الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université de 8 Mai 1945 – Guelma -

Faculté des Mathématiques, d'Informatique et des Sciences de la matière

Département d'Informatique

Mémoire de Fin d'études Master

Filière : Informatique

Option : Système Informatique

Thème :

Vers une détermination des vrais influenceurs sur les réseaux sociaux

Encadré Par :

Dr. Abdelhakim Hannousse

Présenté par :

Bilal Aib

Juillet 2019

Remerciements

Tout d'abord, nous remercions ALLAH le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce Modeste travail.je dois remercier mes parents, mes frères pour leur soutien inconditionnel et leur altruisme tout au long de ces années, et particulièrement pendant mon projet fin d'étude.

Je dois remercier mon encadreur, le docteur Abdelhakim Hannousse, pour leur soutien, leur motivation, leur patience et leur disponibilité. C'est très réconfortant de pouvoir partager des idées et discuter ouvertement de nouvelles façons d'aborder un problème avec une telle facilité.

Je remercie tous les collègues et amis proches qui m'ont accompagné tout au long de ces années, et en particulier ceux qui ont rempli ces deux dernières années de tant de joie, de rires et de camaraderie Donc, à Abdou Guergour, Aymen Boukeskesse et Djabali Lounis.

Ces remerciements vont tout d'abord au corps professoral et administratif de l'université 8 Mai 1945 de Guelma, plus particulièrement au corps professoral du département de l'informatique, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée.

Résumé

Les réseaux sociaux sont devenus rapidement l'outils de communication et de partage d'information les plus indispensables dans la vie quotidienne des millions des personnes dans le monde. Dans un réseau social en trouve des membres qui diffusent des informations et d'autres qui les évaluent et les rediffusent. Cela permet de diffuser le maximum des informations dans un temps et coûts très réduits. Les premiers membres source des informations sont appelés « influenceurs ». La détection des influenceurs est très importante et utile dans différents domaines. En marketing, par exemple, la détection des influenceurs permet de diminuer le coût de publicité en payant un nombre réduit de membres pour partager des avis positifs concernant un produit particulier.

Dans le cadre de ce mémoire, nous menons une étude sur l'identification des internautes influenceurs en Twitter. Pour cet objectif, nous contribuons par la proposition d'un algorithme basé sur le taux de rediffusions des tweets (retweets) capturant le flux d'informations dans le réseau. En réalité, un internaute peut rediffuser un message tel qu'il est, ou rediffuser un message après certaines modifications. Dans ce contexte nous proposons un algorithme qui prend en considération pas seulement le degré de centralité des membres mais aussi le contenu des tweets diffusés. En d'autres termes, en considère un membre comme influenceur s'il est le premier qui a partagé un sujet dans un tweet qui a été apparait dans la plus longue liste des amis et suiveur dans le graph dans une période du temps donné.

Mots-clés : Réseaux sociaux, Détection des influenceurs, Twitter.

Table des matières

Table de figures

Chapitre 1.

Chapitre 2.

Chapitre 3.

Chapitre 4.

Liste des tableaux

Introduction Générale

Un réseau social est une plateforme en ligne où les internautes échangent et partagent l'un avec l'autre des expériences, des actualités, des photos, des blogs, etc…

Parmi les caractéristiques des réseaux sociaux, qui les rendent différents des autres plateformes de communication traditionnelles, sont : le suivi des activités des internautes, la rediffusion des informations, le fait d'aimer et de commenter un post. Aujourd'hui, de nombreuses entreprises utilisent les réseaux sociaux comme un moyen essentiel de commercialisation de leurs produits. Vue le développement rapide des réseaux sociaux en termes de quantité d'informations diffusées et de nombre de ses membres, le développement des méthodes d'analyse de ces réseaux a suscité beaucoup d'intérêt et de curiosité de la part des chercheurs académiques en sciences politiques, économiques, sociologie et informatique. Une grande partie de cet intérêt peut être attribuée à l'attrait de l'analyse des réseaux sociaux en termes de relations entre les internautes, à l'étude de l'autorité et de l'influence sur les réseaux sociaux, ainsi qu'aux modèles et implications de ces relations. Chaque réseau social est constitué logiquement d'un ensemble d'utilisateurs connectés les uns aux autres.

Chaque utilisateur a la possibilité d'étendre son influence à travers le réseau. L'influence propagée par certains utilisateurs est plus grande que d'autres. Il y a eu beaucoup des travaux récents sur l'analyse des réseaux sociaux en concentrant explicitement sur la sociométrie, y compris les mesures quantitatives de l'influence, de l'autorité et de la centralité des internautes.

Dans le cadre de ce mémoire, nous menons une étude sur l'identification des internautes influenceurs en Twitter. Pour cet objectif, nous contribuons par la proposition d'un algorithme basé sur le taux de rediffusions des tweets (retweets) capturant le flux d'informations dans le réseau. En réalité, un internaute peut rediffuser un message tel qu'il, ou rediffuser un message après certaines modifications. Cela a un impact majeur sur la détection des vrais flux d'informations dans le réseau. Dans l'algorithme qui nous propose, les textes, des tweets sont d'abord analysés pour extraire des listes des sujets discutés dans chaque tweet. Cela nous permet de détecter

la rediffusion des tweets sémantiquement similaires au lieu de concentrer sur l'aspect syntaxique des tweets. Vue la difficulté et le temps nécessaire pour la collection des données depuis Twitter, nous avons utilisé des data sets disponibles sur le net pour tester notre algorithme. Les résultats obtenus sont comparés avec les degrés de centralités fréquemment utilisés pour la détection des influenceurs dans les réseaux sociaux.

Ce mémoire est divisé en trois chapitres. Dans le premier chapitre, nous introduisons les concepts fondamentaux des réseaux sociaux. Dans le deuxième chapitre, nous discutons les différentes approches et techniques utilisées dans la détection des influenceurs en Twitter. Dans le troisième chapitre, nous présentons les différentes étapes de notre algorithme ainsi que les résultats obtenus.

Chapitre I

Les réseaux sociaux

Chapitre I

Les réseaux sociaux

Ce chapitre présente les principaux concepts des réseaux sociaux tels que la définition et le rôle qui jouent les réseaux sociaux dans la vie quotidienne, les types des réseaux sociaux et les différentes représentations possibles des réseaux.

1. Qu'est-ce qu'un réseau social ?

Les réseaux sociaux ont été découverts aux États-Unis en 1995, il s'agissait d'un service de réseautage social appelé Classmates [24]. Cependant ils n'ont été connus par tous les continents qu'en 2004. Les réseaux sociaux sont développés sur Internet à partir du début du XXIème siècle suite à l'apparition des nouvelles technologies numériques comme Facebook. L'année 2005 connue le lancement du Youtube comme une première plateforme de partage de vidéos. En 2006 la première plateforme de microblogging Twitter est apparait [29]. Depuis, Internet a révolutionné le monde des ordinateurs et des communications comme rien d'autre auparavant. Internet, aujourd'hui, est à la fois une capacité énorme de diffusion dans le monde entier, un mécanisme de distribution de l'information et un moyen de collaboration et d'interaction entre les individus par le biais de leurs ordinateurs, peu importe leurs emplacements géographiques.

