attente. Son objectif principal est de fournir des modèles, des techniques et des outils pour stocker et organiser des masses d'informations et localiser celles qui seraient pertinentes relativement à la requête de l'utilisateur [Arezki, ].

D'une manière générale, les requêtes sont composées d'un ensemble de mots-clés. Ces mots-clés peuvent être reliés entre eux par des opérateurs booléens, comme ils peuvent être aussi organisés sous forme d'expressions.

La recherche d'information (RI) n'est pas un domaine récent, une des premières définition de la RI a été donnée par Gerard Salton, qui définit la RI comme un domaine qui consiste à acquérir, organiser, stocker, rechercher et sélectionner l'information et cela pour répondre aux besoins des utilisateurs. Ainsi la RI est utilisée dans plusieurs disciplines tel que : la classification/catégorisation, le filtrage d'information, la fouille de textes, .. etc [McGill, 1986].

Dans la partie suivante, nous nous basons sur la recherche d'information sociale.

### III.2.1 La recherche d'information sociale (RIS)

Avec l'explosion des technologies du Web, le rôle des utilisateurs d'Internet a été transformé de consommateurs passifs d'information à producteurs actifs comme expliqué précédemment [Ben Jabeur, 2013].

Des statistiques récentes montrent un taux de participation très fort des utilisateurs dans les réseaux sociaux ainsi qu'une quantité incroyable de contenu généré et publié quotidiennement. Selon [6], le nombre d'utilisateurs actifs sur les réseaux sociaux est estimé à 1,43 milliard en 2012. Ce nombre atteindra 1,85 milliard en 2014 [Ben Jabeur, 2013]. Le tableau III.1 représente le nombre des utilisateurs actifs sur les réseaux sociaux durant les 3 dernières années.

Année	2012	2013	2014
Nombre d'utilisateurs (milliard)	1.43	1.66	1.85

TABLE III.1 – Le nombre d'utilisateurs actifs sur les réseaux sociaux [6].

Les utilisateurs du Web produisent divers contenus, créent des annotations, des mentions, des commentaires, manipulent des documents sur le Web, etc. On observe dans les réseaux sociaux l'existence de l'information générée par l'utilisateur<sup>1</sup> qui peut être comme suit :

1. Les tags, utilisés pour annoter des bookmarks<sup>2</sup>, des pages Web, des images, etc. Ce type d'informations peut être considérées comme un avis de l'utilisateur, généralement positif, à propos des différentes ressources annotées. Il est aussi envisageable d'exploiter ces données pour en extraire des informations thématiques à propos des ressources annotées (ex : ressources liées à un domaine d'intérêt).
2. Les relations entre utilisateurs au sein du réseau social : amis, co-auteurs, followers, fans, etc.
3. Les profils des utilisateurs.

Le tableau III.2 représente un exemple de données exploitables et généré par les utilisateurs dans les différents réseaux sociaux actuels [BADACHE, 2012].

1. Le contenu généré par les utilisateurs (en anglais user-generated content, ou UGC)

2. Bookmark signifie marque page en anglais. Le bookmarking consiste à réunir et partager des liens sauvegardés.

UGC	Unité de mesure	Réseau social
J'aime Partager Commen- taire Mention	Nombre de « J'aime » Nombre de « Partage » Nombre de «Commentaire» Nombre de «Mentions»	Facebook
Tweet	Nombre de « Tweet »	Twitter
Partager	Nombre de « Partage »	LinkedIn, Pinterest
(J'aime +1)	Nombre de « J'aime +1 »	Google plus
Voter	Nombre de « Vote »	IMdb

TABLE III.2 – Exemples des données exploitables dans les réseaux sociaux [BADACHE, 2012].

D'une manière générale, ces données sociales peuvent être exploitées pour intégrer des propriétés sociales dans la RI (ex, à partir des tags, l'intérêt de la ressource est défini), mais aussi pour qualifier les ressources sur le Web par rapport à un thème donné suite aux retours positifs que donne l'utilisateur (ex, mesurer l'importance de la ressource). Cela a pour but de mieux trier les résultats des moteurs usuels de la recherche d'information.

Dans le cadre de notre travail, nous avons choisi d'exploiter le réseau « Twitter ». En effet, ce dernier est particulièrement bien adapté à la diffusion et à la propagation de l'information. Il offre un accès, relativement facile, à ses contenus, ce qui permet de connaître les sujets qui intéressent les utilisateurs ainsi que leurs réactions. Ainsi, son format de message court oblige les rédacteurs à adopter un style synthétiques tout en leur permettant d'inclure des liens vers les sources d'origine [PEPIN, ]. Dans ce qui suit, nous présenterons le réseau Twitter.

### III.3 Le Réseau social : Twitter

Twitter est un service de microblogage et un réseau social, il est à la fois un moyen de communication et un système de collaboration qui permet le partage et la diffusion des messages textuels. Il permet aux utilisateurs de communiquer des informations sur leurs statuts, activités, pensées et opinions [B., 2009].

En comparaison aux services de microblogage disponibles sur le Web, Twitter reste le site le plus populaire avec plus de 240 millions d'utilisateurs actifs. Vu ce succès, la quantité de données issues par Twitter a considérablement augmenté avec un taux qui excède les 504 millions de tweets par jour [13].

De ce fait, Twitter a été rapidement adopté par les blogueurs mais aussi par les stars, les personnes politiques, les journalistes ou encore les grandes marques qui l'utilise pour accompagner leur communication sur les nouveaux produits, les promotions du moment, les jeux concours, etc.

Cependant, les utilisateurs trouvent une difficulté pour accéder aux dernières actualités, masquées par l'énorme quantité des données et le flux soutenu des publications. Ainsi, le réseau Twitter dispose d'une API<sup>3</sup> qui permet de chercher et de télécharger des données du réseau.

3. API : Application Programming Interface

La forte popularité de Twitter et la facilité d'accès aux contenus textuels qui y sont publiés offrent d'énormes opportunités aux chercheurs en informatique, en sociologie, en traitement automatique de la langue. Cela explique le grand nombre d'études qui lui sont dédiées et les nombreuses méthodes envisagées pour analyser les tweets [PEPIN, ].

C'est pour cela, que nous avons choisi de centrer notre étude sur le réseau Twitter. La partie suivante présentera les différentes fonctionnalités du réseau ainsi que la structure de données d'un profil utilisateur.

### III.3.1 Fonctionnalités de Twitter

Les utilisateurs de Twitter publient des courts messages de 140 caractères appelés « tweets » visibles par tout le monde et qui sont envoyés directement à leurs abonnés appelés « followers ».

Les tweets contiennent principalement du texte mais peuvent également contenir des liens et des images, et peuvent être publiés en utilisant un large éventail, à partir d'un téléphone mobile, ordinateurs, ou à partir d'une API Twitter. Le retweet quant à lui est un type spécial de tweet c'est ce que l'on connaît sous le nom de « partage » [PEPIN, ]. Nous détaillons ces fonctionnalités dans le tableau III.3 suivant.

### III.3.2 Structure de données d'un utilisateur Twitter

Sur Twitter, les utilisateurs créent des profils pour interagir entre eux. De ce fait, les tweets d'un utilisateur peuvent être récupérés en utilisant des techniques de récupérations des API REST<sup>4</sup> et Stream<sup>5</sup>.

Ainsi, le profil d'un utilisateur est une source riche d'information. Les informations les plus importantes qu'on peut trouver sur un profil utilisateur sont les suivantes :

**Screen\_name :** C'est un alias qui identifie un utilisateur sur Twitter, qui ne doit pas dépasser 25 caractères, il est toujours suivi par un @ (@Screen\_name).

**Name :** Chaque utilisateur peut choisir un nom à afficher qui est différent du screen\_name. C'est le nom qui s'affiche aux autres utilisateurs pour identifier la personne. Il ne doit pas dépasser 20 caractères.

**Location :** C'est un attribut aidant à localiser la personne, ce dernier est spécifié par l'utilisateur.

**URL :** Un lien vers les autres réseaux sociaux ou le blog de l'utilisateur.

**Description du compte :** c'est la description du profil de l'utilisateur, c'est la où il écrit ses intérêts, sa profession,..etc.

**Informations sur le réseau de l'utilisateur :** On retrouve le nombre de followers et de following de l'utilisateur.

**Nombre de tweets :** C'est le nombre de publication d'un utilisateur.

**La date de création :** La date de création du profil de l'utilisateur.

**Compte certifié :** La certification d'un compte se fait après une étude d'une demande de certification auprès de Twitter, ainsi ces comptes sont validés par l'organisme en se basant sur le niveau de popularité de la personne, et cela dans le but d'empêcher l'usurpation d'identité. Ainsi, un utilisateur ayant un compte certifié gagne beaucoup plus rapidement la confiance des autres utilisateurs, ce

4. <https://dev.twitter.com/rest/public>

5. <https://dev.twitter.com/streaming/overview>

Fonctionnalité	Description
Tweet	Le tweet est un court message de 140 caractères publié par l'utilisateur pour partager ce qu'il fait en se moment ou tout simplement donner son avis sur un sujet bien précis.
Retweet	Le retweet, abrégé par RT, est le fait de relayer ou partager le tweet d'un autre utilisateur. Grâce à cette fonctionnalité, Twitter est devenu un puissant outil de bouche-à-oreille. Quand un utilisateur publie quelque chose d'intéressant et de pertinent, il y a de fortes chances que son message soit relayé.
Abonnements	Les abonnements, ou les followings sont les utilisateurs qu'on décide de "suivre". Ainsi nous recevrons tous leurs tweets sur notre timeline.
Abonnés	Les abonnés, ou les followers sont les utilisateurs qui nous suivent. Ainsi ils recevront nos tweets sur leur timeline.
Timeline	La timeline est la page principale sur laquelle apparaissent le fil d'actualité qui représente les tweets des abonnés.
Mentions	Les mentions sont sous la forme @identifiant ceci représente un lien vers un autre utilisateur, utilisés généralement pour interagir avec les abonnés ou avec les utilisateurs qui ont un profil public.
Hashtags	Un hashtag est constitué d'un symbole # suivi d'un mot-clé, généralement utilisé pour créer un espace de conversation, pour regrouper les tweets sur des sujets à la fois sérieux et humoristiques, allant des principaux événements mondiaux aux dernières actualités en passant par les offres d'emploi et le partage de musique.
Trending Topic	Les trending topics sont les sujets d'actualité. Ce sont des expressions ou des hashtags qui ont été retweetés plusieurs fois dans une période de temps. Il est possible d'afficher les trending topics d'un pays, d'une ville ou encore dans le monde entier.

TABLE III.3 – Fonctionnalités de Twitter [PEPIN, ].

qui lui permet d'influencer son entourage. Un compte certifié est reconnu grâce à une petite icône bleue située à côté du nom de l'utilisateur. La figure III.1 représente l'icône d'un compte certifié.



FIGURE III.1 – Icône d'un compte certifié sur Twitter

Sur Twitter il existe deux types de relations entre les utilisateurs, on trouve les « followers » et les « following ». Quand un utilisateur est suivi par un autre on dit que c'est son follower, et quand un utilisateur suit un autre utilisateur on dit qu'il est son following. La figure III.2 représente le type de relation entre utilisateurs de Twitter.



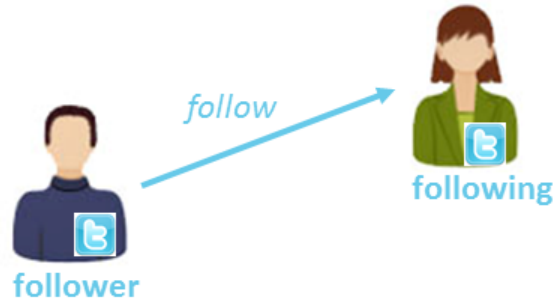


FIGURE III.2 – Relations entre utilisateurs du réseau Twitter.

Dans cette représentation, B qui est suivi par A donc on dit que A est un follower de B, et on a B qui suit C donc C est un following de B.

### III.4 La recherche d'information sociale sur Twitter

La recherche des tweets est une tâche de recherche d'information ad-hoc<sup>6</sup> dont l'objectif est de sélectionner les tweets pertinents en réponse à une requête Q [I., 2010].

Quand on parle de pertinence dans la recherche de tweets, on ne se limite pas à la similarité textuelle mais nous prenons également en compte les interactions sociales dans le réseau. De ce fait la pertinence des tweets dépend aussi de l'importance de l'utilisateur qui les publie [Boughanem, ].

En comparaison avec une recherche classique sur le web, la recherche de tweets permet d'obtenir une information brève, concise et précise sur un sujet actuel [B., 2009]. Elle permet aussi de recevoir en temps réel des informations sur un événement qui vient de se produire quelques secondes auparavant [Boughanem, ]. La recherche de tweets permet également d'accéder aux actualités avec une diversité de points de vue des utilisateurs et à une échéance proche de l'événement [Boughanem, ].

En utilisant l'API de Twitter la recherche des tweets devient plus facile. Selon le type de l'API utilisée, les résultats de recherche seront axés soit sur la pertinence soit sur l'exhaustivité. Ainsi, la recherche s'effectue par mots-clés, et cela suivant la requête de l'utilisateur. La requête peut contenir des opérateurs qui modifient les résultats. Dans le tableau III.4, nous expliquons les opérateurs les plus utilisés parmi les disponibles [12].

Cependant ce qui cause le plus de difficulté ce n'est pas la recherche des tweets, mais plutôt l'indexation en temps réel du flux de données récupérées, ainsi que la mesure de l'importance de l'utilisateur [Boughanem, ].

L'étude [Boughanem, ] associe la pertinence des tweets à l'importance sociale de l'utilisateur. Elle considère l'influence et l'expertise comme les principaux facteurs sociaux qui déterminent l'importance de l'utilisateur et la qualité de ses tweets. Dans la partie suivante nous allons évaluer la notion d'influence sur Twitter.

6. ad-hoc signifie « Qui a été instituée spécialement pour répondre à un besoin ».

Requête	Résultats de recherche
watching now	les tweets dans lesquels il y a le mot "watching" et "now". C'est l'opérateur par défaut.
"happy hour"	les tweets qui contiennent exactement l'expression "happy hour".
love OR hate	les tweets contenant soit "love" ou "hate" (ou les deux).
#Android	les tweets contenant le hashtag "Android".
@BillGates	tous les tweets de l'utilisateur "Bill Gates".

TABLE III.4 – Les opérateurs de recherche sur Twitter.

### III.5 L'influence sur Twitter

Sur Twitter, l'utilisateur interagit avec ces followers à travers leurs tweets, ainsi les utilisateurs peuvent s'échanger énormément d'informations, et en plus de la possibilité de rediffuser le message, un utilisateur peut commenter et donner son avis sur un sujet. Une étude [Gummadi, 2010] montre que les utilisateurs qui ont beaucoup de mentions, en plus d'un grand nombre de tweets rediffusés sont considérés comme de potentiels influenceurs. En plus d'avoir un large public d'audience, ces derniers ont la capacité d'engager les autres dans une conversation.

Ainsi, l'importance de l'utilisateur est déterminée par sa capacité d'affecter les autres utilisateurs et la proportion de ses tweets rediffusés [Boughanem, ] [Gummadi, 2010].

En effet, quand un autre utilisateur rediffuse ces tweets cela veut dire que ce dernier adopte la même idée si une opinion y est exprimée. La retransmission d'un tweet reflète donc l'importance du message communiqué.

D'après [Boughanem, ], l'influence d'un utilisateur dépend de ses relations de rediffusion et elle est estimée selon sa position dans le réseau. Ainsi, pour évaluer cette influence, une modélisation du réseau Twitter a été faite en se basant uniquement sur les relation de rediffusion. Le réseau social d'influence est modélisé par un graphe  $G = (U, E)$  où  $U$  est l'ensemble des utilisateurs et  $E = U \times U$  représente l'ensemble des relations d'influence entre eux.

Une relation d'influence  $e(u_i, u_j) \in E$  est définie par  $u_i \in U$  vers  $u_j \in U$  si et seulement s'il existe au moins un tweet publié par  $u_j$  et retweeté par  $u_i$ . Pour cela le poid  $w(u_i, u_j)$  de la relation d'influence est calculé par la formule III.1.

$$w(u_i, u_j) = \frac{\text{nb tweets publiés par } u_j \text{ et retweetés par } u_i}{\text{nb tweets retweetés par } u_i} \quad (\text{III.1})$$

Dans ce qui suit, nous allons présenter les mesures d'influence sur Twitter.

### III.6 Les mesures d'influence sur Twitter

La question de mesurer l'influence sur les réseaux sociaux se pose de manière récurrente, ces derniers temps, surtout pour les réseaux sociaux asymétriques tels que Twitter [VIGNOLLES, 2012].