Un réseau social est une composition sociale qui vous permettent de communiquer avec des amis, des lieux, des organisations ou des individus. Tous les amis peuvent être connectés à un ou plusieurs autres amis pour construire éventuellement une structure qui représente le lien social de cet individu. Dans lesquels les nœuds représentent les individus et les arcs représentent une forme d'interaction sociale entre les nœuds, notamment l'amitié, la parenté, les informations relatives à une relation, etc…

1.1. Pour quoi utiliser les réseaux sociaux ?

Les réseaux sociaux est un moyen indispensable dans notre société actuelle, et dans nos relations avec les autres. Les réseaux sociaux visent à créer un tissu relationnel ; ils permettent de rester disponible avec la famille, les amis (qu'ils soient anciens ou nouveaux), les clients (dans le milieu professionnel ou même au sein d'une même entreprise) ainsi que les fans (pour les différentes célébrités). Ce sont des outils idéaux pour envoyer des messages, partager des idées avec une grande communauté de personnes appelée parfois "amis" par exemple sur Facebook.

Il est possible aussi d'utiliser les réseaux sociaux dans le domaine commercial ou en marketing pour vendre des produits ou des services et bien évidemment, augmenter vos ventes sera sûrement votre objectif prioritaire.

Les réseaux sociaux permettent aussi de travailler à votre réputation pour gagner des nouveaux visiteurs connaissent et considèrent votre marque et l'image qu'elle dégage, ils interagiront plus facilement sur vos réseaux sociaux. C'est là que le bouche à oreille pourra commencer à faire son travail et que les internautes vous recommanderont à leurs connaissances.

1.2. Les différents types des réseaux sociaux

Les réseaux sociaux peuvent être classés selon différentes types, chacun a son utilité et son spécialité :

1.2.1. Les réseaux personnels et généralistes

Ce type est tourné vers des sujets généraux et divers (sport, musique, politique, etc.). L'objectif est de faire partager des passions et des idées avec le reste de la communauté.

Facebook

Facebook est le site de réseautage social le plus populaire sur Internet créée en 2004 par Mark Zuckerberg. Au premier trimestre 2019, Facebook comptait plus de 2,38 milliards d'utilisateurs actifs chaque mois et 1,56 milliard d'utilisateurs actifs chaque

jour dans le monde[1]. Chaque internaute peut inscrivez et faire un réseau illimité d'amis (des personnes proches ou même inconnus) qu'il accepte. Il est aussi utilisé par les entreprises, les artistes et les politiciens [37], il permet de partager des statuts, des photos, des liens, des vidéos et d'envoyer des messages directs, des appels vocaux ou vidéos [28].

Twitter

Twitter est réseau sociale de microblogging créée en 2006 et qui permet d'envoyer des messages courts appelés *tweets* diffusés par les membres inscrits aux internautes qui suivent chaque compte. Depuis 2017 la taille d'un tweet est passée de 140 à 280 caractères. Twitter est classé parmi les réseaux les plus populaires avec des milliers de nouveaux utilisateurs chaque jour. Au premier trimestre 2019, Twitter compte 330 millions d'utilisateurs actifs mensuels dans le monde [29].

En Twitter comme en Facebook, on distingue deux types de relations entre les différents membres de réseaux. Les amis (*friends en anglais*) et les suiveurs (*followers en anglais*). Les amis peuvent partager, les uns avec les autres, toutes leurs conversations, par contre un suiveur à un accès limité autorisé par la personne qui suive ; donc un partage limité à un seul sens. Pour joindre une conversation, un utilisateur en Twitter utilise simplement l'Hashtag (#) pour participer à un sujet de conversation ; l'hashtag est utilisé comme un tag rattaché au mot clé du sujet de conversation.

1.2.2. Les réseaux de partage

Ce sont des plateformes dédiées au partage des multimédias (photos, sons, vidéos, etc.) entre internautes. L'objectif est de faciliter l'accessibilité aux sources multimédias pour les internautes d'une communauté. Parmi ces plateformes les plus populaires on

[1] https://www.journaldunet.com/ebusiness/le-net/1125265-nombre-d-utilisateurs-de-facebook-dans-le-monde/

peut citer Youtube [26] (une plateforme pour envoyer, regarder, commenter, évaluer et partager des vidéos) et FlikR (une plateforme pour partager des photos) [30].

1.2.3. Les réseaux professionnels

Ce sont les réseaux les plus performants au sens propre du terme. Ils offrent la possibilité de se connecter et de partager des informations en mode professionnelle. Parmi ces réseaux en peut citer LinkedIn qui permet de publier et partager des curriculums vitæ (CV) et de chercher d'embauche dans les entreprises et les organisations qui publient leurs annonces de travail dans le réseau [31].

1.2.4. Les réseaux personnels thématiques

Les réseaux personnels thématiques peuvent être vue comme des réseaux généralistes mais sont orientés autour d'une thématique (voiture, musique, etc.)

Les modèles de représentation des réseaux sociaux

Une première étape pour l'analyse des réseaux sociaux consiste à utiliser une représentation adéquate. Il existe différents modelés de représentation des réseaux sociaux. Dans cette section nous présentons quelques modèles les plus répandus.

2. Représentations des graphes

2.1. Représentation graphique

Un réseau social peut être représenté sous forme d'un graphe constitué des nœuds et des arcs. Chaque nœud modélise un utilisateur ou membre de réseau et un arc modélise une des relations disponibles par le réseau (amitié, suivis, etc.). Vue le nombre énorme des utilisateurs, cette représentation, bien qu'il soit clair pour les êtres humains, elle ne peut pas être utilisé efficacement par les ordinateurs ou manipulés efficacement à l'aide des outils mathématiques. Pour cela, différentes autres représentations cherchent à obtenir des fromes qui peuvent stocker l'ensembles de nœuds et de bords d'une manière qui facilite la manipulation par ordinateurs.

2.2. Matrice d'adjacence

La matrice d'adjacence est une matrice carrée (aussi appelée matrice de contiguïté) qui permet de représenter un graph fini G par une relation entre ses différents nœuds. Dans la matrice d'adjacence, une valeur de 1 indique une connexion entre les nœuds Vi et Vj de graph, et un 0 indique qu'il n'y a pas de relation directe entre les deux nœuds [2].

$$M=\begin{bmatrix}0 & 0 & 1 & 0 & 0 & 1\\ 1 & 0 & 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 1 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 1\\ 1 & 0 & 1 & 0 & 1 & 0\end{bmatrix}$$

2.3. Liste d'adjacence

Une liste d'adjacence est un ensemble de listes utilisées pour représenter un graph fini où chaque nœud est lié à une liste de tous les nœuds qui y sont liés. Pour chaque nœud $v \in V$ nous maintenons une liste des nœuds adjacents à v. Avec une liste d'adjacence, nous pouvons éliminer les relations inutiles (les valeurs 0) qui se présentes dans la matrice d'adjacence et d'accéder directement à la liste de tous les nœuds reliés à un nœud particulier par une entrée correspondante [2].

From	**bords**
Bilal	Sara
Wassim	Mary
. . .	. . .

2.4. Liste des arcs

Cette liste contient la liste des arcs du graph G. C'est une autre façon simple et courante pour stocker de grands graphs est d'enregistrer tous les bords du graph [2].

(Bilal, Sara)

(Sara, Bilal)

(Bilal, Wassim)

(Wassim, Mary)

(Mary, Wassim)

3. Les différents types de graphes

Il existe différents types de graphs de base. Dans cette section, nous présentons ces différents types.

3.1. Graphe

Un graph G = (V, E) contient un ensemble de nœuds V et de arcs E. Les nœuds sont reliés les uns aux autres par un ensemble de liens appelés arcs. Dans un réseau social, chaque nœud désigne un membre de réseau et un arc entre deux nœuds désigne une relation entre deux membres de réseau comme le montre la figure 1.

3.2. Graphe nul

Un graph nul est un graph où l'ensemble de ces nœuds est vide. Évidemment, puisqu'il n'y a pas de nœuds, il n'y a pas non plus des arcs. Formellement, G (V, E), V =E =0.

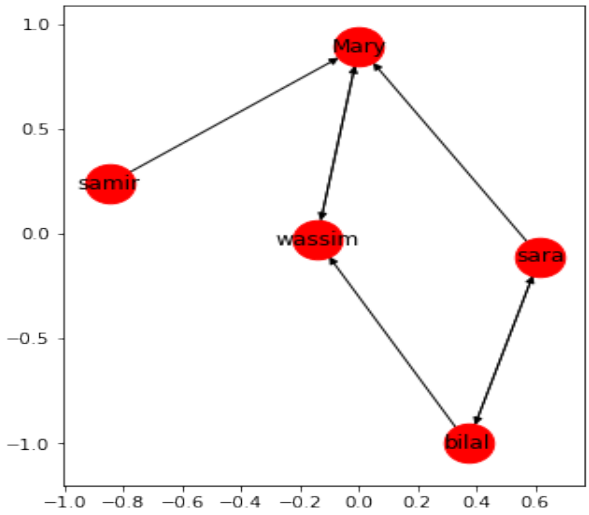


Figure 1. Exemple d'un graph représentant un réseau social

3.3. Graphe vide

Un graph vide est un graphe dont l'ensemble de ses arcs est Vide : G (V, E), E =0. Notez que l'ensemble des nœuds peut être non vide. Donc, un graph nul est un graph vide mais pas l'inverse [6].

3.4. Graphe orienté, non orienté et mix

Les graphes qui n'ont que des arcs dirigés sont appelés des graphes orientés et ceux qui n'en ont que des arcs non dirigés sont appelés des graphes non orientés. Les graphes mixtes ont des arcs orientés et non orientés. Dans les graphes orientés, nous pouvons avoir deux arcs entre Vi et Vj (un de Vi à Vj et un de Vj à Vi), alors que dans les graphes orientés, un seul arc peut exister. Par conséquent, la matrice d'adjacence des graphes orientés n'est pas en général symétrique (Vi connecté à Vj ne signifie pas que Vj est connecté à Vi, c'est-à-dire la valeur de Ai,j peut être différentes à Aj,i), alors que la matrice d'adjacence des graphes non orientés est symétrique ($A = A^T$) [6]. Dans les réseaux sociaux, il existe de nombreux réseaux orientés, non orientés et mixtes. Par

exemple, Twitter est un réseau mixte, où les relations suiveuses ne sont pas bidirectionnelles alors que les relations d'amitiés sont bidirectionnelles.

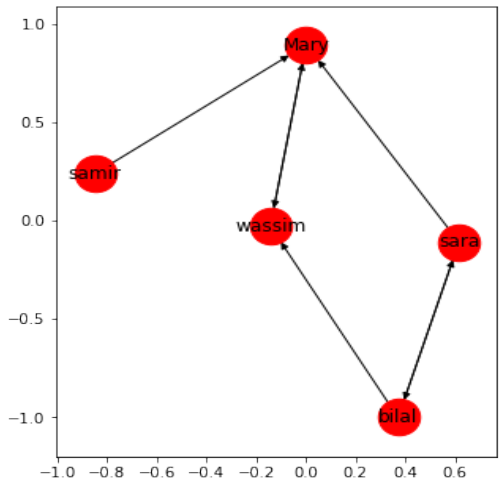


Figure 2. Exemple d'un graph orienté

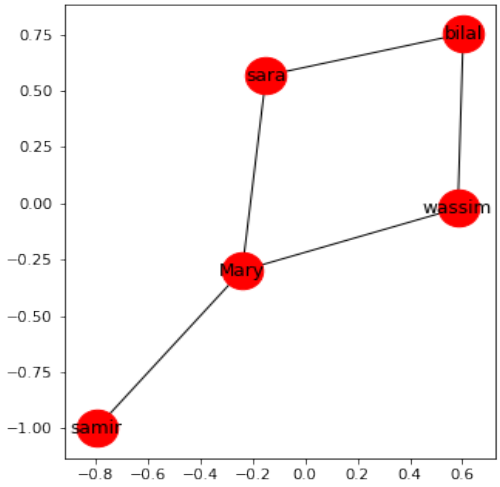


Figure 3. Exemple d'un graph non orienté

3.5. Graphe pondéré

Les graphes pondérés sont des graphes ayant des poids ou des valeurs associées aux arcs. Les poids des arcs peuvent représenter un concept tel que le coût de connexion, la longueur, la capacité, la similarité, la distance, etc., qui dépend de l'utilisation spécifique de ce graph. Dans un réseau social, un poids peut désigner, par exemple, la fréquence de communication entre deux membres d'un réseau [2].

L'arc avec une linge continue = poids >0.5].

L'arc avec une ligne pointillée =poids <= 0.5].

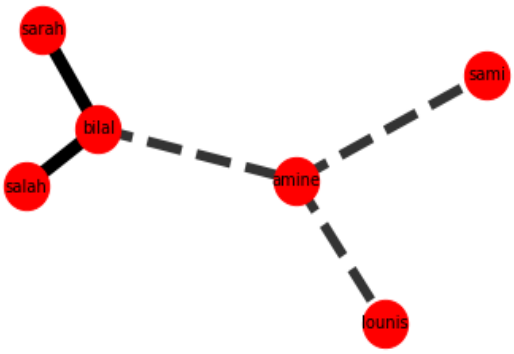


Figure 4 Exemple d'un graph pondéré

3.6. Graphe simple

Un graphe simple est un graphe non orienté sans boucles ni arcs multiples.

3.7. Graphe régulier

Un graphe régulier (ou uniforme) est un graph dans lequel tous les nœuds ont le même nombre de voisins, qui est le degré de chaque nœud [6].

3.8. Graphe complet

Un graphe complet est un graph dans lequel n'importe quelle paire de nœuds est connectée [2].

3.9. Graphe connecté et déconnecté

Un graphe est appelé connecté si chaque paire de ses nœuds distincts sont connectés ; sinon, on parle d'un graphe déconnecté [6].

4. Propriétés d'un réseau social

Les réseaux sociaux sont des graphes particuliers possédant des propriétés spécifiques. Ces propriétés sont présentées dans cette section.

4.1. Réseau invariant d'échelle

Les réseaux sociaux sont des graphes dont la distribution des degrés obéit à une loi de puissance. Cela veut dire que dans un réseau social, un nombre limité des membres sont bien connectés, alors que la majorité des membres sont mal connectés [1].

$$P(k) \sim K^{-\gamma}$$

La distribution des degrés est calculée comme suit :

$$PK = 1 / N \#\{I|Ki = k\}$$

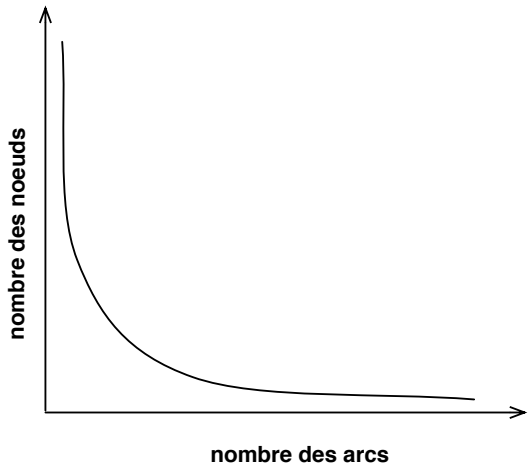


Figure 5. Réseau invariant d'échelle

4.2. Étude du petit monde

Dans les réseaux sociaux, la plupart des nœuds sont homogènes et peuvent être atteints par un petit nombre d'étapes. Cela signifie qu'un réseau social se composent des communautés denses qui se relier de façon lâche par des liens indirects [2].