Ainsi, sur les réseaux asymétriques, les mesures d'influence sont bien différentes que sur les autres réseaux. Nous retrouvons dans ce qui suit les mesures pour évaluer l'influence d'un utilisateur sur Twitter.

1. **IN/OUT Degree** : Sur un réseau asymétrique, comme Twitter, il est plus facile de repérer les individus présents dans un grand nombre de communautés, ceci reviendra à calculer la mesure de centralité (betweenness centrality). Ainsi, plus un individu a d'abonnés (followers = indegree), plus il pourrait être influent [VIGNOLLES, 2012].  
 Pourtant, l'étude [alan J.-Ph. Vignolles A., 2010] montrent que le nombre d'abonnés est significativement lié au nombre d'abonnements (following = out degree). En d'autres termes, même si le réseau n'est pas symétrique bon nombre de liens sont réciproques. Par conséquent, l'utilisation du nombre d'abonnés comme indicateur d'influence est peu pertinent.
2. **RETWEET** : Une fonctionnalité spécifique à Twitter et peut être considéré comme indicateur d'influence, en effet, un influenceur aurait son contenu plus fréquemment répété et rediffusé par la communauté [VIGNOLLES, 2012].
3. **Ratio(Abonnés/Abonnements)** : Il y a aussi le calcul du ratio du nombre d'abonnés par rapport au nombre d'abonnements comme indicateur d'influence. Les influenceurs sont définis comme ayant beaucoup d'abonnés et peu d'abonnements, soit un rapport abonnés/abonnements élevé [alan J.-Ph. Vignolles A., 2010].

Pour conclure, de nombreuses études ont été réalisés pour établir des indicateurs d'influence, mais hélas, l'estimation de l'influence sur les réseaux sociaux n'est pas une chose facile car cette notion n'est pas stable dans le temps. Cependant, de nombreux auteurs sont arrivés aux mêmes résultats et suggèrent que l'influence s'acquiert progressivement avec le temps et des efforts de la part de l'utilisateur [et Dodds P.S., 2007] [alan J.-Ph. Vignolles A., 2010].

### III.7 L'estimation de l'influence sur Twitter

Au niveau des réseaux sociaux, l'influence reste une notion difficilement quantifiable, car cette dernière dépend de plusieurs mesures, comme ceux citées précédemment, qu'il faut établir au préalable. Cependant chaque mesure quantifie un critère d'influence précis. Le but c'est de trouver un moyen pour réunir le plus de critères en exploitant les mesures d'influences existant sur le réseau social qui changent d'un réseau à un autre. Ainsi dans le cas du réseau Twitter nous pouvons estimer l'influence d'un individus à travers ses activité [Gummadi, 2010] :

1. **Indegree influence** : le nombre d'adeptes d'un utilisateur, indique directement la taille de l'auditoire pour cet utilisateur.
2. **Retweeter influence** : que nous mesurons par le nombre de retweets contenant son nom, indique la capacité de l'utilisateur de générer du contenu capable d'être rediffusé.
3. **Mention influence** : que nous mesurons par le nombre de mentions contenant son nom, indique la capacité de l'utilisateur d'engager les autres dans une conversation.

Cependant dans notre étude, nous nous sommes basées sur d'autres critères, citées plus loin dans ce mémoire, pour l'évaluation de l'influence d'un utilisateur sur ce réseau social. Par ailleurs, des fonctions

ont été utilisées et ceci dans le but d'établir des scores qui varient selon le degré d'influence d'un individu dans le réseau.

### III.8 Calcul du score d'influence

Dans cette partie nous nous intéressons à l'évaluation de l'influence, compte tenu que cette dernière reste une notion qualitative, son évaluation dans le monde réel reste subjective et pas facilement identifiable. Cependant la présence de la notion d'influence au niveau des réseaux sociaux est guidée par les fonctionnalités du réseau, ainsi un individu influenceur peut facilement être identifier par le biais de son comportement et du comportement de ses relations sociales. De ce fait, un individu influenceur peut être détecté par le biais de sa popularité ou de son impact dans le réseau. L'influence devient alors une notion quantifiable et un score d'influence peut être attribué à chaque individu en étudiant son comportement dans le réseau social.

Ainsi, pour évaluer l'influence dans le réseau social nous nous somme penchées à l'évaluation de deux critères d'influence : popularité et impact. A chaque individu sera affecté un score d'influence, ce score est calculé suivant ces deux critères mais aussi suivant l'ensemble des influenceurs existant dans le réseau social. De ce fait nous avons utilisé une fonction de score appelée Z-Score pour déterminer l'influence d'un individu par rapport aux autres membres du réseau. Cette fonction est expliquée plus en détails dans ce qui suit.

#### III.8.1 La fonction de score Z-Score

Dans la fin des année 60, le professeur Edward Altman créa un fonction de scoring (Z-Score) qui évalue la probabilité qu'une entreprise fasse faillite. Pour exprimer les résultats, on peut utiliser différents « langages » ou manière. Ainsi le Z-Score donne l'écart type relatif entre une valeur mesurée et une valeur cible. Le Z-score est une grandeur sans unité. Un signe négatif indique que la valeur mesurée est inférieure à la valeur cible, un signe positif indique que la valeur mesurée est supérieure à la valeur cible.

La valeur cible est estimée selon différentes méthodes [Wilder, 1977] :

- La valeur est connue : associée à un échantillon de données connu lors de la préparation.
- La valeur de référence est certifiée : elle est déterminée par des méthodes de références.
- La valeur est calculée : déterminée après mesure des données.
- Valeur de consensus : prise par consensus et validation d'entités expertes.

Ainsi le Z-Score est calculé par la formule III.2 suivante :

$$Z = \frac{x - V_{cible}}{SD} \quad (III.2)$$

Où  $x$  est le résultat de nos calculs,  $V_{cible}$  est la valeur cible, et  $SD$  l'écart-type ou déviation standard.

#### III.8.2 Paramètres statistiques utilisée pour le Z-Score

Dans le but de déterminer et de calculer la valeur cible ou l'écart-type utilisé précédemment, il est nécessaire d'utiliser des méthodes statistique lorsqu'on connaît a priori la distribution des valeurs. Une distribution normale est entièrement caractérisée par [CSCQ, 2007] :

**La moyenne :**

$$m = \frac{1}{n} \sum_{i=1}^n x_i \quad (\text{III.3})$$

ou  $x_i$  est le résultat de calcul des mesures et  $i$  et  $n$  sont le nombres d'utilisateurs dans le réseau étudié.

**L'écart-type SD :** Où Standard Deviation,

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2} \quad (\text{III.4})$$

**Le Coefficient de variation (CV) :** est déduit des grandeurs

$$CV = 100 * \frac{SD}{m} \quad (\text{III.5})$$

Le calcul de ces grandeurs à l'avantage d'être simple mais il est limité lorsque la distribution des mesures s'éloigne significativement de la distribution normale en particulier en présence de valeurs aberrantes.

### III.8.3 Les limites du Z-Score

Bien que le Z-score est censé être plus efficace qu'une analyse comparative des résultats, le Z-score n'en reste pas moins imparfait. Cet outil se base sur une base de données qui se doit être totalement complète. Si la qualité des informations n'est pas complète, le résultat du Z-score est faussé.

## III.9 Conclusion

Dans cette partie de notre mémoire, nous avons présenté le réseau Twitter qui représente le réseau social sur lequel notre approche a été définie. Nous avons ainsi défini les principales entités sur lesquelles notre étude se penche à savoir les utilisateurs et plus particulièrement les influenceurs. Dans la prochaine partie de notre mémoire, nous détaillerons l'approche de détection des utilisateurs influenceurs que nous proposons dans le cadre du présent travail.

Deuxième partie

Notre contribution

# Chapitre I

## Conception de l'approche

### I.1 Introduction

Selon [et Dodds P.S., 2007], la réussite d'une tendance ne dépend pas de la personne qui commence, mais de la façon dont la société est sensible à la tendance globale. En d'autres termes, une tendance peut être engagée par toute personne, et si l'environnement est bon, celle-ci se propage. Ce phénomène de tendance est principalement observé dans les réseaux sociaux, ainsi les adopteurs précoces<sup>1</sup> d'une tendance sont vus comme étant des individus influenceurs car ils arrivent facilement à provoquer un changement dans le comportement des autres personnes.

C'est pour cela que les entreprises cherchent à entrer en contact avec ce genre d'individus dans le but d'améliorer leur image de marque et de promouvoir de nouveaux produits. Cependant, trouver les bons individus influenceurs dans les réseaux sociaux n'est pas une chose facile étant donnée la vélocité avec laquelle les données sont générées quotidiennement sur ces réseaux, la recherche de personnes avec une grande influence et un fort impact devient donc difficile.

Ainsi, si nous devons schématiser ce problème : toute entreprise souhaitant commercialiser un nouveau produit P, doit trouver un sous ensemble d'individus, dit influenceurs. Ces derniers pourront contribuer à la propagation d'un contenu référençant le produit P dans le réseau afin de le promouvoir et de sensibiliser les futurs consommateurs.

Dans le cadre de notre travail, nous avons choisi de centrer notre approche sur le réseau Twitter car ce dernier est très utilisé pour les campagnes marketing. Nous entamons une présentation du méta-modèle qui regroupe les fonctionnalités du réseau étudié : Twitter. Puis nous enchaînerons avec une schématisation générale de notre approche que nous détaillerons partie par partie.

L'approche que nous proposons consiste à détecter en temps réel les individus influenceurs, à partir d'un corpus construit à l'aide de mots-clés appartenant à un domaine précis. Vu la taille gigantesque des données circulant sur Twitter, une première étape de construction du corpus est nécessaire. Ainsi, notre approche se compose essentiellement de deux étapes :

- Une première partie de préparation du corpus et ceci en appliquant la méthode de recherche d'information par mots-clés citée auparavant dans le but de construire notre corpus de base. Ce même corpus est analysé et une extraction d'autres entités est effectuée, ces dernières ont une relation

---

1. Adopteurs précoces : "early adopters", expression anglosaxonne utilisée en marketing pour désigner les individus les plus prompts à adopter une nouvelle technologie ou une innovation.

directe avec les entités du corpus de base.

- La deuxième étape est l'étape de traitement qui consiste en la détection des influenceurs. Pour cela, chaque entité utilisateur est évaluée selon deux critères d'influence, à savoir la popularité et l'impact. Un score d'influence est alors calculé pour chaque entité à l'aide de la fonction de « Z-Score ». Étant donné que le corpus d'étude est construit en temps réel, le score attribué pour chaque utilisateur est recalculé à chaque fois que le corpus subit une modification. Comme la fonction Z-Score évalue l'ensemble du corpus, la modification de ce dernier n'altère pas les résultats générés par la fonction ainsi elle est la mieux adaptée pour notre approche.

Notre travail consiste alors à détecter les individus influenceurs dans un réseau social et de calculer leur score d'influence dans un domaine donné dans le but de guider les entreprises dans le processus de recherche de potentiels ambassadeurs pour leurs images de marque dans les réseaux sociaux numériques.

## I.2 Méta-modèle d'un réseau social

Notre travail consiste à analyser les données issues d'un réseau social, qui est constitué principalement d'utilisateurs et de leurs activités (publication, statut, commentaire,...) et de relations entre utilisateurs (amitié, abonné, cercle,...).

Nous commençons par analyser les activités des utilisateurs et plus précisément leurs publications. Pour cela il est nécessaire d'avoir un aperçu global des entités et des fonctionnalités du réseau étudié « Twitter » avant de détailler notre approche. La figure I.1 représente le méta-modèle de Twitter.

Un utilisateur peut publier un message, l'action de publication d'un message est dite « tweeter » le message publié est appelé « tweet ». Ce dernier peut être un tweet simple, un retweet ou une mention. L'ensemble des publications d'un utilisateurs sont appelées « tweets » ou « posts ». Un tweet peut aussi contenir des ressources tel que : les URLs, les hashtags et les mentions.

Pour le retweet, quand une personne rediffuse un tweet on dit qu'elle a « retweeter » ou en utilisant l'acronyme « RT ». Dans le cas d'une réponse à un tweet (commentaire) on dit qu'elle a « mentionné ».

Un utilisateur peut signaler ou bloquer un autre utilisateur. Un utilisateur a des relations avec d'autres utilisateurs, il peut suivre et peut être suivi par d'autres utilisateurs, il a donc respectivement des « followings » et des « followers ». En l'occurrence nous allons plutôt employer le terme « relations ».



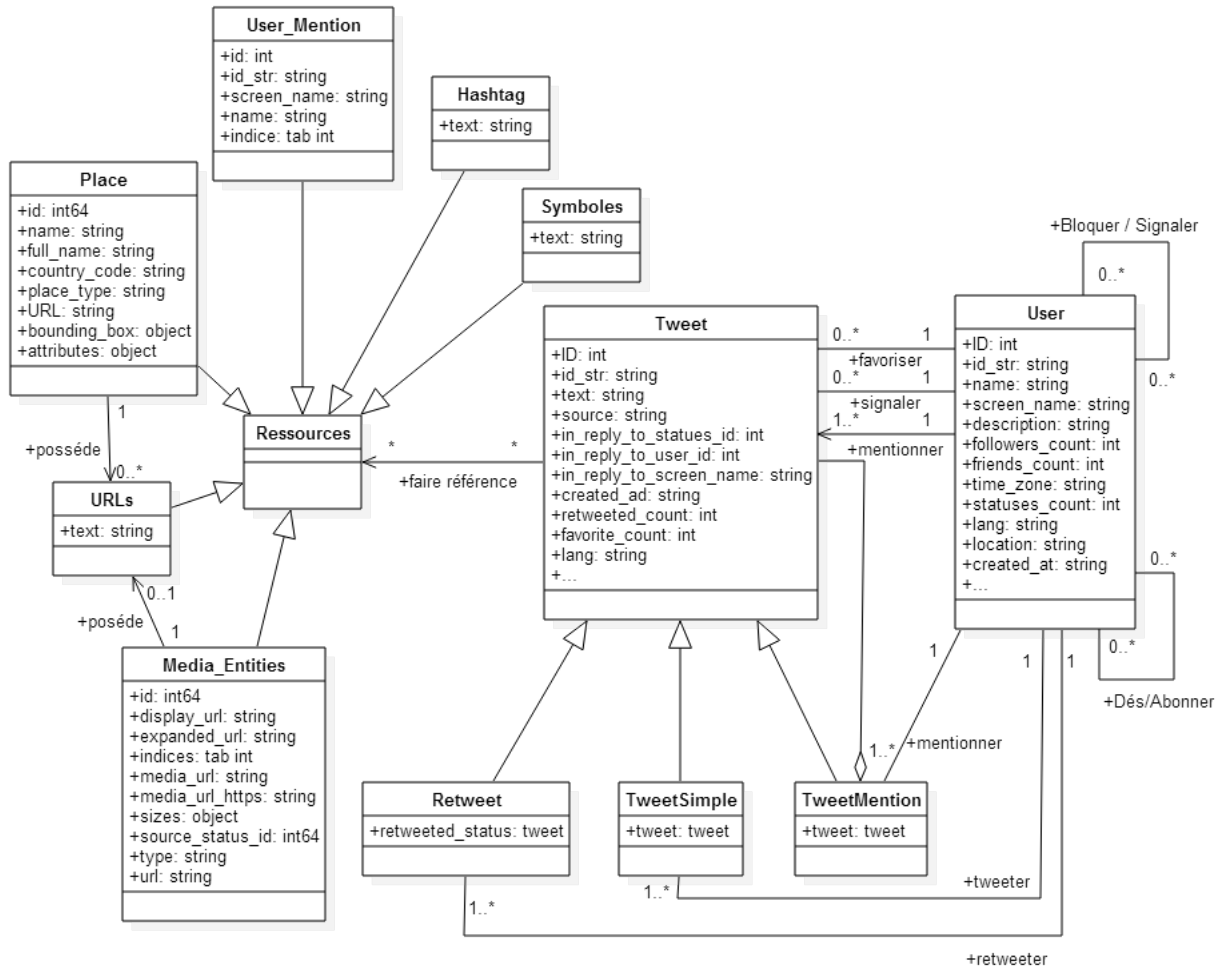


FIGURE I.1 – Méta-modèle du réseau social Twitter.

### I.3 Description de l'approche

Pour répondre à la problématique générale, nous avons conçu une approche composée de deux étapes essentielles :

1. **La Préparation du corpus** : Cette première étape permet de construire le corpus d'étude, et ceci ce fait en deux phases :
  - (a) Une première phase pour la construction du corpus de base qui consiste en la recherche et la récupération des entités "publications" et ceci via un processus de recherche par mots-clés.
  - (b) Une deuxième phase pour constituer notre corpus d'étude. Le corpus de base ne contenant pas les entités nécessaire pour notre étude. Ce dernier est complété par d'autres entités issues du réseau social, à savoir les utilisateurs qui ont postés ou retweetés les publications précédemment récupérées. La récupération des entités engendre des entités superflues, c'est pour cela que ces dernières subiront une phase d'élagage, pour garder que les entités pertinentes.

Comme les réseaux sociaux sont généralement représentés sous forme de graphe et comme la détection des influenceurs repose sur les relations entre les individus, une base de données graphe

permet de mettre en valeur ce type de relations, ainsi nous sauvegardons notre corpus dans une base de données orientée graphes (NoSQL).

2. **La détection des influenceurs** : Cette étape consiste à calculer le score d'influence de certaines entités du corpus à savoir les entités "utilisateurs". Ces dernières sont d'abord évaluées selon deux critères : leurs popularités et leurs impacts dans le réseau, viendra par la suite l'étape d'affectation du score d'influence. La figure I.2 schématise notre approche.

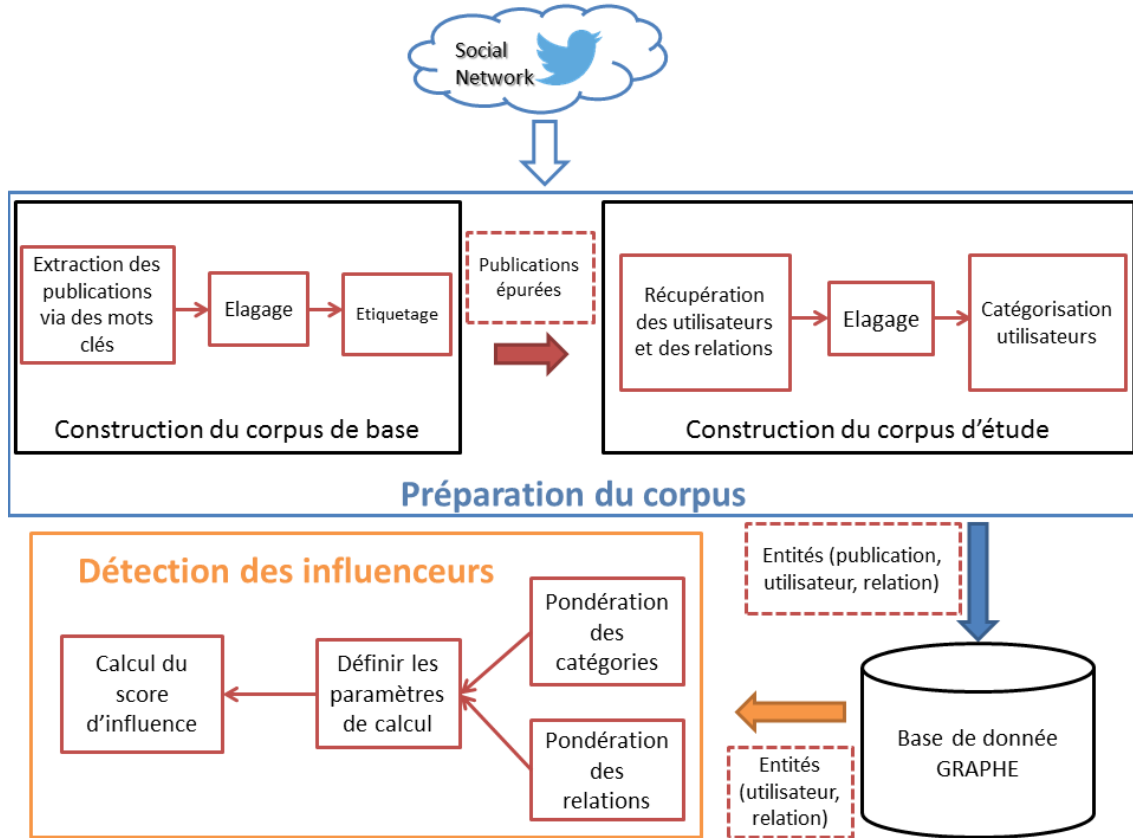


FIGURE I.2 – Schématisation de notre approche

Dans ce qui suit, nous allons détailler chaque étape de notre approche.

### I.3.1 Étape 1 : Préparation du corpus

Au niveau de cette étape, nous allons extraire les données nécessaires du réseau social dans le but de construire notre corpus d'étude. Cela se passe en deux phases essentielles :

1. Construction du corpus de base.
2. Construction du corpus d'étude.

Ces deux phases seront expliquées plus en détails dans ce qui suit. La figure I.3 résume de cette étape de notre approche.

L'étape de préparation du corpus est importante, elle constitue une étape de validation des données qui seront apte à être analysées et étudiées dans la suite de notre approche et ceci dans le but de garantir de meilleurs résultats.

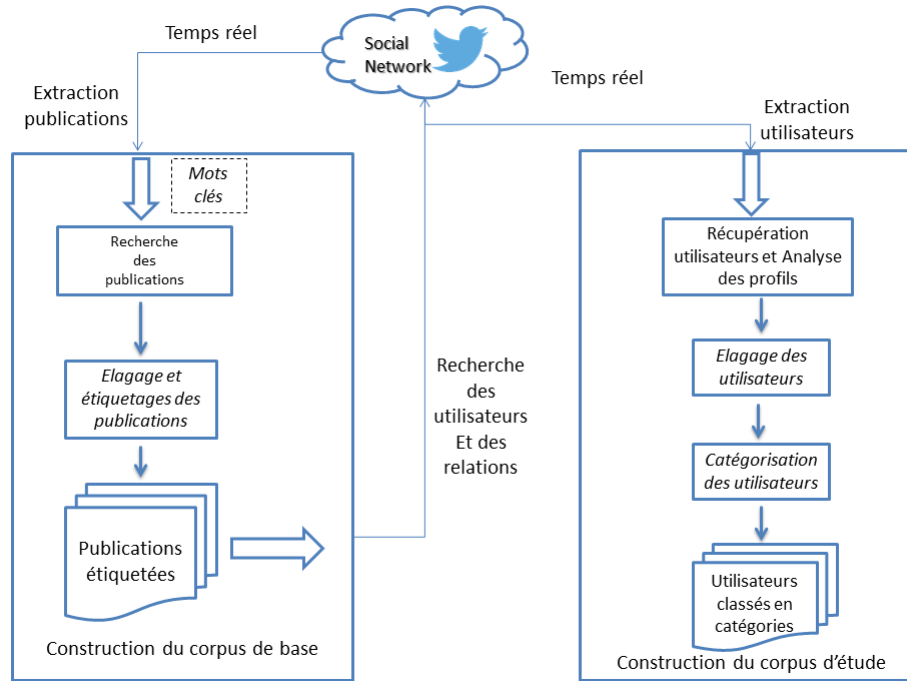


FIGURE I.3 – Préparation du Corpus

### Phase 1 : Construction du corpus de base

1. **Extraction des entités « publications »** : L'extraction des entités "publications" du réseau social se fait en temps réel. Chaque entité est une donnée récoltée à partir de plusieurs mots-clés à l'aide de l'API du réseau social. Chaque récolte de données est référencée par les mots-clés introduits et appartient à un domaine connu au préalable.

Comme nous n'avons pas effectué la catégorisation automatique des textes, il est difficile de savoir à quel domaine appartient chaque mot-clé utilisé. Cependant pour garantir de meilleurs résultats d'exécution de notre approche, il est préférable de respecter le contexte des domaines prédéfinis.

2. **Élagage et étiquetages des entités « publications »** : Les données une fois récoltées subissent un traitement d'élagage, ce dernier consiste à :

- Éliminer les publications non conformes à la langue de l'utilisateur, dans notre cas nous nous basons essentiellement sur les publications en anglais.
- Éliminer les publications dont le nombre de rediffusion ne dépasse pas un nombre N, car plus une publication est rediffusée plus elle est considérée comme pertinente.

Une fois les publications élaguées, ces dernières sont étiquetées par le biais des mots-clés « hashtags » qui y figurent. Cette phase peut être effectuée suivant l'algorithme 1 présenté ci-dessous.

---

**Algorithm 1** Construction du corpus de base

---

**Input :** Domaine + Mots-clés

**Output :** Corpus de base

- 1: Sélectionner un domaine ;
  - 2: Choisir mots-clés ;
  - 3: Recherche et récupération des entités "publications" référencées par les mots-clés ;
  - 4: **if** *Publications non conformes* **and** *nombre\_rediffusion* < *N* **then**
  - 5:   Élagage des entités inutiles
  - 6: **end if**
  - 7: Étiquetages des entités ;
- 

**Phase2 : Construction du corpus d'étude**

A partir des entités publications récupérées pour la construction du corpus de base, nous récupérons d'autres entités (utilisateurs et relations), soit les utilisateurs qui ont postés ou retweetés les publications précédemment récupérées lors de la première phase ainsi que leurs relations. Cette phase consiste à préparer les données pour les traitements. Pour ce faire, les entités utilisateurs sont d'abord filtrées avant d'être analysées et mises en catégories.

1. **Récupération des entités utilisateurs et des relations** Le corpus de base créé dans la première phase sera complété par d'autres entités extraites elles aussi à partir du réseau social. Ces entités ont un lien direct avec les entités publications du corpus de base. En effet, ces dernières sont essentiellement des utilisateurs "retweeters" ou "posteurs" des publications ainsi que leurs relations (indegree et outdegree) dans le réseau social.
2. **Élagages des entités « utilisateurs »** : Les filtres appliqués correspondent à l'élimination des utilisateurs non pertinents pour notre approche suivant certains critères énumérés dans ce qui suit :
  - (a) Élimination des utilisateurs ayant des profils incomplets. Les informations tel que la description du profil sont essentielles pour bien cibler leur domaine d'activité.
  - (b) Élimination des utilisateurs ayant un nombre de publications inférieur à  $m$  publications durant toute leur présence sur le réseau social. Ils sont automatiquement reconnus comme des utilisateurs non-actifs.
  - (c) Élimination des utilisateurs n'ayant pas de nom d'utilisateur « screen\_name » valide. Cette information est cruciale pour déterminer le nombre de fois qu'un utilisateur a été « mentionné » ou « retweeté » par les autres.
3. **Catégorisation des utilisateurs** : Les utilisateurs sont classés en trois principales catégories (Célébrités, Experts, Consommateurs) et cela en se basant sur les propriétés intrinsèques de leurs profils. D'autre part la classification des utilisateurs est un moyen de gérer les profils et ainsi déterminer un des deux critères d'influence à savoir le critère de la « popularité » de l'utilisateur, et avoir un aperçu de sa position dans la pyramide de l'influence vue précédemment.

Dans ce qui suit nous présentons ces trois catégories créées pour la classification.

(a) **Catégorie 1 : Les célébrités**

Cette catégorie est affectée à chaque utilisateur disposant d'un profil de type « certifié ». Ces

utilisateurs sont souvent connus par le grand public et dispose souvent d'un degré de crédibilité important dans leur domaine d'activité. Ils sont donc au sommet de la pyramide et sont souvent les personnes les plus suivies.

(b) **Catégorie 2 : Les experts**

Cette catégorie comprend généralement les bloggeurs, les journalistes et les experts dans un domaine. Ces derniers oeuvrent dans la publication d'articles faisant référence aux marques ou aux entreprises. Cette catégorie d'utilisateurs prodigue à la marque une plus grande visibilité au sein du réseau social.

(c) **Catégorie 3 : Les consommateurs**

Cette dernière catégorie se situe au niveau de la base de la pyramide d'influence. Elle représente de ce fait les consommateurs qu'on appelle ainsi car ces derniers consomment (adoptent) l'opinion des autres d'une part. D'une autre part, ils représentent souvent la majeure partie de notre corpus et sont souvent des ponts de connexion entre utilisateurs de plus haut niveau favorisant ainsi la propagation de l'information.

L'algorithme 2, présenté ci-dessous, synthétise les différentes étapes de cette phase.

L'opération 2 correspond à la récupération des profils utilisateurs. Les opérations 8, 11, 14 correspondent au filtrage de ces profils alors que les opérations 18, 20, 22 correspondent à la catégorisation des utilisateurs.

---

**Algorithm 2** Construction corpus d'étude

---

**Input :** Corpus de base

**Output :** Entités (publications, utilisateurs, relations) filtrées et classées

```

1: for each publication do
2:   if nombre_rediffusion  $\geq N$  then
3:     Récupérer les utilisateurs retweeters et posteurs ;
4:     Récupérer les relations entres les utilisateurs ;
5:   end if
6: end for
7: for each utilisateur do
8:   if Profil_incomplet then
9:     Élagager utilisateur ;
10:  end if
11:  if nombre_publication  $< N$  then
12:    Élagager utilisateur ;
13:  end if
14:  if Screen_Name non valide then
15:    Élagager utilisateur ;
16:  end if
17:  if utilisateur_certifié then
18:    catégorie  $\leftarrow$  catégorie1 : célébrités ;
19:  else if nb_rediffusion  $\geq N$  then
20:    catégorie  $\leftarrow$  catégorie2 : experts ;
21:  else
22:    catégorie  $\leftarrow$  catégorie3 : consommateurs ;
23:  end if
24: end for

```

---

Ainsi une fois le corpus d'étude bien défini, on passe à la deuxième étape de notre approche qui est la détection des influenceurs. Cette étape est néanmoins précédée par l'étape de sauvegarde du corpus dans une base de données.

Le stockage des entités récoltées dans la base de données se fait en parallèle avec l'étape de préparation du corpus ainsi chaque entité analysée et catégorisée au cours des phases précédentes est chargée dans la base de données graphe, les principales entités chargées sont :

- Les publications : Ces entités sont représentées par des **noeuds** au niveau de la base de données. Ces noeuds possèdent des labels qui représentent les propriétés de la publication (son auteur, sa date de publication, le nombre de retweet, ..etc).
- Les utilisateurs : Ces entités sont aussi représentées par des **noeuds** labellisés avec des informations sur l'utilisateur (son nom, sa localisation, ses followers, ses followings, ...etc).
- Les relations : sont représentées par des **arcs** dirigés « indegree » et « outdegree » et labellisés, constituent les liens entre les **noeuds** de la base de données cités avant.

La figure I.4 représente l'étape de sauvegarde des entités dans la base de données graphes.

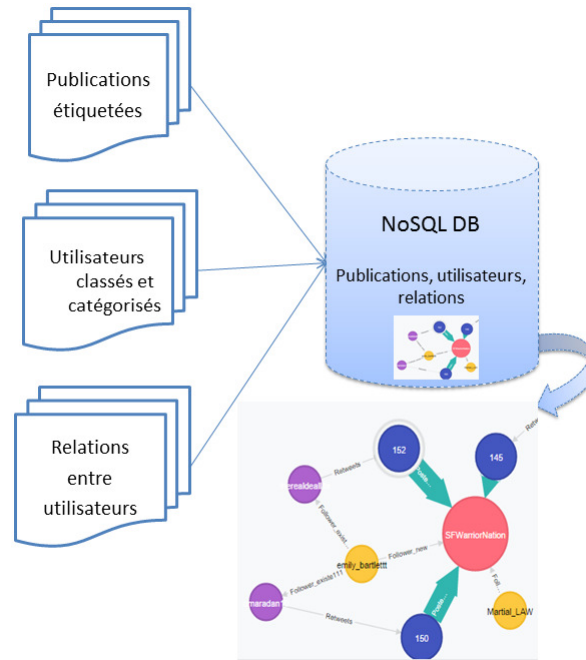


FIGURE I.4 – Sauvegarde des entités dans la base de données

### I.3.2 Étape 2 : Détection des l'influenceurs

Cette étape constitue la dernière partie de notre approche. Un score d'influence est attribué à chaque utilisateur. Pour cela, nous nous basons sur deux critères d'influence : la popularité et l'impact.

Pour rappel la popularité de l'utilisateur se distingue par le biais de sa position dans la pyramide de l'influence citée auparavant, les célébrités ayant le degré de popularité le plus important. Cependant, l'impact d'un utilisateur se distingue à travers son réseau de relation au sein du réseau social.

Ainsi, nous nous basons sur ces deux principaux critères car l'impact augmente le degré de visibilité de la marque ou l'entreprise, la popularité leur procure une image de marque durable dans le temps. De ce fait, les individus qui disposent d'un score d'influence élevé au niveau de ces deux critères sont les plus susceptibles d'être reconnus comme influenceurs. La figure I.5 résume l'étape de détection d'influenceurs.

Les principales entités traitées au niveau de cette partie de notre approche sont essentiellement :

- Les utilisateurs : qui constituent les individus influenceurs auxquelles un score d'influence est attribué.
- Les relations : qui constituent les relations « followings » et « followers » des utilisateurs dans le réseau social.