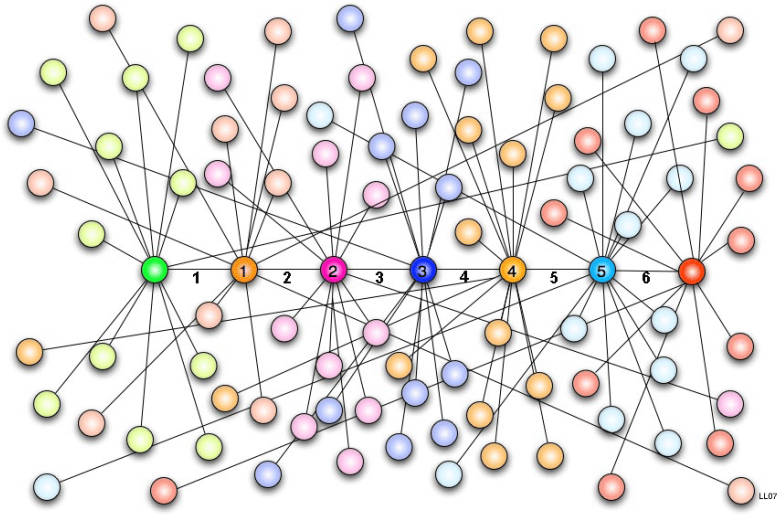


Figure 6. Exemple d'un réseau du petit monde[2]

5. Mesures sur les réseaux sociaux

Dans cette section nous présentons les différentes mesures utiles pour l'analyse des réseaux sociaux pour la détection des nœuds importants. Ces mesures sont appelées les mesures de centralité.

5.1. Degré de centralité

Nous considérons souvent que les personnes ayant de nombreux liens sont importantes. Le degré de centralité transfère la même idée dans une mesure. La mesure

[2] Image libre disponible sur Wikipédia : https://commons.wikimedia.org/wiki/File:Six_degrees_of_separation_01.png

du degré de centralité classe les nœuds ayant le plus grand nombre de connexions. Le degré de centralité Cd pour un nœud vi dans un graphe non orienté est :

$$C_d(v_i) = d_i \text{ (di est le nombre des arcs reliés à vi)}$$

di dans la formule désigne le nombres des arcs reliés à v_i dans le graphe G. Cette valeur peut être normalisée simplement par le biais de sa division sur nombre total des nœuds (n) dans le graphe.

$$C_d^{norm}(v_i) = \frac{d_i}{n-1}$$

5.2. Centralité intermédiaire

Tandis que le degré de centralité se base sur le nombre des amis qu'un membre de réseau possède, la centralité intermédiaire se base sur le concept de l'accessibilité de la personne. La centralité intermédiaire repose sur l'idée qu'une personne est plus importante si elle est plus intermédiaire dans le réseau. Cette mesure est basée sur la notion de géodésie, ce qui signifie qu'un acteur peut devenir davantage important dans le réseau s'il est situé sur les géodésiques entre plusieurs paires des acteurs du réseau.

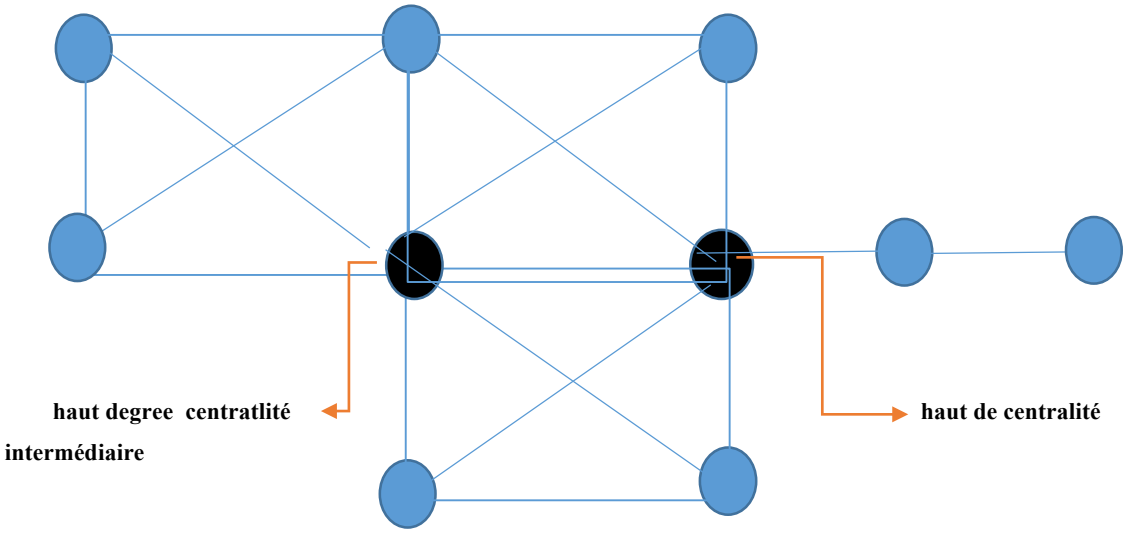


Figure 7. Exemple sur le degré de centralité des nœuds

5.3. Centralité de proximité

Une autre mesure de centralité est la centralité de proximité. L'intuition de cette mesure est que plus les nœuds centraux sont nombreux, plus ils peuvent atteindre rapidement les autres nœuds. Formellement pour déterminer les nœuds centraux dans les réseaux, la mesure de centralité de proximité prend en compte les nœuds qui ont la plus petite longueur moyenne de chemin (séquence de liaisons) pour les nœuds qui sont liés à d'autres nœuds. La centralité de proximité est importante parce qu'il prend en compte non seulement les connexions immédiates d'un acteur, mais aussi les liens indirects de tous les autres nœuds de la chaîne de valeur de réseau. C'est une mesure de la portée, de la vitesse à laquelle l'information se propagera à tous les autres utilisateurs.

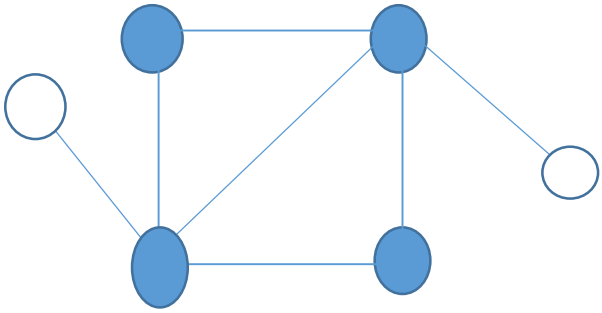


Figure 8. Exemple de centralité de proximité

5.4. Centralité de Katz

Un problème majeur avec la centralité des vecteurs propres se pose lorsqu'on considère les graphes acycliques. Dans un graphe acyclique orienté, la centralité devient zéro, même si le nœud peut avoir de nombreux arcs qui lui sont reliés. Dans ce cas, le problème peut être corrigé en ajoutant un terme de *bias* à la valeur de centralité. Le terme de *bias* est ajouté aux valeurs de centralité pour tous les nœuds, quelle que soit leur nature [2].

5.5. PageRank

Tout comme la centralité des vecteurs propres, la centralité de Katz est confrontée à certains défis. Un défi qui se produit dans les graphes orientés est qu'une fois qu'un nœud devient une autorité (haute centralité), elle passe toute sa centralité le long de tout de ses liens externes. Cela est moins souhaitable, parce que tout le monde n'est pas connu par la personne bien connue est bien connue. Pour atténuer ce problème, la valeur de la centralité est divisée par le nombre des arcs sortants de ce nœud de sorte que chaque voisin connecté obtienne une fraction de la centralité du nœud source [2].

6. Conclusion

Dans ce premier chapitre, nous avons introduits les définitions et les éléments de bases et les propriétés des réseaux sociaux en focalisant sur les concepts directement liés à la réalisation de notre projet. Dans le prochain chapitre, nous continuons l'introduction des concepts en discutant les approches similaires pour la détection des influenceurs dans les réseaux sociaux.

Chapitre II

Identification des influenceurs dans les réseaux sociaux

Chapitre II

Identification des influenceurs dans les réseaux sociaux

Dans ce chapitre, nous présentons et détaillons quelques travaux similaires réalisés dans le cadre d'identification des membres influenceurs dans les réseaux sociaux. L'identification des membres influenceurs est un sujet très actif dans cette dernière décennie. La détection des influenceurs nécessite à utiliser les techniques de l'intelligence artificielle, notamment les techniques de data mining et de la recherche d'informations afin de caractériser et déterminer quels sont les membres d'un réseau qui exercent le plus d'influence sur les autres, c.-à-d. les membres qui incitent les autres à avoir des comportements particuliers.

1. Qu'est-ce qu'un influenceur dans un réseau social ?

L'influence sur les réseaux sociaux est un phénomène très important, non seulement du point de vue de la circulation de l'information, mais aussi de point de vue du control d'envies des autres utilisateurs. Notamment dans le domaine de politique et marketing. Les influenceurs sont aussi appelés les leaders des opinions d'aujourd'hui.

Pour définir un influenceur, de nombreux auteurs ont leurs propres définitions. Watts & Dodds [14] définissent une personne influente ou un leader d'opinion comme une personne qui fait partie d'une minorité et qui exerce une influence sur un grand nombre de population.

La détection des influenceurs apporte une valeur importante dans la pratique. Dans l'exemple de la propagation d'une épidémie, si nous connaissons les membres influents, cela peut aider à prédire la propagation de la maladie et à la contrôler. Dans les réseaux criminels, l'identification des membres influenceurs permette la localisation rapide des chefs des gangs.

2. Les différentes approches et techniques existantes est leurs limites

Au cours des dernières années, de nombreuses approches ont été proposées pour l'identification des nœuds influents dans les réseaux sociaux. Chaque approche se repose sur une mesure particulière appliquée sur les différents membres d'un réseau. Notamment, WFCA [8], Leader Rank [21], K-Shell [18], centralité de proximité [20], centralité d'intimité [19], PageRank [17], indice H et HITS [16]. Dans cette section, nous choisissons de focaliser sur trois approches prometteuses dans la détection des influenceurs en Twitter.

2.1. La méthode WFCA

La méthode WFCA se repose sur un algorithme qui est proposé pour identifier les nœuds influents à l'aide d'une analyse formelle pondérée des concepts (WFCA). L'idée de base est de quantifier l'importance des nœuds via la technique WFCA [8]. Dans cette méthode, une analyse formelle pondérée du concept est appliquée. Cette méthode tient compte de l'information globale concernant un réseau donné ; ensuite, convertit les relations binaires entre les membres de réseau dans une hiérarchie. En fin, les nœuds sont agrégés en fonction de leurs attributs pour classer l'importance des nœuds.

2.2. Algorithme de l'influence-passivité (IP)

Romero et ses collaborateurs [15], sont parvenus à la conclusion que, si un utilisateur doit être considéré comme influent, alors il le fait non seulement qu'il doit être populaire et attirer l'attention de ses pairs, mais il doit aussi surmonter la passivité, un état dans lequel un utilisateur reçoit de l'information mais ne la propage pas sur le réseau.

L'utilisation de passivité dans l'algorithme proposé vient de la preuve que les utilisateurs de Twitter sont généralement passifs. Ainsi, lors de la détermination de l'influence d'un utilisateur, en tenant compte de la passivité de toutes les personnes qui

sont à l'origine de l'influence d'un utilisateur sont influencés par lui est également très important. Les hypothèses suivantes sont prises en compte par les auteurs :

1. Le score d'influence d'un utilisateur dépend du nombre de personnes qu'il influence, ainsi que de leur passivité.
2. Le score d'influence d'un utilisateur dépend du degré de dévouement des personnes qu'il influence. Le dévouement se mesure par le degré d'attention qu'un utilisateur accordé à un autre utilisateur, par rapport à d'autres utilisateurs.
3. Le score de passivité d'un utilisateur dépend de l'influence de ceux auxquels il est exposé, mais pas de ceux auxquels il est influencé par.
4. Le score de passivité d'un utilisateur dépend de la mesure dans laquelle il rejette l'influence d'un autre utilisateur par rapport à celle des autres.

Compte tenu de ces hypothèses, il convient de noter que le graphe du réseau pour cet algorithme est un graphe pondéré.

2.3. Algorithme de centralité efficace

L'algorithme de centralité efficace [13] est un algorithme qui propose une nouvelle mesure centrée sur l'influence de chaque nœud contribue à l'efficacité de l'ensemble du réseau. L'idée de base est de mesurer l'efficacité du réseau considérant le grand impact des nœuds les plus influents d'un réseau. L'efficacité est mesurée pour chaque nœud en le supprimant puis comparer le degré de changement de l'efficacité de réseau avant et après le retrait. La suppression de chaque nœud apporte en même temps l'enlèvement des arcs qui y sont reliés. Une fois que les nœuds clés sont supprimés et les arcs n'existent plus, la communication et le contact entre de nombreux nœuds de l'ensemble du réseau disparaîtront. En d'autres termes, l'élimination de ces nœuds pivots modifiera de façon importante le processus d'évaluation de l'efficacité de tout le réseau. Afin d'évaluer le rendement de ce programme, ils utilisent le modèle Sensible-Infecté-Recouvert (SIR) pour évaluer l'efficacité de l'algorithme et la

capacité d'étalement des nœuds avec différentes mesures de centralité. Basé sur quatre réseaux réels, les simulations montrent que la méthode proposée est plus performante en général, par rapport à l'utilisation de degré de centralité et la centralité de proximité.

2.4. Twitter Rank

Dans le contexte de Twitter. Les travaux de Weng [24] ont donné naissance à Twitter Rank, une extension de l'algorithme du PageRank qui prend en compte à la fois la similarité du sujet échangés entre utilisateurs et la structure des liens du réseau social. Cependant, l'influence d'un utilisateur peut varier selon les sujets, car un utilisateur de Twitter peut avoir des intérêts ou une expertise dans plusieurs domaines distincts. En [24], le nombre de suiveurs, c'est-à-dire, le nombre total de personnes qui suivent un utilisateur particulier a été interprété comme un bon indicateur d'influence.

Weng et ses collaborateurs [24] ont observé que 72,4 % des utilisateurs suivent plus de 80 % de leurs adeptes et que 80,5 % des utilisateurs ont 80 % de leurs amis (c.-à-d. les utilisateurs de Twitter dont les mises à jour sont suivies) qui les suivent. Le cadre général proposé pour Twitter Rank est décrit à la figure 10.