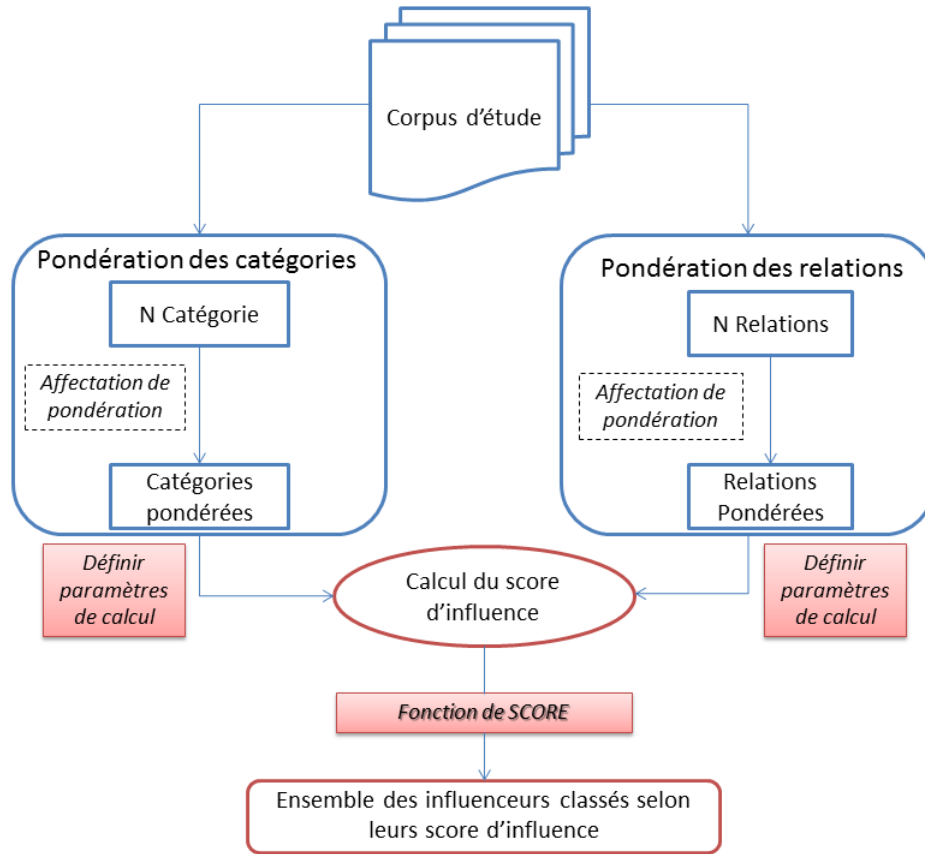


FIGURE I.5 – Étape de détection des influenceurs

A ce niveau, chaque entité est traitée suivant plusieurs étapes expliquées en détails dans ce qui suit.

1. **Pondérations des catégories :** Dans cette partie nous nous intéressons seulement aux utilisateurs. Après la phase de catégorisation des utilisateurs dans la première partie de notre approche, chaque utilisateur est affecté à une catégorie spécifique expliquée précédemment. A ce niveau ces catégories représentent déjà une mesure « qualitative » du niveau d'influence d'un utilisateur à savoir son degré de « **popularité** ». Cependant pour pouvoir évaluer chaque catégorie, il est primordial de « quantifier » cette mesure d'influence. Dans ce qui suit nous procédons à l'évaluation quantitatif de chaque catégorie.

Dans notre approche les catégories sont hiérarchisées, ainsi chaque individu dont la position est plus proche du sommet de la pyramide d'influence se voit attribué une note plus importante que celui qui se trouve plus près de la base de la pyramide.

Pour rappel la pyramide comprend 3 niveaux cités auparavant, c'est à partir de ces niveaux que nous distinguons trois catégories d'utilisateurs. Et selon la catégorie à laquelle s'identifie l'utilisateur, une note attribuée correspondante au niveau de popularité de ce dernier. Les catégories d'utilisateurs sont les suivantes :

- **Catégorie 1 : Célébrités.** Cette catégorie se voit octroyer la plus haute note.
- **Catégorie 2 : Experts.**
- **Catégorie 3 : Consommateurs.**



Ainsi l'évaluation des noeuds se fait suivant les étapes de l'algorithme 3

---

**Algorithm 3** Pondération des catégories

---

**Input :** Entités utilisateurs

**Output :** Entités utilisateurs avec score de popularité

```

1: if Entité_utilisateur  $\in$  Catégorie1 then
2:   ScorePopularité  $\leftarrow$  ScoreCatégorie1;
3: end if
4: if Entité_utilisateur  $\in$  Catégorie2 then
5:   ScorePopularité  $\leftarrow$  ScoreCatégorie2;
6: end if
7: if Entité_utilisateur  $\in$  Catégorie3 then
8:   ScorePopularité  $\leftarrow$  ScoreCatégorie3;
9: end if

```

---

**2. Pondération des relations :** Dans cette partie nous nous intéressons seulement aux relations. Les relations sont comme expliquées auparavant des liens sociaux qui existent entre les utilisateurs du réseau social. Ces liens définissent « **l'impact** » que pourrait avoir un utilisateur sur un autre. C'est donc une mesure d'influence importante pour notre étude.

De ce fait, nous procédons à l'évaluation de cette dernière par le calcul d'un ratio qui définit le nombre de relations sociales d'un utilisateur. Ce nombre est déterminé par le calcul des « arcs entrants » (indegree) et des « arcs sortants » (outdegree) d'un utilisateur dans la base de données graphe, soit le ratio (outdegree/indegree).

Nous avons regroupé les différents cas de figures des résultats de calcul du ratio dans le tableau suivant I.1.

Ratio(outdegree/indegree)	Valeur approximative	Signification
	>1	L'utilisateur a plus tendance à suivre le fil d'actualité des autres utilisateurs.
	<1	L'utilisateur est suivi par beaucoup de personnes.
	=1	Il y a un équilibre entre followers et les following

TABLE I.1 – Calcul du ratio(outdegree/indegree)

La pondération des relation se fait suivant les étapes de l'algorithme 4

---

**Algorithm 4** Pondération des relations

**Input :** Entités "utilisateur" + relations

**Output :** Entités "utilisateur" avec score d'impact

---

```

1: if  $ratio(outdegree/indegree) \geq 1$  then
2:    $ScoreImpact \leftarrow ScoreRelation1$ ; //  $ScoreRelation1 = ratio(outdegree/indegree)$ 
3: end if
4: if  $ratio(outdegree/indegree) \leq 1$  then
5:    $ScoreImpact \leftarrow ScoreRelation2$ ; //  $ScoreRelation2 = ratio(outdegree/indegree)$ 
6: end if
7: if  $ratio(outdegree/indegree) = 1$  then
8:    $ScoreImpact \leftarrow ScoreRelation3$ ; //  $ScoreRelation3 = ratio(outdegree/indegree)$ 
9: end if

```

---

3. **Calcul du score d'influence** Une fois l'évaluation des deux mesures d'influence « popularité » et « impact » effectuées, un calcul du score d'influence est établi et ceci en prenant en compte la pondération des catégories et la pondération des relations.

(a) **Paramètres de calcul :** Toute fois avant de procéder au calcul du degré d'influence d'un utilisateur dans le corpus, nous avons établi un coefficient pour chaque mesure d'influence. Ceci dans le but d'offrir une flexibilité dans cette partie de notre approche pour le choix du paramétrage des deux critères « popularité » et « impact ».

Soit  $\alpha$  et  $\beta$  les paramètres de calcul du score d'influence.

$\alpha$  correspond au paramètre de calcul du critère « popularité ». Soit l'équation suivante I.1.

$$ScorePopularity = \alpha \cdot ScoreCatégorie \quad (I.1)$$

$\beta$  correspond au paramètre de calcul du critère « impact ». Soit l'équation suivante I.2.

$$ScoreImpact = \beta \cdot ScoreRelation \quad (I.2)$$

(b) **Fonction de Score :** Après avoir défini un paramètre pour chaque critère d'influence, on passe au calcul du score final de chaque utilisateur. Ce calcul se fait par le biais d'une fonction de score présentée dans le chapitre précédant et qui garantit une évaluation des utilisateurs selon les critères d'influence établis et ceci au niveau de tout le corpus d'étude dont nous disposons. Un score d'influence est alors affecté à chaque utilisateur dans notre corpus. Les résultats sont présentés sous forme de tableau où chaque utilisateur lui est associé un score d'influence correspondant. L'algorithme 5 montre l'étape de l'évaluation de l'influence des utilisateurs.

---

**Algorithm 5** Calcul du score d'influence

---

**Input :** Entités utilisateurs**Output :** Entités utilisateurs avec score d'influence

```
1: for each utilisateur do  
2:    $ScoreInfluence \leftarrow ScorePopularité + ScoreImpact$  ;  
3: end for
```

---

Notre avons opté pour une approche incrémentale. Une fois les résultats affichés, le corpus peut être incrémenté en temps réel par d'autres entités ainsi de nouvelles entités seront extraites, pour cela le score d'influence de chaque utilisateur sera recalculé et cela dans le but de recherche d'autres influenceurs.

## I.4 Conclusion

Au niveau de ce chapitre, nous avons présenté notre approche qui se résume en deux parties. La première étape de préparation du corpus, consiste en la préparation du corpus de base et du corpus d'étude à partir de données récoltées du réseau social. Ensuite vient la deuxième étape qui consiste à l'étude du corpus préparé précédemment afin de déterminer le niveau d'influence des utilisateurs. Pour cela, un score d'influence est calculé suivant deux mesures d'influence : la popularité et l'impact de l'utilisateur dans le réseau. Dans le prochain chapitre nous présenterons sa modélisation que nous avons réalisé avec le langage UML.

## Chapitre II

# Modélisation de l'approche

### II.1 Introduction

Dans cette partie du mémoire nous allons présenter la conception de notre outil qui permet de réaliser, la construction du corpus ainsi que la détection des influenceurs. La détection des influenceurs se fait par le calcul d'un score d'influence pour chaque utilisateur.

Nous avons opté pour une solution incrémentale, de ce fait, le corpus d'étude sera incrémenté à la demande, c'est alors qu'un score d'influence sera recalculé pour chaque entité utilisateur. Les données seront stockées dans une base de données NoSQL dédiée aux graphes, nous avons choisi la base de données Noe4j connue pour son extensibilité et sa facilité d'utilisation.

La conception de notre outil a été réalisée avec le langage UML<sup>1</sup> qui est dédié à la modélisation des systèmes informatiques. Dans ce qui suit, nous présenterons nos différents diagrammes de conception. Nous commençons par le diagramme de cas d'utilisation suivi du diagramme de séquence et du diagramme d'activité. Et pour finir, nous présenterons notre diagramme de classe.

### II.2 Diagramme de cas d'utilisation

Un diagramme de cas d'utilisation décrit les fonctionnalités principales et nécessaires aux utilisateurs du système. Les cas d'utilisation décrivent sous la forme d'actions et de réactions le comportement d'un système du point de vue d'un utilisateur. Ils permettent de définir les limites du système et les relations entre le système et l'environnement [Gaertner, 2003].

Pour notre système, les principaux acteurs vont des simples blogueurs<sup>2</sup> jusqu'au décideurs d'entreprise en passant par les spécialistes en marketing qui cherche à lancer une campagne marketing sur le web. Ainsi, nous distinguons les cas d'utilisation suivants :

- Construction du corpus.
- Consultation des entités récupérées.
- Détection des influenceurs.
- Incrémentation du corpus.
- Suppression du corpus.

---

1. Unified Modeling Language

2. Personne qui détient et gère son blog.

La figure II.1 illustre notre diagramme de cas d'utilisation qui englobe toutes les fonctionnalités externes pour l'utilisateur. Le tableau II.1 représente les cas d'utilisations avec leurs descriptions.

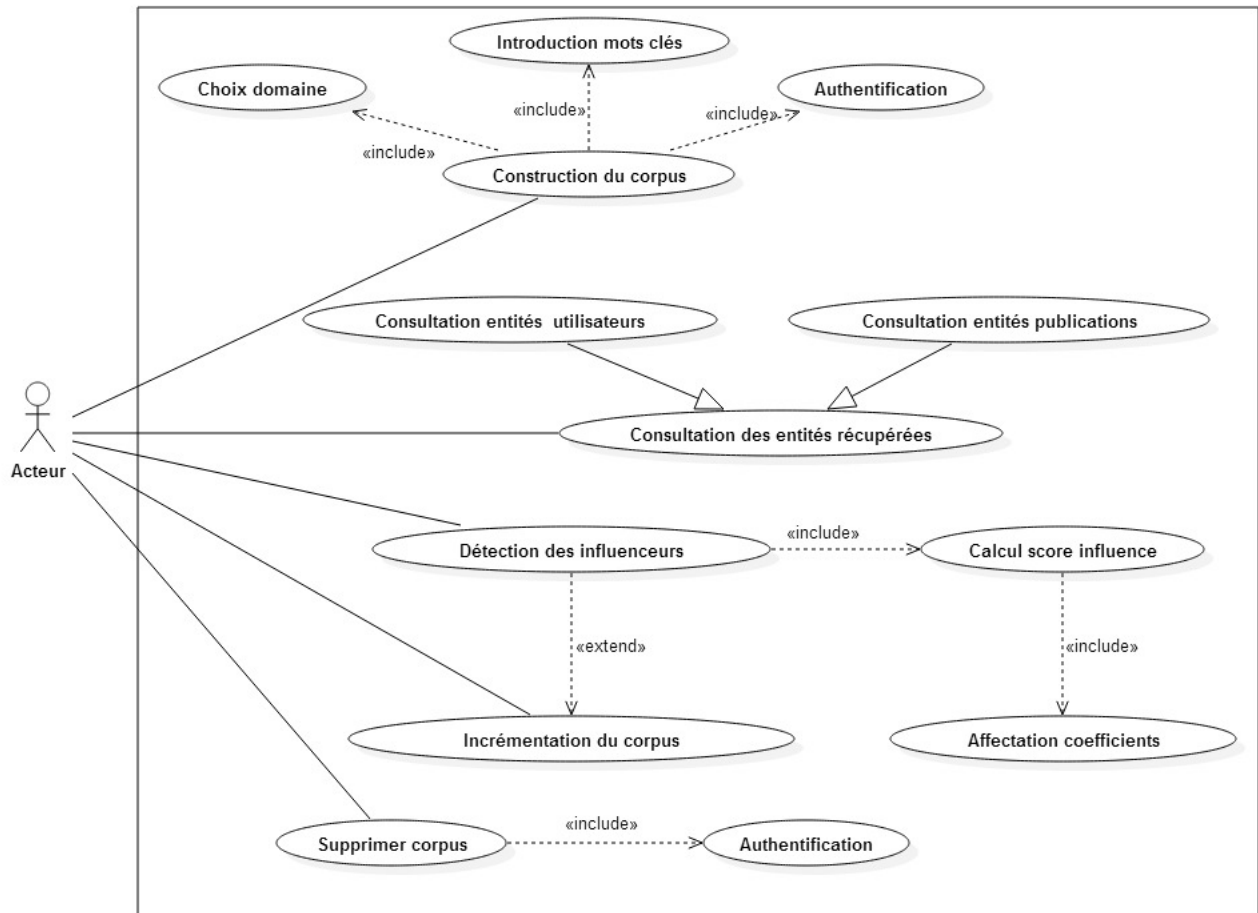


FIGURE II.1 – Diagramme de cas d'utilisation.

### II.3 Diagramme de séquence

Les diagrammes de séquence montrent les interactions et les échanges entre acteur et objets du système manipulés par l'utilisateur, en mettant en valeur la succession chronologique des opérations. Le diagramme de séquence est un moyen de capturer le comportement des fonctionnalités du système présentées dans le diagramme de cas d'utilisation. La représentation se concentre donc sur l'expression des interactions [Gaertner, 2003].

Les diagrammes de séquence de notre outil sont représentés dans ce qui suit. Nous commençons par présenter le diagramme d'authentification.

Cas d'utilisation	Description
Construction du corpus	Le système donne la main à l'utilisateur pour la construction du corpus à partir des données du réseau social. Pour cela, il doit s'authentifier, choisir un domaine et introduire des mots-clés, selon le domaine choisit. L'authentification d'un utilisateur nécessite son enregistrement au niveau du réseau Twitter, afin d'obtenir les clés d'accès pour l'extraction des données qui se fait selon un domaine choisit. Dans le cadre de notre travail, nous avons choisit 3 domaines : Technologie, Sport et Voiture.
Consultation des entités extraites	Permet à l'utilisateur de consulter les entités récoltées après extraction à partir du réseau social. Ces entités sont des publications et des profils utilisateurs.
Détection des influenceurs	Dans cette étape on parle de détection des influenceurs avec la génération du graphe des influenceurs et le calcul du score d'influence. Le calcul du score d'influence est calculé selon deux critères, la popularité et l'impact et cela en attribuant un coefficient à chaque critère.
Incrémentation corpus	L'utilisateur peut incrémenter le corpus dans le but de trouver de nouveaux influenceurs. Et cela, en extrayant de nouvelles entités ainsi le graphe sera régénéré et le score d'influence recalculé.
Suppression du corpus	L'utilisateur peut supprimer le corpus afin de régénérer un nouveau selon d'autres mots-clés et domaine choisi.

TABLE II.1 – Description des cas d'utilisation

### II.3.1 Digramme de séquence d'authentification

L'acteur de cette opération doit être inscrit au réseau social Twitter, ainsi en entrant ces informations d'authentification, il obtient des clés d'accès pour l'extraction des entités. Ce protocole d'authentification est détaillé davantage dans le chapitre implementation de l'outil. La figure II.2 illustre cette étape.

### II.3.2 Diagramme de séquence de préparation du corpus

Cette opération vient juste après l'authentification, l'acteur doit choisir un domaine et introduire des mots-clés en rapport avec le domaine choisit afin d'extraire les entités à partir du réseau social. Une fois les entités récupérées du réseau social, une opération d'élagage des entités récupérées est effectuée avant d'entamer le stockage dans la base de données. L'élagage des entités consiste en :

- La suppression des utilisateurs ayant un nombre de followers inférieur à 100.
- La suppression des publications qui ont été rediffusées moins de 50 fois.

La figure II.3 illustre l'opération de préparation du corpus.

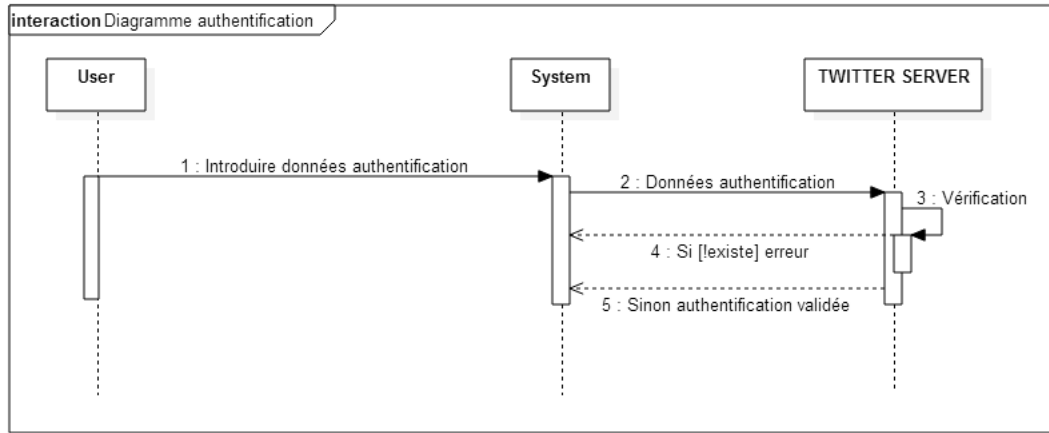


FIGURE II.2 – Diagramme de séquence authentification.

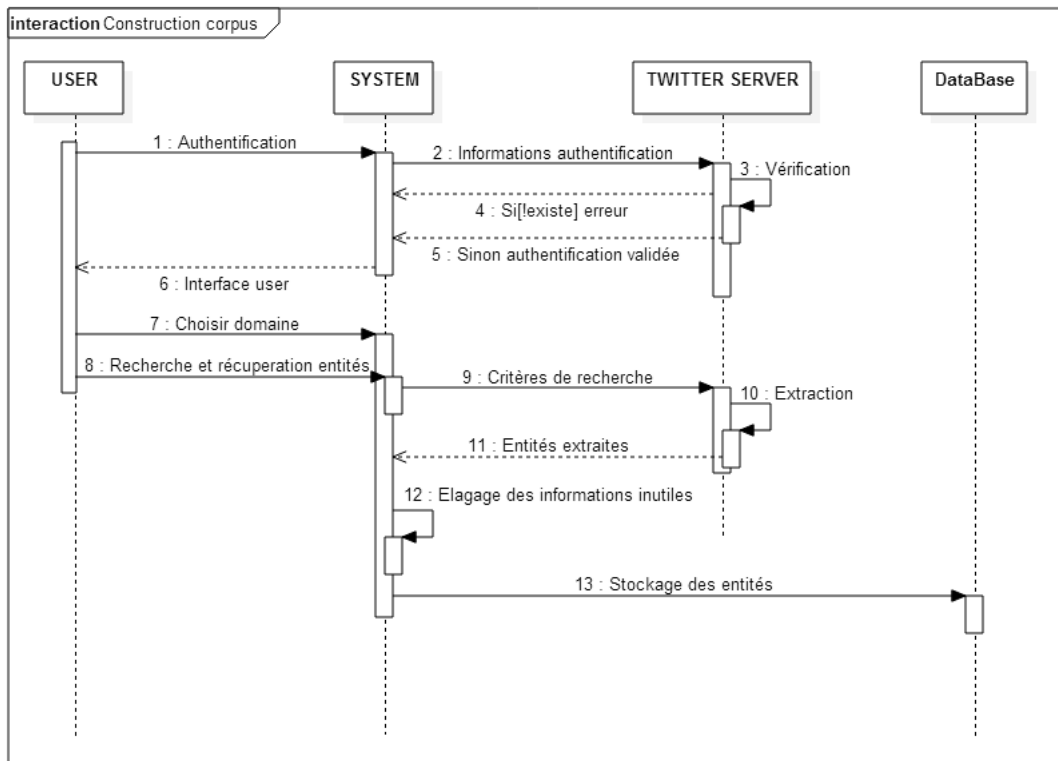


FIGURE II.3 – Diagramme de séquence de préparation du corpus

### II.3.3 Diagramme de séquence de consultation des entités

L'acteur peut consulter les différentes entités extraites, une fois l'extraction et l'élagage des informations inutiles effectué. L'acteur peut donc consulter les publications et les profils utilisateurs extraits qui correspondent au domaine choisit et aux mots-clés introduis précédemment. La figure II.4 représente cette étape.

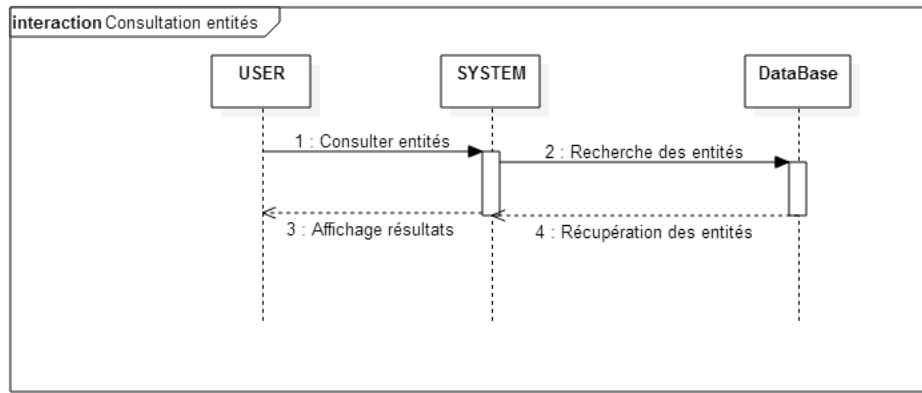


FIGURE II.4 – Diagramme de séquence de consultation des entités.

### II.3.4 Diagramme de séquence de détection des influenceurs

Cette opération permet à l'acteur de détecter les utilisateurs influenceurs et cela en calculant le score d'influence. Le score d'influence est calculé selon deux critères la popularité de l'utilisateur et son impact dans le corpus. La figure II.5 illustre cette fonctionnalité.

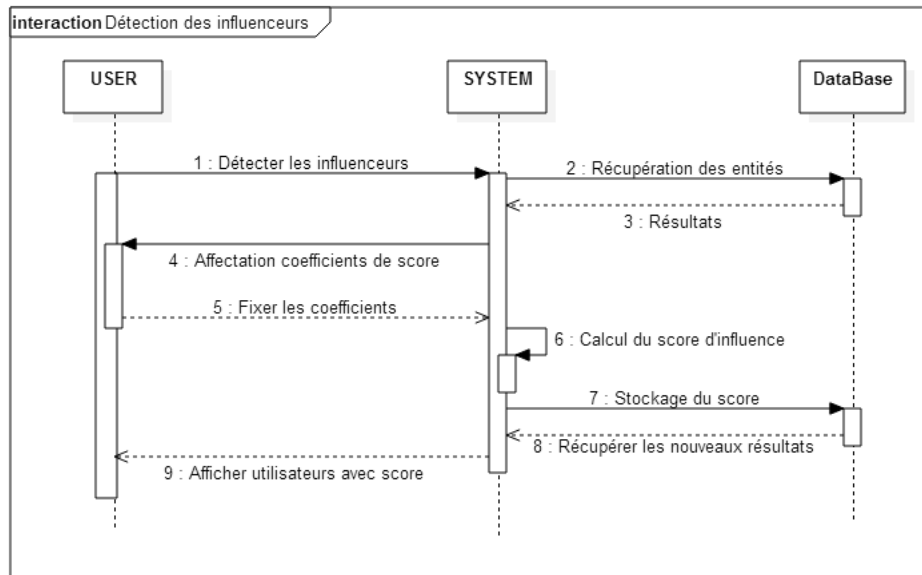


FIGURE II.5 – Diagramme de séquence détection des influenceurs.

### II.3.5 Diagramme de séquence d'incrémentation du corpus

L'acteur peut aussi incrémenter le corpus, pour cela, il doit faire une nouvelle extraction de données, ainsi le graphe des influenceurs sera régénéré et le score d'influence recalculé. La figure II.6 représente cette opération.



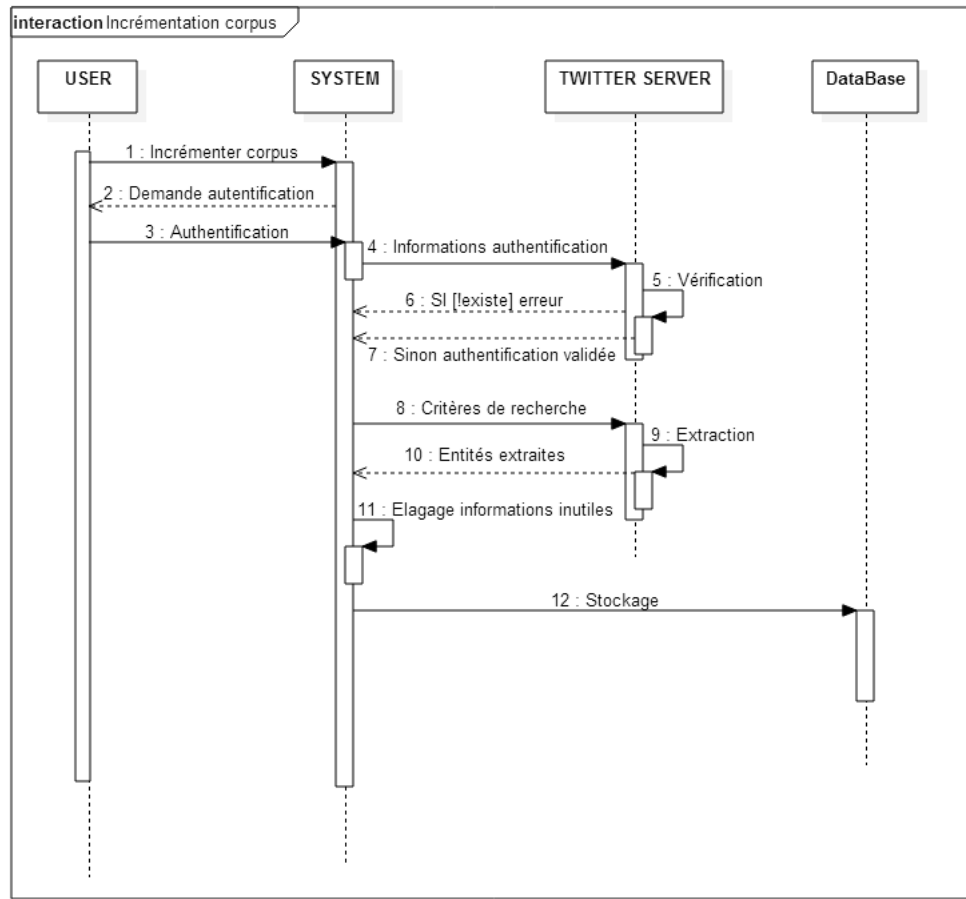


FIGURE II.6 – Diagramme de séquence d'incrémentation du corpus.

### II.3.6 Diagrammes d'activité

Le diagramme d'activité est une variante des diagrammes d'états transitions, organisé par rapport aux actions et principalement destiné à représenter le comportement interne d'une méthode (la réalisation d'une opération) ou d'un cas d'utilisation. Le but de ce diagramme est de mettre en évidence les contraintes de séquentialité et de parallélisme qui pèsent sur le processus global [Gaertner, 2003].

Dans notre cas, nous avons illustré les deux étapes qui constitue notre approche à savoir la préparation du corpus et la détection des influenceurs. La figure II.7 illustre l'enchaînement général de notre processus.

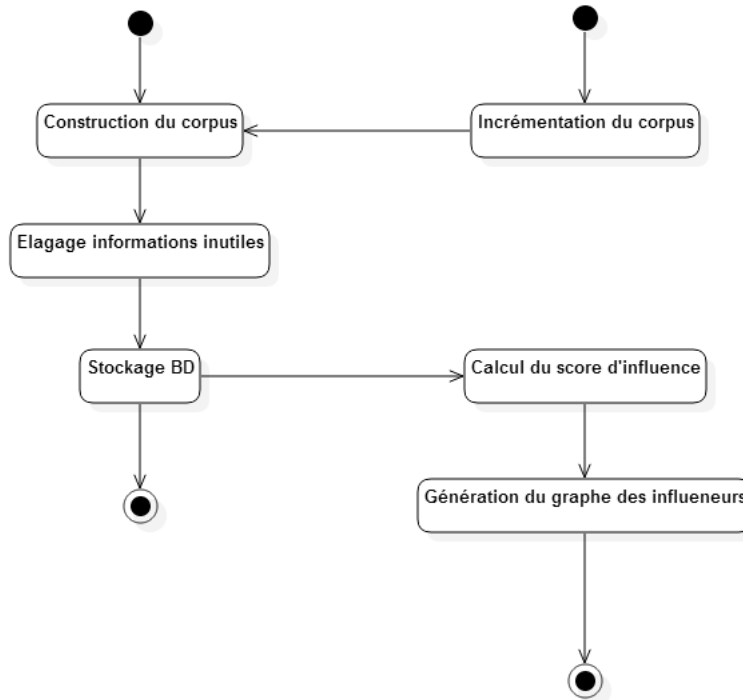


FIGURE II.7 – Diagramme d'activité.

### II.3.7 Diagramme de classes

Un diagramme de classe décrit de manière générale la structure d'un système, en termes de classes et de relations entre les classes. On distingue deux principaux types de relations entre objets [Gaertner, 2003] :

- Les associations, connues depuis les vieux modèles entité/association utilisés dans la conception des bases de données.
- Les sous-types exprimés à l'aide de l'héritage qu'on retrouve dans la conception orienté objets.

Dans la figure II.8 nous allons présenter le diagramme de classe de notre outil.

Notre corpus contient des utilisateurs et des publications et est généré selon plusieurs mots-clés. Chaque mot-clé appartient à un domaine bien précis. Un utilisateur appartient à une catégorie. Il publie et rediffuse une ou plusieurs publications. Un utilisateur peut être le friend ou le follower d'un autre utilisateur. Pour une session donnée plusieurs mots-clés sont utilisés afin de calculer le score d'influence des utilisateurs et cela suivant plusieurs critères.