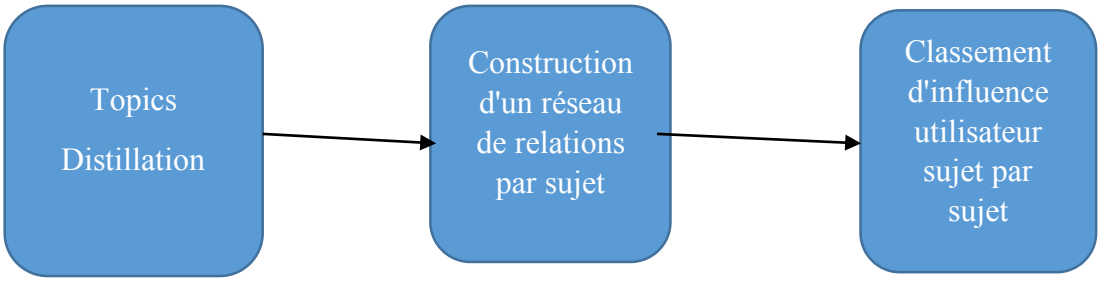


Figure 1. Le cadre général Twitter Rank (adapté de Weng et al. (2010))

Tout d'abord, dans la phase de d'extraction, les sujets qui intéressent les twitteurs sont extraits à partir de ce qu'ils tweetent. Ensuite, un réseau de relations spécifique au sujet est construit, basé sur les sujets précédemment rassemblés. Enfin, l'algorithme

TwitterRank est appliqué pour mesurer l'influence d'un twitter sensible au sujet, en tenant compte à la fois des sujets qui ont été extraits et de la structure du réseau de relations spécifique au sujet.

2.5. L'algorithme de recherche thématique induite par hyperliens (HITS)

L'algorithme HITS utilise une méthode de classement des pages Web mise au point par Kleinberg [16]. Cela est basé sur les principes d'autorités et de pôles. Les autorités, c'est-à-dire les pages qui ont un plus grand nombre de liens, ont une relation qui se renforce mutuellement avec les pages qui ont des liens vers de nombreux sites connexes.

3. Conclusion

Dans ce deuxième chapitre, nous avons introduits les différentes approches existantes dans la détection des membres influenceurs dans les réseaux sociaux. Chaque approche se repose sur une mesure d'influence appropriée. Ces mesures peuvent être modifiées d'un modèle d'un réseau à un autre. En plus, la majorité des techniques proposées sont basées sur l'aspect structurelle et statique des réseaux (la position des nœuds dans le réseau). Dans le dernier chapitre nous présentons notre contribution dans ce domaine ; en se basons sur l'aspect sémantique associé aux conversations. À fin, nous comparons les résultats obtenus avec les différentes techniques utilisées sur des données de Twitter.

Chapitre III
Détection des influenceurs dans les réseaux sociaux

Chapitre III

Détection des influenceurs dans les réseaux sociaux

Dans ce chapitre nous présentons et détaillons notre contribution à la détection des influenceurs dans les réseaux sociaux. Dans cette étude nous proposons et développons une technique pour l'identification des personnes influents sur Twitter. Pour l'élaboration de notre projet, deux expériences distinctes ont été menées, chacune avec un type différent. Dans la première expérience, nous avons utilisé des datasets de Twitter disponible sur le net comme celui de SNAP (un dépôt de données de l'Université de Stanford) [23]. Dans la deuxième expérience, nous avons recueilli les informations manquantes par le biais de Twitter API [22]. Un sous-ensemble de l'ensemble de données est utilisé pour construire des graphes orientés dont notre analyse a été basé.

1. Détection des influenceurs en Twitter

Pour la détection des influenceurs en Twitter, nous avons proposé une approche basée sur la sémantique des tweets diffusés sur le réseau. En Twitter, un utilisateur a le droit de poster des tweets à ses différents amis et suiveurs dans le réseau. Un suiveur ou un ami, dans le réseau, au doit de rediffuser les tweets reçus tel quels sont ou après des modifications ou ajouts des commentaires. Pour une bonne couverture des flots de diffusion des tweets, nous proposons de suivre la diffusion des tweets en comparant la liste des topics discutés dans les tweets consécutifs. Pour cela, notre technique consiste à :

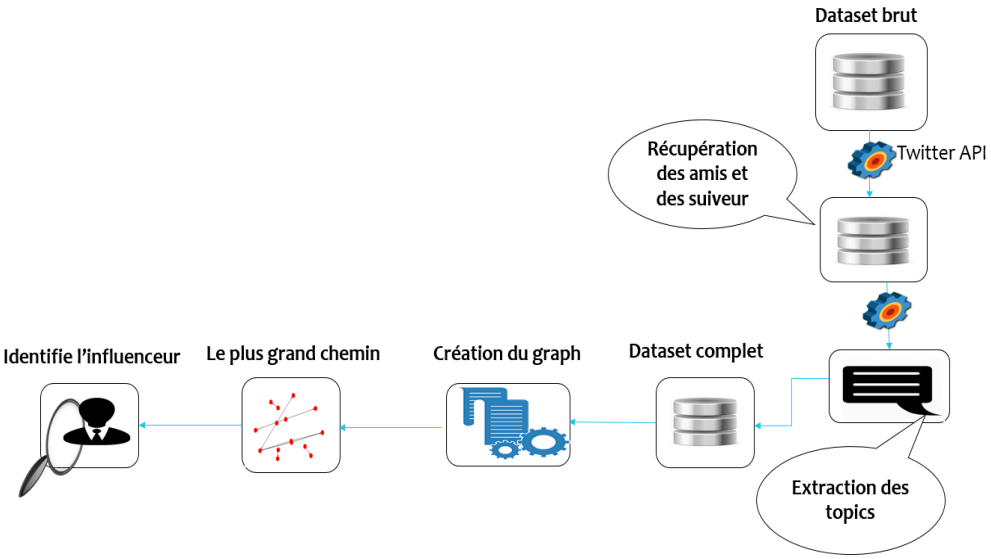


Figure 1. La représentation de notre algorithme

2. La collection des datatsets

Collecter des datatsets de Twitter disponible sur le net est une tâche lourde est difficile. Notamment dans le choix des bons datasets qui inclusent tous les attributs qu'on a besoin dans notre projet. En plus, les fichiers des datasets doivent êtres sous une format standard Json. Cela facilite considérablement la manipulation des données.

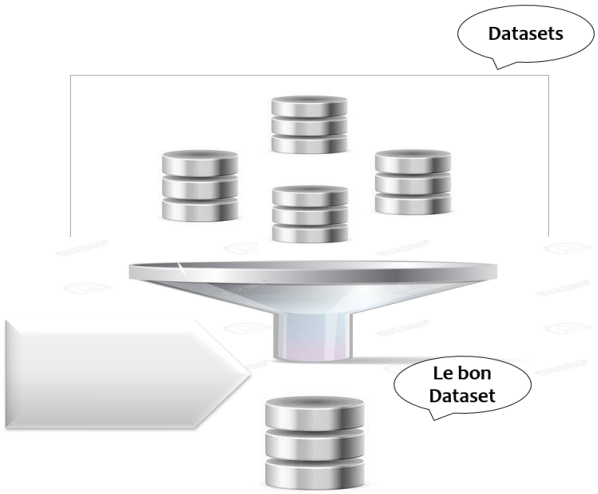


Figure 2. Trouver le bon dataset

3. Tri des dataset par temps

Dans cette phase nous cherchons à trier les tweets des datatsets par temps de création si la liste des tweets ne sont pas triés. Cette étape est plus importante afin de trouver le premier utilisateur qui a partagé le tweet propagé.