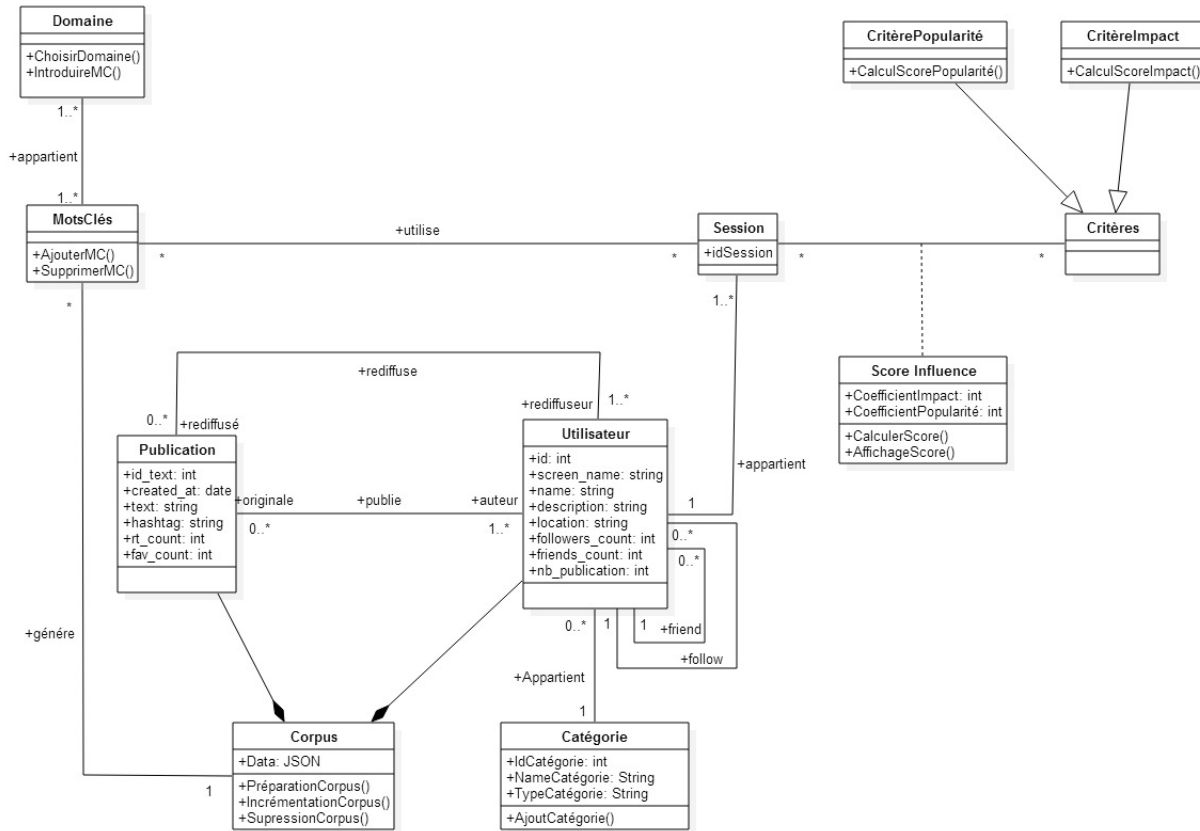


FIGURE II.8 – Diagramme de classe de notre outil.

## II.4 Conclusion

Dans cette partie du mémoire, nous avons présenté la modélisation de notre approche réalisée en UML. Dans le chapitre suivant, nous allons présenter la dernière étape de notre projet qui est la réalisation de notre outil.

# Chapitre III

## Implémentation de l'outil

### III.1 Introduction

Ce chapitre sera consacré à la réalisation de notre outil de détection d'influenceurs dans le réseau social Twitter. Nous allons dans ce qui suit présenter les modules qui ont été réalisés en se basant la conception préalablement présentée. Nous allons par la suite décrire l'environnement du travail et les principales technologies utilisées pour le développement de notre outil. On terminera par des captures écran de l'outil développé.

- Un module d'authentification, ce module comprend l'accès à l'API twitter.
- Un module de récupération des données à partir de Twitter par le biais de mots-clés fournis par l'utilisateur. La récolte de données est une étape très importante dans notre système et cela à fin de construire et de préparer le corpus pour les traitements.
- Un module pour la sauvegarde et la représentation des entités récoltées dans une base de données. Notre choix de base de données s'est posé sur une base de données orientée graphe Neo4J.
- Un module de consultation et d'affichage des entités récoltées qui a pour but principal la mise en relation entre les besoins de l'utilisateur exprimés sous formes de requêtes et les informations précédemment récupérés et disponibles dans notre base de données.
- Un module de détection d'influenceur qui comprend la partie évaluation des entités et le calcul du score d'influence. Ce module propose aussi une visualisation graphique de la base de données.

### III.2 Outils de développement

Dans cette section, nous présentons les différentes plates-formes logicielles nécessaires pour le fonctionnement de notre système. Nous passerons en revue les environnements de développement utilisés ainsi que les arguments relatifs aux choix de leurs utilisation. Enfin nous proposerons une visualisation d'une instance de notre base de données graphe Neo4J.

Les outils que nous avons adoptés pour implémenter les différentes composantes de notre système sont :

- Le langage Python sur un environnement de développement Pydev d'Eclipse.
- Base de donnée : Neo4j.
- API : Tweepy

Tous ces outils vont être détaillés, en mettant l'accent sur leurs utilités ainsi que les atouts qu'ils présentent.

### III.2.1 Langage de programmation

Pour assurer un meilleur déploiement de notre système de recherche d'influenceurs dans un réseau social, nous avons décidé de le rendre indépendant vis-à-vis des différentes plateformes (systèmes d'exploitation) qui opèrent sur la machine. L'implémentation de notre application doit prendre en considération un paramètre primordial : la portabilité. Pour cela, nous avons fait le choix d'un langage multiplateforme (portable) qui est : Python. Et plus précisément la version 2.7.

#### Les modules Python utilisés

Python est un langage riche qui offre de très nombreuses bibliothèques et fonctions pour réaliser des tâches courantes. Un module peut être appelé depuis plusieurs programmes, il s'agit d'un fichier avec l'extension .py. N'importe quel fichier .py peut donc être appelé depuis un autre comme un module [8]. Il peut contenir, du script, des fonctions, des classes. Parmi les modules essentiels qu'on a utilisés :

- **Tweepy 2.2** : Module qui donne accès à l'API de Twitter, et permet d'obtenir un objet et d'utiliser n'importe quelle méthode que l'API Twitter. Il soutient l'accès à Twitter via l'authentification de base avec la méthode OAuth.
- **Py2neo** : c'est une bibliothèque de Python qui permet d'accéder à la base de données orienté graphe Neo4j, via son interface de service Web 'RESTful', et d'effectuer des requêtes en langage Cypher grâce à son module cypher.
- **Matplotlib** : Bibliothèque du langage de programmation Python destinée pour la visualisation graphique. Elle peut être combinée avec les bibliothèques Python de calcul scientifique NumPy et SciPy.
- **Numpy** : Bibliothèque du langage de programmation Python qui contient des modules pour l'optimisation, l'algèbre linéaire, les statistiques, utilisée lors du calcul de la fonction de score.

### III.2.2 Environnements de développement

Dans ce qui suit, nous citons les environnements de développement de notre outil.

#### Editeur de code : PyDev

PyDev<sup>1</sup> est un plug-in tiers pour Eclipse. Il est un environnement de développement intégré (IDE) utilisé pour la programmation en Python. Il dispose de l'autocomplétion qui permet d'accéder facilement aux fonctions natives du langage Python, le débogage graphique, analyse de code et de nombreuses autres fonctionnalités. C'est également un éditeur de code très pratique qui permet de créer des projets Python, et d'éditer des scripts Python.

#### Editeur d'interface graphique : PyQt

PyQt est un module libre qui permet de lier le langage Python avec la bibliothèque Qt distribué sous deux licences : une commerciale et la GNU GPL. Il permet ainsi de créer des interfaces graphiques en

---

1. <http://pydev.org/>

Python. Une extension de QtDesigner (utilitaire graphique de création d'interfaces Qt) permet de générer le code Python d'interfaces graphiques.

### Environnement de travail :

Nous avons travaillé sur le système d'exploitation Windows 7, 64 bit, installé sur des machines doté d'un processeur Intel i5, CPU 2.50 Ghz, et une mémoire vive de 6GO.

### III.2.3 Les API utilisés

Dans le cadre de notre travail, nous avons principalement utilisé l'API de Twitter pour la récupération des entités étudiées.

Les APIs d'accès aux données de Twitter peuvent être classées selon deux types, en fonction de leur conception et leur méthode d'accès :

- **REST API** : basée sur l'architecture REST couramment utilisée pour la conception des APIs Web. Ces APIs utilisent la stratégie « PULL » pour la récupération des données. Pour recueillir des informations à partir de Twitter, un utilisateur doit explicitement envoyer une demande (requête).
- **STREAM API** : basée sur le principe de « Streaming ». Elles fournissent un flux continu d'informations publiques de Twitter. Ces APIs utilisent la stratégie « PUSH » pour la récupération de données (en fournissant les mises à jour sans intervention de l'utilisateur).

Dans notre cas nous avons opté pour l'utilisation de l'API Streaming de twitter car comme notre approche se base sur la récolte de retweets en temps réels il est préférable de suivre les retweets comme ils se produisent plutôt que d'essayer de les trouver dans le passé. Par ailleurs l'API REST mentionnée précédemment fournit un nombre limité des utilisateurs qui ont effectués le retweet.

Les APIs basées sur le Streaming comportent trois types de paramètres :

- flux publics : représentent des flux contenant les tweets publiques sur Twitter.
- flux d'un compte : représentent des flux mono-utilisateur, avec tous les tweets d'un utilisateur donné.
- flux d'un site : représentent des flux multi-utilisateurs destinés à des applications accédant au tweets relatifs à un ensemble d'utilisateurs.

### Accès à Twitter API

Les APIs Twitter sont accessibles seulement via des requêtes authentifiées. Twitter utilise le protocole Open Authentication « OAuth » et chaque demande doit être signée avec des informations d'identification valable d'un compte Twitter.

Open Authentication (OAuth) est un standard ouvert pour l'authentification, adopté par Twitter pour fournir l'accès à des renseignements des comptes des utilisateurs non sécurisé.

L'authentification des demandes d'API sur Twitter est effectuée en utilisant 'OAuth' qui fournit une alternative plus sûre que les approches traditionnelles d'authentification. En plus de ça le mot de passe de compte Twitter de l'utilisateur n'est jamais demandé ou partagé avec des applications tierces, l'API n'est accessible qu'avec des applications.

Les applications sont aussi connues sous le nom de "customer" et elles doivent toutes être enregistrées sur Twitter developers<sup>2</sup>, après l'enregistrement, Twitter fournit deux clés d'accès à l'application, les clés d'accès secrète « Consumer Key » et « Consumer Secret », ces deux clés doivent être utilisées par l'application pour s'authentifier auprès de Twitter applications. L'utilisateur utilise ses deux clés pour authentifier son application sur Twitter.

Twitter vérifie l'identité de l'utilisateur et lui délivre un PIN. L'utilisateur fournit ce PIN à son application pour demander un « Access Token » et un « Access Secret », unique à cet utilisateur. En utilisant les jetons d'accès Access Token et Access Secret l'application authentifier l'utilisateur auprès de Twitter, et lance des appels API au nom de l'utilisateur.

Après avoir traité les données de Twitter avec les différents modules cités précédemment, nous les sauvegardons dans une base de données orientée graphes, pour cela nous avons choisi la BD Neo4J de Neotechnologie. Le processus d'authentification à Twitter est illustré dans la figure III.1

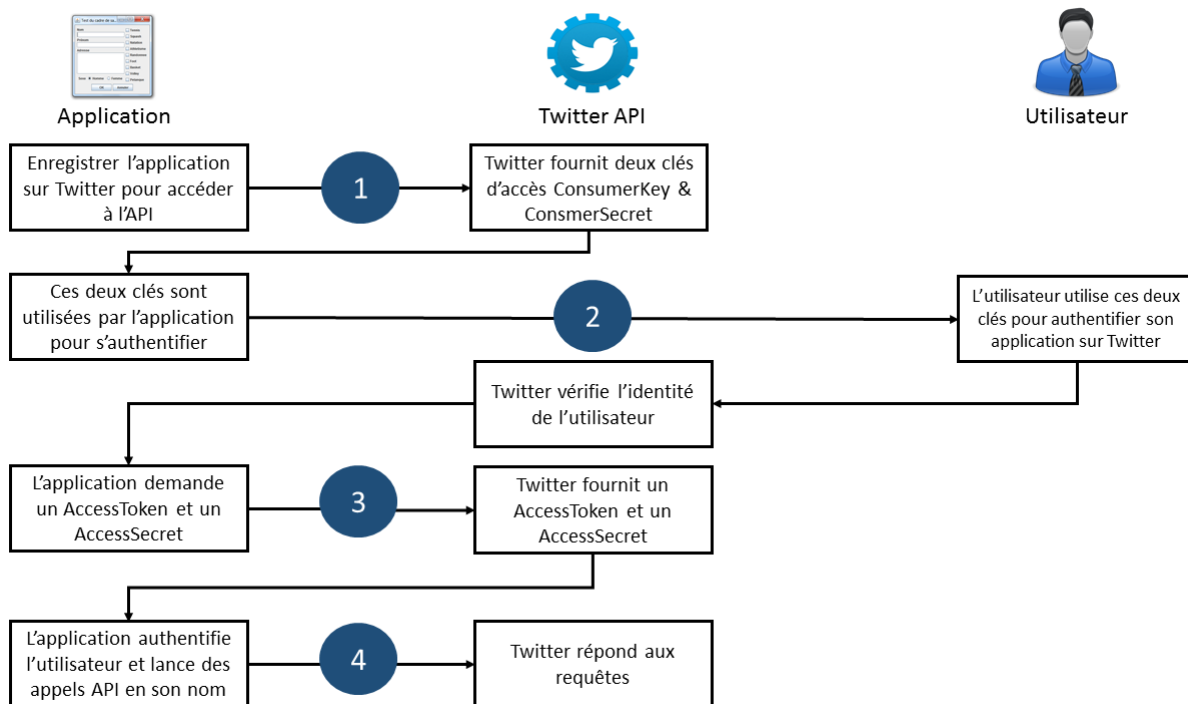


FIGURE III.1 – Les étapes d'authentification sur Twitter API.

2. <http://dev.twitter.com/>

### III.2.4 Base de données orientée graphes Neo4j

Neo4j est un système de gestion de base de données orienté graphes, au code source libre, développé en Java par la société suédo-américaine Neo Technology. Son modèle de données est un graphe de propriétés et son langage de requêtage est Cypher.

#### Avantages de Neo4j

1. Des requêtes haute performance : Le modèle de données du graphe permet l'exécution de requêtes complexes avec une haute performance, inhérentes aux données connectées des applications d'aujourd'hui. En un mot le bon outil pour le bon besoin.
2. Des livrables en temps record : La Modélisation d'une base de donnée est simple et facile. Les entreprises peuvent capturer rapidement toutes sortes de données, structurées, semi-structurées et déstructurées et ainsi les stocker dans Neo4j. Ceci résultant dans une réduction des temps de développement, une réduction de coûts de maintenance.
3. Fiabilité garantie par le respect des propriétés ACID.
4. Système de stockage sur disque rapide et durable.
5. Extensible (peut contenir des milliards de noeuds, relations et propriétés).
6. Expressif avec un langage (CYPHER) proche du langage humain.
7. Simple et accessible par une interface REST ou une API Java orientée objet.

En plus de ses avantages, Neo4j se compose principalement de :

- Noeuds : Les noeuds peuvent contenir des propriétés. Les noeuds peuvent avoir 0 à plusieurs labels.
- Relations : Les relations peuvent également contenir des propriétés. Une relation relie deux noeuds entre eux. Elle possède nécessairement un noeud de départ et un noeud d'arrivée. Une relation est toujours orientée (entrante ou sortante), cependant, elle peut être traversée dans les deux directions. Un noeud peut être relié à lui-même. Toute relation a un type qui peut être vu comme un label.
- Propriétés : Les propriétés sont des couples clé-valeur. La clé est toujours de type String. La valeur peut être de type primitif (boolean, byte, int, etc ) ou un tableau de primitifs. NULL n'est pas une valeur valide. Il est modélisé par l'absence de clé.
- Labels : Les labels regroupent des noeuds pour former des ensembles. Les labels sont utilisés pour définir des contraintes et ajouter des indexes pour les propriétés. Les labels peuvent être utilisés pour améliorer les performances d'une requête. Exemple : Mettre un label Utilisateur sur tous les utilisateurs pour faire une recherche de tous les utilisateurs dont le nom est X. Les labels peuvent être modifiés pendant l'exécution, ils peuvent donc être utilisés pour marquer des états temporaires sur les noeuds.

#### Cas d'usage de Neo4j

Neo4j est très utilisé, parmi les exemples de cas d'usage, nous citons :

1. **Gestion de réseau et analyse d'impact** : dans le but d'avoir le contrôle sur son réseau et d'identifier en temps réel les clients affectés par une panne.