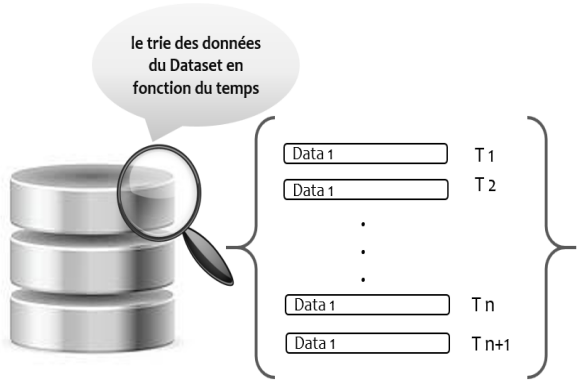


Figure 3. La 2eme étape d'algorithme

4. Récupération des amis et des suiveurs

Les datatsets trouvé sur le net inclus des informations générales sur le tweet mais peu d'informations sur les utilisateurs. Notamment, la liste des amis et suiveurs de chaque utilisateur. Ces dernies informations sont pertinentes pour la réalisation de notre projet. Pour cela, nous avons utilisé le Twitter API pour récupérer la liste des amis et suiveurs de chaque utilisateur ayant un au moins un tweet dans le dataset choisi. Ensuite, la liste amis de chaque membre doit être filtrée pour ne garder que des utilisateurs qui sont déjà dans le dataset. Notons qu'il y a aussi des utilisateurs privés qui ne peuvent pas être récupéré par Twitter API.

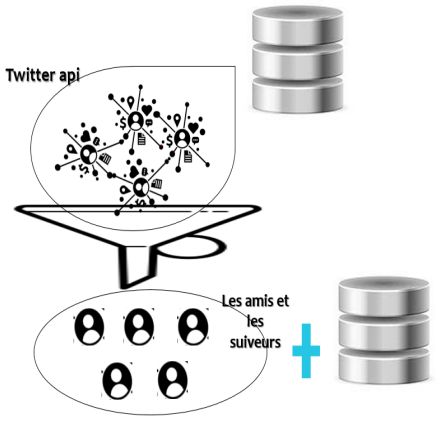


Figure 4. Récupération des amis et des suiveurs

5. Extraction des topics

Faire extraire la liste des sujets discutés dans chaque tweet en dataset est une étape pertinente dans ce projet. Un sujet est un terme construit par un ou plusieurs mots appelé aussi n-grammes. Ces termes sont soit des adjectifs soit des noms. Les hashtags sont transformés en sujet par la suppression du symbole (#) et la séparation des mots en chaque lettre majuscule.

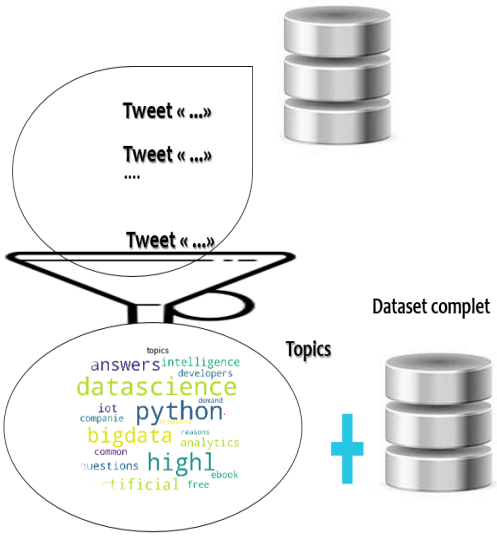


Figure 5. Extraction des topics

6. Recherche des topics propagés

Dans cette phase, nous construisons un graphe de propagation des sujets entre les différents utilisateurs dans le dataset choisi. Pour cela, nous choisissons d'abord un utilisateur et nous parcourons la liste de ses amis et suiveurs partageant au moins un des topics discutés dans son tweet. Si c'est le cas, un arc entre l'utilisateur et son ami est ajouté à un graphe indiquant que le tweet a été propagé par son ami.

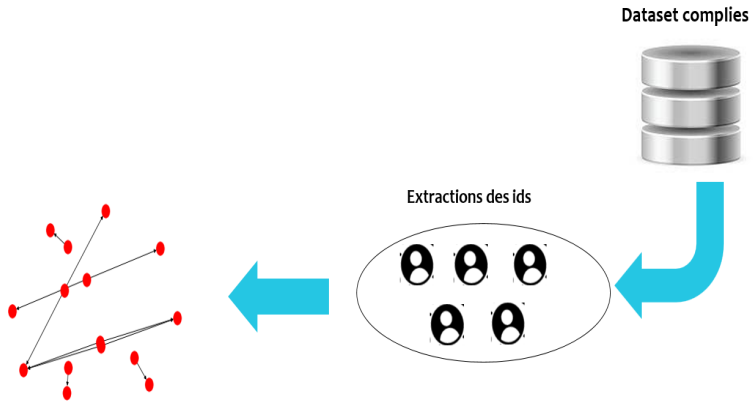


Figure 6. Recherche des topics propagés

7. Répétition de la recherche

La recherche des sujets propagés est répétée pour tous les utilisateurs dans le datatset. Cela nous permet de construire un graphe orienté indiquant le flux de propagation de chaque tweet dans le réseau.

8. Extraction du plus long chemin

Intuitivement, l'influenceurs est celui qui dans son sujet de tweet a été propagé le maximum dans le réseau. Pour cela, il suffit de parcourir le graphe des sujets propagés et de trouver le plus long chemin c.à.d la plus longue séquences listes des membres partageant un sujet. On peut trouver plusieurs longs chemins, cela veut dire que

plusieurs sujets ont été propagés dans le réseaux avec le même degré. En raison de simplicité nous choisissons le premier.

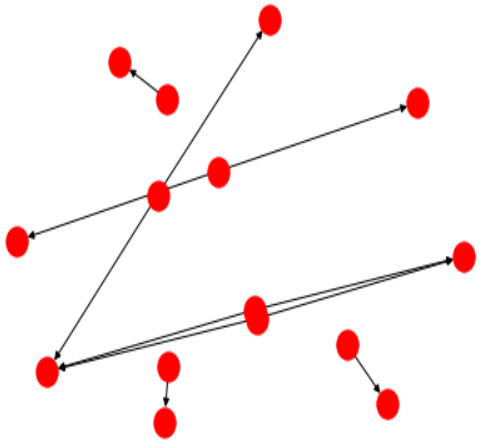


Figure 7. Extraction du long chemin

9. Identification d'influenceur

Intuitivement, l'influenceur est celui qui a déclenché le débat sur le sujet propagé. Cela indique que l'influenceur est représenté par de premier nœud dans le plus long chemin trouvé dans l'étape précédente.

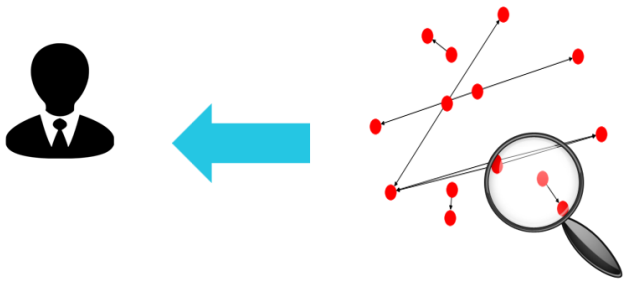


Figure 8. Identification d'influenceur

10. Conclusion

Dans ce chapitre, nous avons présenté les différentes phases de notre algorithme proposé pour la détection des influenceurs en Twitter. Par notre algorithme nous déterminons l'influenceur par l'étude de propagation des sujets discutés dans les tweets pour une période de temps donné.

Chapitre IV
Implémentation de notre algorithme

Chapitre IV

Implémentation de notre algorithme

Dans ce chapitre nous présentons les différentes phases d'implémentation de notre algorithme, les résultats obtenus ainsi que la liste des outils utilisés dans l'implémentation.