2. **Logistiques** : calculer le meilleur chemin pour des livraisons de colis ou autres.
3. **Collaboration sociale** : rechercher très facilement qui sont les amis de mes amis.
4. **Recommandation** : définir en temps réel la liste des produits achetés par mes amis que je n'ai pas moi même achetés.
5. **Géo-Spatial** : modélisation d'une carte routière et calculs d'itinéraires.
6. **Biologie, interactions moléculaires** : réduire les risques d'effets secondaires des médicaments en calculant en temps réel les interactions entre une protéine et une future molécule.

### Langage d'interrogation Cypher

Cypher est le langage de requête de Neo4j proche du langage SQL. Il est également très intuitif, basé sur la représentation graphique du modèle graphe. Quelque exemple de requête Cypher :

**Création d'un noeud** : `CREATE (ee :Person name : 'Neo', from : 'Matrix' );`

- **CREATE** la clause utilisée pour la création des données.
- **()** parenthèse pour indiquer que c'est un noeud.
- **ee :Person** 'ee' c'est une variable et 'Person' c'est le label du nouveau noeud.

**Recherche d'un noeud** : `MATCH (ee :Person) WHERE ee.name = 'Neo' RETURN ee ;`

- **MATCH** la clause pour spécifier un modèle de noeuds et de relations.
- **(ee :Person)** un motif de noeud unique avec l'étiquette «personne» qui assignera les matchs à la variable 'ee'.
- **WHERE** la clause utilisée pour limiter les résultats.
- **ee.name = 'Emil'** compare la propriété name à la valeur 'Emil'.
- **RETURN** clause utilisée pour demander des résultats particuliers.

### III.2.5 Stockage et modèle physique de données

Dans le cadre de notre travail, on s'intéresse au deux entités, les utilisateurs, et leurs tweets. Chaque utilisateur aura comme attributs ses informations qui se retrouve sur son profil utilisateur dans le réseau social Twitter ainsi que d'autres informations tel que la catégorie à laquelle il appartient et son score d'influence. Chaque tweet aura comme attributs le texte du tweet et d'autres attributs comme le nombre de retweets et la source d'où il a été Tweeté. Comme on parle d'une base de données graphe, les entités citées précédemment seront représentées sous forme de noeuds, liés avec des relations possédant des propriétés, voici un graphe explicatif représentant le modèle physique des données représenté par la figure III.2

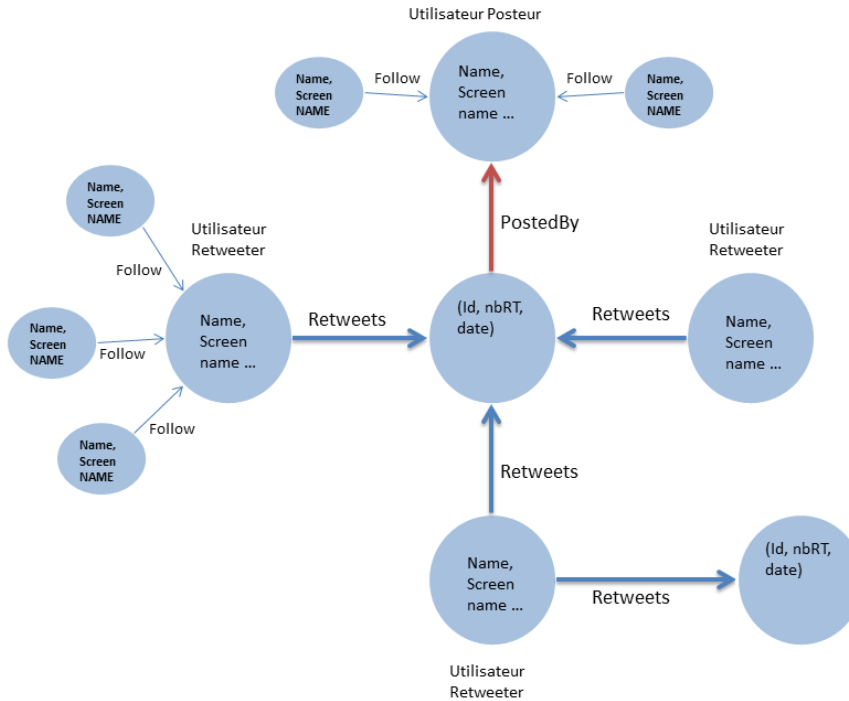


FIGURE III.2 – Le modèle physique de données.

Maintenant on expliquera les différents cas qu'on a traité pour la construction de la base de données graphe.

#### Cas I : Le tweet existe

- **Cas 1 :** L'utilisateur posteur et l'utilisateur retweeteur n'existe pas :  
Dans ce cas l'utilisateur retweeteur et posteur n'existe pas dans la base de donnée, ce dernier sera automatiquement ajouté dans la base de données. Ce cas est représenté par la figure III.3
- **Cas 2 :** L'utilisateur posteur existe et retweeteur n'existe pas :  
Dans ce cas l'utilisateur retweeteur est ajouté dans la base de données.
- **Cas 3 :** L'utilisateur posteur n'existe pas et le retweeteur existe.  
Dans ce cas l'utilisateur posteur est ajouté à la base de données.

#### Cas II : Tweet n'existe pas

- **Cas 1 :** L'utilisateur posteur n'existe pas et l'utilisateur retweeteur n'existe pas  
Dans ce cas le tweet, l'utilisateur posteur et le retweeteur vont être ajouté à la base de donnée. Ce cas est illustré dans la figure III.4.
- **Cas 2 :** L'utilisateur posteur existe, l'utilisateur retweeteur n'existe pas  
Dans ce cas le tweet, et l'utilisateur retweeteur vont être ajouté à la base de donnée.
- **Cas 3 :** L'utilisateur posteur n'existe pas et l'utilisateur retweeteur existe  
Dans ce cas le tweet, et l'utilisateur posteur vont être ajouté à la base de donnée.

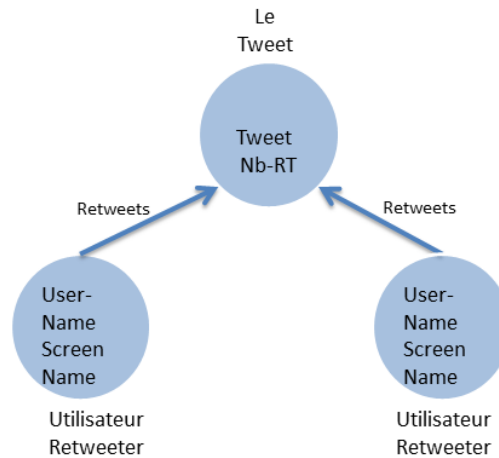


FIGURE III.3 – Nouveau utilisateur posteur, nouveau utilisateur retweeteur.

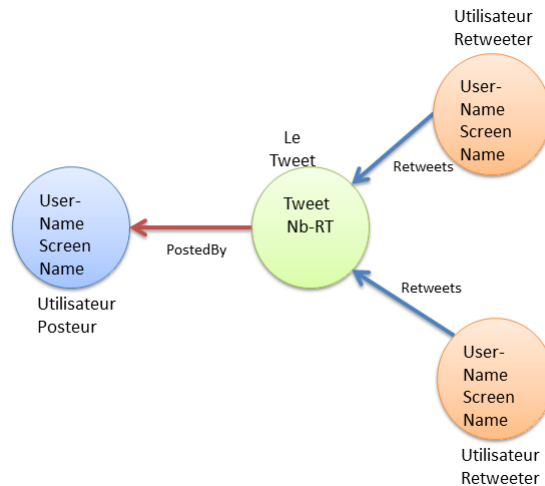


FIGURE III.4 – Nouveau tweet, nouveau utilisateur posteur, nouveau retweeteur.

### Cas III : Le tweet existe, L'utilisateur posteur existe, L'utilisateur retweeteur existe

Dans ce cas nous allons procéder à l'ajout des Follower des utilisateurs posteurs et retweeteur et cela en évitant la redondance des noeuds utilisateurs.

- **Cas 1** : Nouveau Follower de l'utilisateur retweeteur. La figure III.5 représente ce cas.
- **Cas 2** : Nouveau Follower de l'utilisateur posteur. La figure III.6 représente ce cas.
- **Cas 3** : Follower du posteur existe, Follower du retweeteur n'existe pas. Dans ce cas il existe une relation entre le Follower et l'utilisateur retweeteur mais ce meme utilisateur figure aussi parmi les followers de l'utilisateur posteur, alors une relation entre le Follower de l'utilisateur posteur et l'utilisateur retweeteur sera ajoutée. La figure III.7 représente ce cas.
- **Cas 4** : Follower du posteur n'existe pas, Follower du retweeteur existe. Dans ce cas il existe une relation entre le Follower et l'utilisateur posteur mais ce meme utilisateur figure aussi parmi les followers de l'utilisateur retweeteur, alors une relation entre le Follower de l'utilisateur retweeteur et l'utilisateur posteur sera ajoutée.

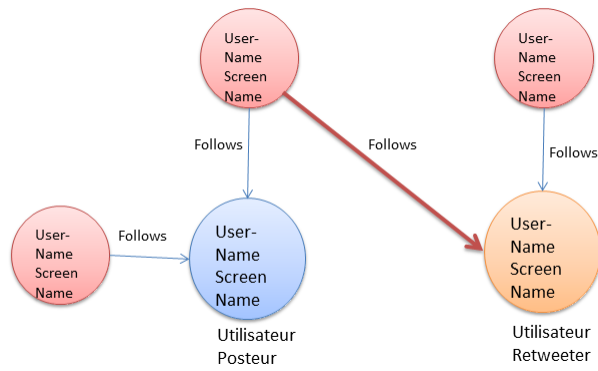


FIGURE III.5 – Nouveau Follower de l'utilisateur retweeter.

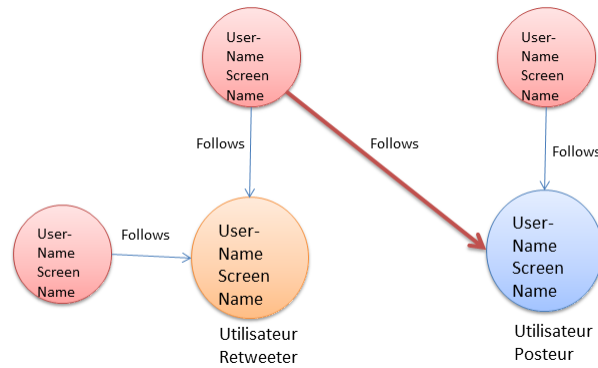


FIGURE III.6 – Nouveau Follower de l'utilisateur posteur.

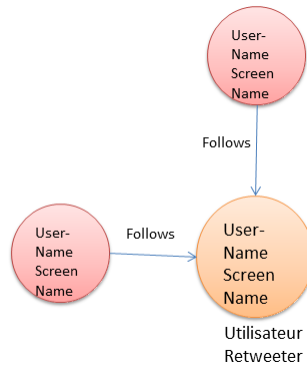


FIGURE III.7 – Follower du posteur existe, Follower du retweeter n'existe pas.

### III.3 Réalisation de notre outil de détection des influenceurs

Dans ce qui suit, nous détaillerons les différentes composantes de notre outil en présentant quelques interfaces graphiques relatives aux fonctionnalités offertes. La figure III.8 illustre l'architecture globale de notre application.

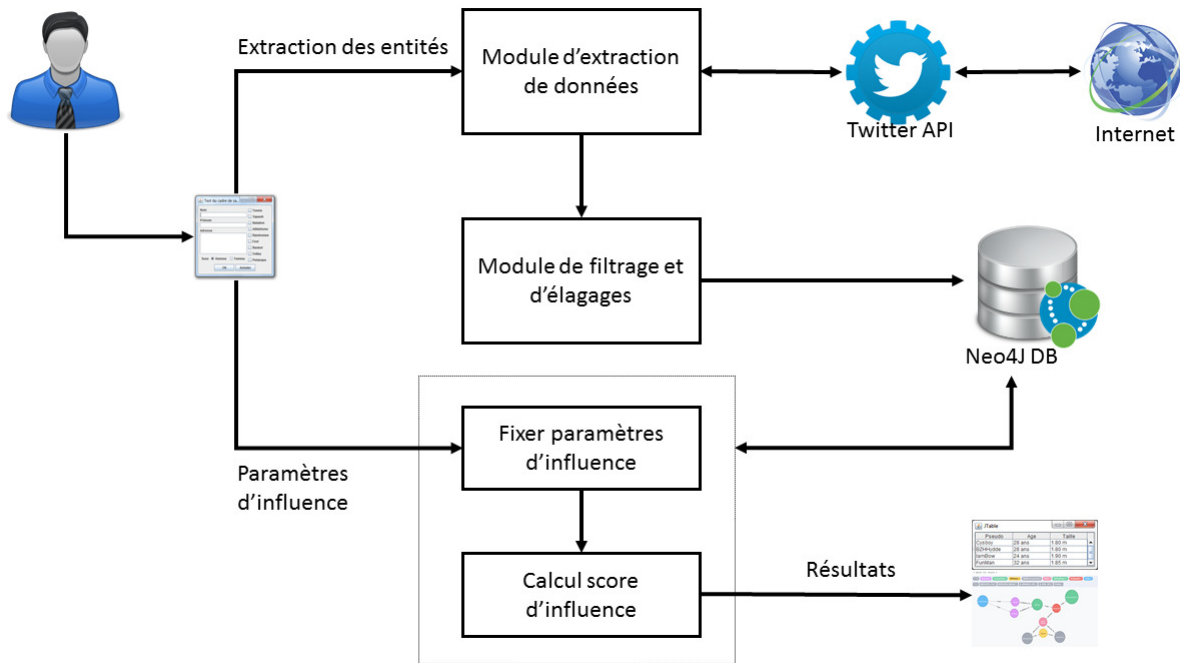


FIGURE III.8 – Architecture globale de notre application.

Notre outil se compose de deux (02) parties principales :

- La partie de construction du corpus, qui comprend les modules d'extraction de données et de filtrage et d'élagages des entités inutiles.
- La partie détection des influenceurs qui comprend le module de calcul du score d'influence ainsi que la génération du graphe des influenceurs.

#### III.3.1 La partie construction du corpus

En premier lieu, l'utilisateur doit s'authentifier, la figure III.9 illustre cette opération. Ainsi pour procéder à la recherche et à la récupération des informations à partir du réseau Twitter, l'utilisateur doit avoir des jetons d'accès à l'API de Twitter et cela après une opération d'authentification.

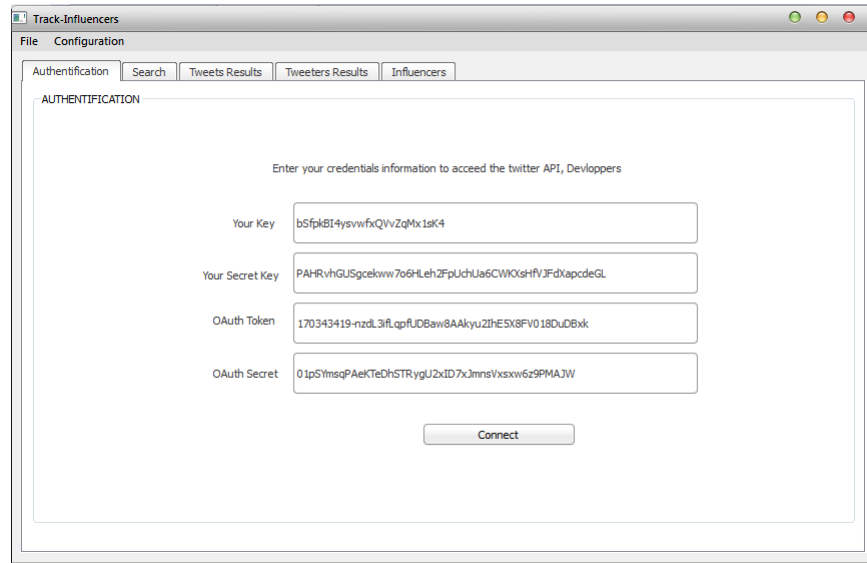


FIGURE III.9 – Interface d'authentification.

Une fois authentifier, l'utilisateur procède à la recherche et à la récupération des données, et cela suivant certain mots-clés. Pour cela, l'utilisateur devra choisir un domaine et des mots-clés relatifs au domaine sélectionné pour garantir de meilleurs résultats de recherche et d'extraction en temps réel de données à partir du réseau social. L'utilisateur a la possibilité d'arrêter l'opération de recherche et de relancer une autre pour cela des informations sur la date de dernière opération de recherche ainsi que le nombres d'entités extraites sont fournies à l'utilisateur . La figure III.10 montre cette opération.

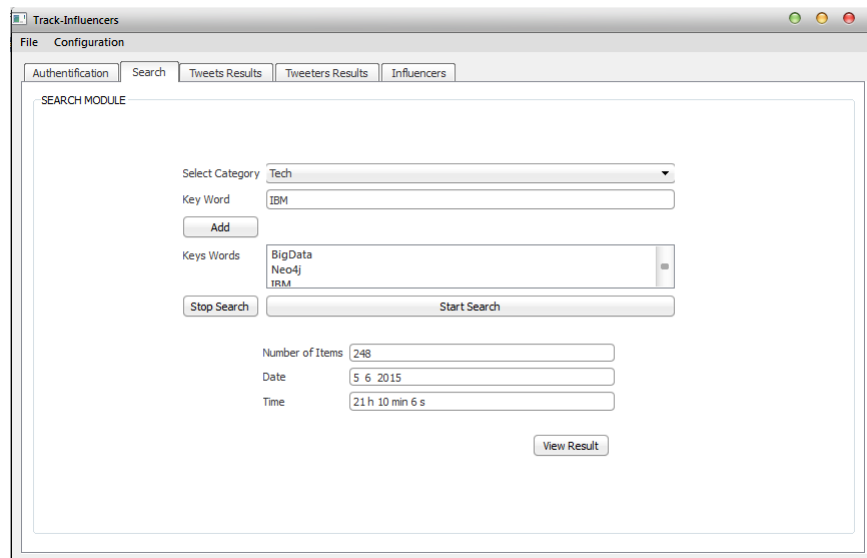


FIGURE III.10 – Interface de recherche et récupération des données du réseau social.

### III.3.2 La partie détection des influenceurs

Avant de calculer le score d'influence pour chaque utilisateur présent dans le corpus, l'utilisateur de notre outil doit fixer deux paramètres correspondant à deux critères d'influence : la popularité et l'impact.

Ces paramètres seront pris en compte pour le calcul de score d'influence des entités utilisateurs dans la base de données. La figure III.11 montre cette fonctionnalité.

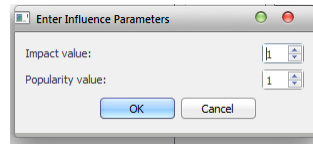


FIGURE III.11 – Réglage paramètres d'influence.

Après le calcul de score, une table regroupant l'ensemble des utilisateurs influenceurs avec leurs pourcentages d'influence évalué dans l'ensemble du corpus est affichée à l'utilisateur, comme le montre la figure III.12. Ce dernier pourra filter les résultats obtenus et choisir lui même les influenceurs qui lui conviennent et pourra par la suite contacter ces influenceurs en accédant directement à leurs profils dans le réseau social Twitter.

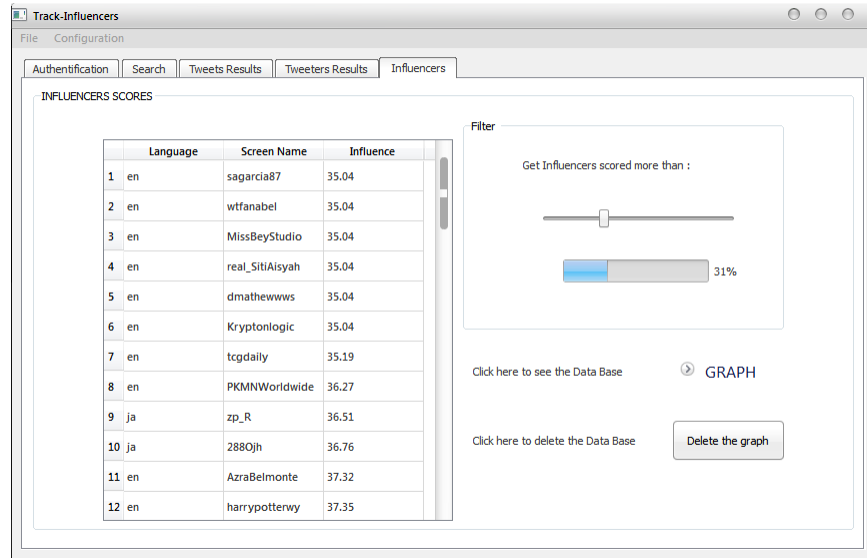


FIGURE III.12 – Affichage des résultats de calcul de score.

### III.4 Présentation de la base de données graphe

Après la construction de notre corpus, les entités récupérées, en temps réel, seront stockées sous forme de graphe dans la base de données. Et cela pour faciliter l'exploitation et la visualisation. Pour mieux exploiter cette base de données, l'utilisateur aura la possibilité d'y accéder directement, afin de visualiser les entités de notre corpus. Dans la base de données chaque noeud représente des tweets postés ou rediffusés par d'autres noeuds qui représentent les utilisateurs. Voici un aperçu de notre base de données dans la figure III.13.

### Légende :

Noeud Rosé : Utilisateur qui a posté le tweet, dit Utilisateur posteur.

Noeud Bleu : Tweet .

Noeud Violet : Utilisateur qui a retweeté le tweet, dit Utilisateur retweeteur.

Noeud Jaune : Utilisateur suiveurs de Utilisateur posteur.

Noeud Orange : Utilisateur suiveur de Utilisateur retweeteur.

Flèches : Les relation entre Noeuds : « PostedBy », « RetweetedBy », « Followed By ».

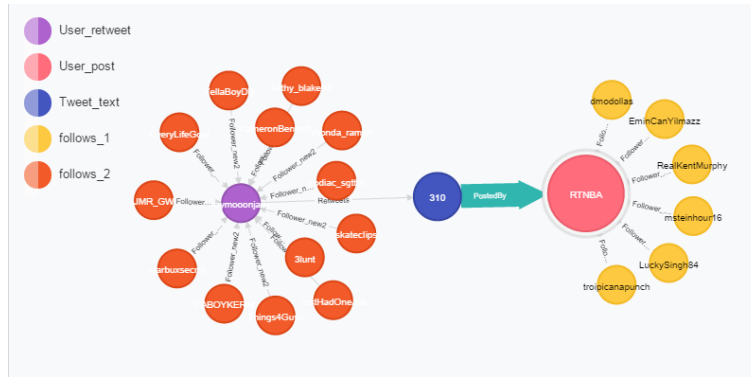


FIGURE III.13 – Aperçu de la base de données Neo4j.

## III.5 Conclusion

Dans cette partie de notre document, nous avons présenté les outils utilisés durant notre réalisation, ainsi que la technique d'authentification utilisée par Twitter pour l'extraction des données du réseau. Nous avons aussi présenté notre base de données graphe ainsi que des captures de notre outil de détection d'influenceurs.



# Conclusion générale

Nous clôturons notre modeste travail par un bilan général de ce qui a été réalisé et sur ce qui reste à réaliser comme perspectives.

Le système que nous avons proposé va permettre à des décideurs d'entreprise ou des spécialistes en marketing de collecter des données issues du réseau social, de les consolider dans une base de donnée orientée graphe dans le but de détecter des potentiels influenceurs qui pourront devenir des futurs ambassadeurs pour une marque ou une entreprise. Dans le cadre de notre travail, nous nous sommes intéressées au réseau Twitter car très utilisé dans les campagnes marketing, il dispose d'un apanage d'utilisateurs actifs qui prodiguent à la marque ou à l'entreprise une plus grande visibilité à travers leurs publications sur le réseau social.

La première étape de la réalisation de notre outil fut l'étude de l'existant. Après une brève synthèse bibliographique dans le domaine du marketing, nous avons découvert que toute entreprise qui désire lancer un produit donné opte de le faire en premier lieu dans les médias sociaux. Ces derniers considérés comme un atout qui garantit une grande visibilité au niveau national voir international.

Dans la même idée, nous avons pu ajusté la solution entreprit par les entreprises en ciblant un type des médias sociaux, à savoir : les réseaux sociaux, et ceci dans le but d'offrir une solution rapide et efficace aux entreprises dans leurs processus de lancement de leurs produits. Durant cette étape nous avons recensé des études qui ont été faites dans le domaine de l'analyse des réseaux sociaux. Ces derniers regorgent d'individus de toute sorte, nous avons choisis de centrer notre étude sur la détection d'individus influenceurs, considérés comme un levier de propagande et de persuasion d'idée et d'avis aux sein des réseaux sociaux.

De ce fait, nous avons proposé une approche de recherche d'influenceurs en temps réel dans un domaine d'activité précis. Nous avons développé un outil permettant de : rechercher, récupérer, stocker, analyser et présenter les informations et données issues du réseau Twitter. A travers l'utilisation d'APIs, nous avons pu extraire un ensemble d'informations relatives aux publications et aux profils utilisateurs. La recherche et la récupération de publications se fait par mots-clés selon un domaine. L'analyse des profils utilisateurs se fait dans le but de quantifier leur influence dans le réseau et de leur affecter un score.

Pour conclure, ce projet nous a permis, de mettre en pratique les connaissances acquises durant notre cursus et d'acquérir un nouveau savoir dans le domaine de la recherche d'information. Il nous a permis également de :

- Nous initier dans le domaine de l'analyse des réseaux sociaux.

- Manipuler de grand volume de données issue des réseaux sociaux comme Twitter.
- Consolider et développer nos compétences en programmation avec le langage Python et dans la modélisation avec UML.
- De connaître de nouveaux concepts liés à la gestion du processus de marketing d'une entreprise.

## Perspectives

Bien que notre outil soit fonctionnel, il est toujours sujet à des améliorations, ainsi comme perspectives de notre travail, nous proposons :

- Génération du graphe des influenceurs.
- Ajouter un module de catégorisation automatique de texte.
- Ajout d'un module d'analyse des sentiments afin de déterminer le type d'influence, influence positive et influence négative.
- Etablir un système de recommandation d'influenceurs.
- Développer une application web et mobile pour faciliter la portabilité et la flexibilité de l'outil pour les spécialistes en marketing.
- Exploitation d'autres réseaux sociaux.

# Bibliographie

- [alan J.-Ph. Vignolles A., 2010] alan J.-Ph. Vignolles A. (January 2010). Identification des leaders d'opinion sur internet : utilisation des données secondaires issues de twitter.
- [Arezki, ] Arezki, H. *Recherche d'Information : un modèle de langue combinant mots simple et mots composés*. PhD thesis, Université Mouloud Mammeri Tizi-Ouzou.
- [Augure, ] Augure. Guide pour mettre en place une stratégie d'influence.
- [B., 2009] B., J. A. . S. X. . F. T. . T. (2009). *Advances in Web Mining and Web Usage Analysis*, volume p. 118–138. Springer-Verlag, Berlin, Heidelberg.
- [BADACHE, 2012] BADACHE, I. (2012). *Master 2 ASIC, Memoire De Stage, RECHERCHE D'INFORMATION SOCIALE*. PhD thesis, Université Paul Sabatier.
- [Ben Jabeur, 2013] Ben Jabeur, L. (2013). Leveraging social relevance : using social networks to enhance literature access and microblog search.
- [Boughanem, ] Boughanem, L. B. J. . L. T. . M. Un modèle de recherche d'information sociale dans les microblogs : cas de twitter.
- [CSCQ, 2007] CSCQ (2007). Fiche technique 18 : Comment lire facilement les rapports de cqe du cscq.
- [Dewing, 2013] Dewing, M. (3 février 2013). *Les médias sociaux - introduction*.
- [E. and P., 1955] E., K. and P., L. (1955). The part played by people in the flow of mass communications. *New York : The Free Press*.
- [Ellison, 2008] Ellison, D. M. B. . N. B. (2008). Social network sites : Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 22.
- [et Dodds P.S., 2007] et Dodds P.S., W. D. (2007). Networks, influence, and public opinion formation. *Journal of Consumer Research*.
- [F. Filliettaz, 2011] F. Filliettaz, M. G. (septembre 2011). Comprendre les réseaux sociaux numériques. un enjeu pour l'enseignement. 17.
- [Fatmi, 2010] Fatmi, A. (2010). *Les avantages et les inconvénients du marketing viral*.
- [France, 2010] France, I. (2010). *Livre Blanc : Les médias sociaux*.
- [G. ERETEO, 2009] G. ERETEO, F. GANDON, M. B. . P. G. (2009). Analyse des reseaux sociaux et web semantique : un etat de l'art knowtex.

- [Gaertner, 2003] Gaertner, P.-A. M. . N. (11 décembre 2003). *Modélisation objet avec UML*. Best of Eyrolles.
- [Gummadi, 2010] Gummadi, M. C. . H. H. . F. B. . K. P. (May 2010). Measuring user influence in twitter : The million follower fallacy.
- [Hanneman and Riddle, 2005] Hanneman, R. A. and Riddle, M. (2005). Introduction to social network methods.
- [HEINRICH, 2012] HEINRICH, L. (2012). Not only sql.
- [I., 2010] I., O. I. . M. C. . L. J. . S. (November 2010). Overview of the trec 2011 microblog track. *Text REtrieval Conference TREC*.
- [Jonathan, ] Jonathan, D. Comment réussir sa stratégie de marketing viral en remplaçant le consommateur au cœur de la communication ?
- [L., 2004] L., V. E. . F. (2004). Communiquer avec les leaders d’opinion en marketing, comment et avec quels médias. *Décisions Marketing*, 35.
- [McGill, 1986] McGill, G. S. . M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc. New York, NY, USA.
- [Merklé, 2004] Merklé, P. (2004). *Sociologie des réseaux sociaux*. Paris : la Découverte.
- [PEPIN, ] PEPIN, A. S. . P. S. . L. Twitter : Extraction, regroupement et visualisation pour la veille strategique.
- [Tarquinio, 2006] Tarquinio, C. (2006). *Les concepts fondamentaux de la psychologie de la santé*. Paris. Dunod.
- [UNG, 2010] UNG, R. (Septembre 2010). Les mutations du rôle et des missions du planning stratégique face à l’essor du marketing d’influence.
- [VIGNOLLES, 2012] VIGNOLLES, E. V. . L. B. . J.-P. G. . A. (January 2012). Le rôle et l’identification des leaders d’opinion dans les réseaux sociaux traditionnels et virtuels : controverses marketing et pistes de recherche.
- [Wilder, 1977] Wilder, J. T. . J. (1977). Exploratory data analysis.

# Webographie

- [1] Cambridge advanced learners dictionary. <http://dictionary.cambridge.org/dictionary/british/social-media>.
- [2] Définition influence. [http://fr.wikipedia.org/wiki/Influence\\_\(psychologie\)](http://fr.wikipedia.org/wiki/Influence_(psychologie)).
- [3] Définition influenceur. <http://www.definitions-marketing.com/Definition-Influenceur-ou-Influencer>.
- [4] Définition leader d'opinion. <http://www.definitions-marketing.com/Definition-Leader-d-opinion>.
- [5] Définition marketing digital. <http://www.definitions-marketing.com/Definition-Marketing-digital>.
- [6] emarketer : Market research on digital media, internet marketing. <http://www.emarketer.com>.
- [7] Facebook statistics directory. <http://www.socialbakers.com/statistics/facebook/>.
- [8] Introduction au langage python. <http://www.jchr.be/python/manuel.htm>.
- [9] Klout. <http://www.surfandbiz.com/article/img/klout-influence-matrix.jpg>.
- [10] Le nosql. <http://www-igm.univ-mlv.fr/~dr/XPOSE2010/Cassandra/nosql.html>.
- [11] Médias sociaux. [http://fr.wikipedia.org/wiki/Medias\\_sociaux](http://fr.wikipedia.org/wiki/Medias_sociaux).
- [12] The search api. <https://dev.twitter.com/rest/public/search>.
- [13] Twitter usage statistics. <http://www.internetlivestats.com/twitter-statistics/>.
- [14] Voir les sites. [www.skype.com](http://www.skype.com), <https://www.viber.com>, <https://www.whatsapp.com>.
- [15] Wikipédia : Statistiques. <http://fr.wikipedia.org/wiki/Wikipedia:Statistiques>.
- [16] Pierre Bertucat. Pourquoi le marketing viral fonctionne? <http://www.pierrebertucat.com/le-marketing-viral-fonctionne>.
- [17] Rémi Giraudier. Un compte twitter certifié, comment ça marche? <http://blog.neocamino.com/un-compte-twitter-certifie-comment-ca-marche/>.
- [18] David Lefèvre. Internet : Pourquoi utiliser le marketing viral? <http://www.kalipub.com/blog/weborama/10-raisons-d-utiliser-le-marketing-viral.html>.
- [19] NICOLAS RAULINE. Les réseaux sociaux attirent les investissements publicitaires. [http://www.lesechos.fr/journal20150309/lec2\\_high\\_tech\\_et\\_medias/](http://www.lesechos.fr/journal20150309/lec2_high_tech_et_medias/)

0204208946033-les-reseaux-sociaux-attirent-les-investissements-publicitaires-1100109.php#.

- [20] Todd Wasserman. Man buys promoted tweet to complain about british airways. <http://mashable.com/2013/09/02/man-promoted-tweet-british-airways/>.