1. Outils de réalisation du projet

Dans notre étude, nous présentons la liste des outils utilisés dans notre technique pour la détection des influenceurs en Twitter.

1.1. Liste des datasets expérimentées

Une datatset est un ensemble de données collectée et fournis par des fournisseurs diverses (personnes ou organisations) pour être utilisées pour des fins académiques ou commerciales. Dans notre projet, nous avons utilisé des datasets Twitter ces datasets regroupent un certain nombre de tweets collecté de Twitter de différentes façons. Tableau 2. Indique l'ensemble des datasets utilisées dans notre projet.

Dataset	**Nbre d'utilisateurs**	**Source**
Trump2008-2018	264616	http://www.trumptwitterarchive.com/data/realdonaldtrump
SNAP	81306	https://snap.stanford.edu/data/ego-Twitter.html
Twitter_API-Billux	317	https://drive.google.com/drive/u/0/my-drive
Twitter samples	20 000	

Tableau 1. Liste des datasets expérimentées

Chaque tweet en data set est un objet complexe. Le tableau suivant fournit une liste de tous les attributs d'un tweet avec une brève description de leurs significations.

Nom d'attribut	Description
created_at	Indique l'heure et la date de création du compte utilisateur.
contributors_enabled	Indique si le contributeur est en mode active ou pas
default_profile	Indique si l'utilisateur n'a pas modifié son profil.
Description	Un chaîne de caractères qui décrire le profil de l'utilisateur
default_profile_image	Indique si l'utilisateur n'a pas d'image de profil personnalisée
entities	La liste des entités de l'URL ou de la description
followers_count	Indique le nombre de suiveurs.
follow_request_sent	Indique si une demande de suivi a été envoyée.
favourites_count	Le nombre de tweets favoris par l'utilisateur
following	Indiquer si l'utilisateur authentifié suit ou non
friends_count	Indique le nombre d'amis de l'utilisateur
geo_enabled	Indique si la géolocalisation est activée
Id	Identificateur unique de l'utilisateur
id_str	Identificateur de l'utilisateur sous la forme d'une chaîne de caractères
is_translator	Indique si l'utilisateur fait partie de l'équipe de Twitter.
lang	La langue préférée par l'utilisateur
listed_count	Le nombre de listes publiques dont l'utilisateur est membre
Location	L'emplacement déclaré par l'utilisateur sous forme de chaîne de caractères
Name	Nom de l'utilisateur
profile_*	La quantité d'informations relatives au profil

Protected	Indique si l'utilisateur protège ses tweets.
Status	Il s'agit d'un objet incorporé avec le dernier tweet
screen_name	Pseudo nom Twitter de l'utilisateur
statuses_count	Le nombre de tweets
time_zone	Le fuseau horaire déclaré par l'utilisateur
utc_offset	Le décalage horaire par rapport à GMT/UTC en secondes
url	L'URL fournie par l'utilisateur associé au profil
Verified	Indique si l'utilisateur est vérifié.

Tableau 2. La structure d'un tweet

Comme vous voyez dans le tableau précédent ; un tweet comporte des informations générales sur son utilisateur et la date de sa création. Malheureusement, ces informations ne comportent pas la liste des amis et suiveurs du tweet. Quand cette information est pertinente pour l'élaboration de notre projet, une technique de récupération de ces informations a été mise en place.

Sachant que chaque tweet posté par un utilisateur est visible à la liste des amis et suiveurs de l'utilisateur, il suffit de récupérer la liste des identificateurs des amis et suiveurs de chaque utilisateur en data set. Pour cela nous avons utilisé Twitter API comme moyen de récupération automatique des informations manquantes.

1.2. Le langage de programmation Python

Python est un langage de programmation général de haut niveau créé par Guido van Rossum et publié en 1991. Il est important de noter qu'il s'agit d'un langage interprété Python a une philosophie de conception qui met l'accent sur la lisibilité du code. Python utilise un système de type dynamique de type Data Science Technologie Stack et de gestion automatique de la mémoire et prend en charge de multiples paradigmes

de programmation (orienté objet, impératif, programmation fonctionnelle et procédurale). Grâce à son succès mondial, il dispose d'une vaste bibliothèque standard complète. Le Python Package Index (PyPI) [25] fournit des milliers de modules tiers prêts à l'emploi pour vos projets informatiques.

Nous vous suggérons d'installer aussi Anaconda. C'est une distribution open source de Python qui simplifie la gestion des paquets et le déploiement des fonctionnalités [26]. La version utilisée dans notre projet est celui de 3.7.

1.3. Qu'est-ce que Anaconda Navigator ?

Anaconda Navigator est une interface utilisateur graphique de bureau (GUI) incluse dans la distribution Anaconda qui vous permet de lancer des applications et de gérer facilement les paquets, environnements et canaux conda sans utiliser de commandes en ligne de commande. Navigator peut rechercher des paquets sur Anaconda Cloud ou dans un dépôt Anaconda local. Il est disponible pour Windows, macOS et Linux.

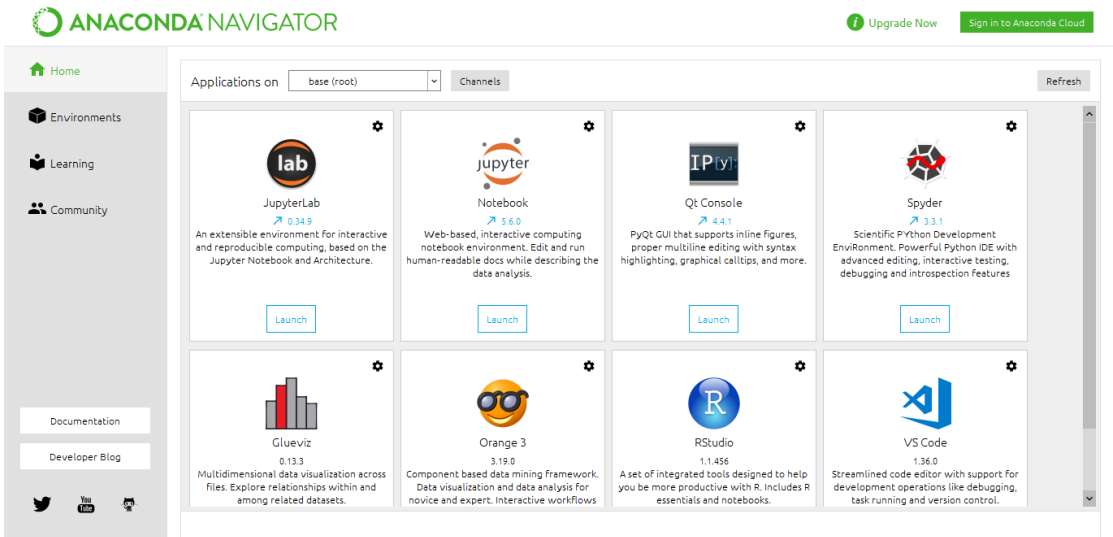


Figure 1. Représentation de l'interface d'anaconda

1.4. Spyder

Spyder est un environnement scientifique puissant écrit en Python, pour Python, et conçu par et pour des scientifiques, ingénieurs et analystes de données. Il offre une combinaison unique des fonctionnalités avancées d'édition, d'analyse, de débogage et de profilage d'un outil de développement complet avec l'exploration de données, l'exécution interactive, l'inspection approfondie et les belles capacités de visualisation d'un logiciel scientifique.

Au-delà de ses nombreuses fonctionnalités intégrées, ses capacités peuvent être encore étendues grâce à son système de plugins et son API. De plus, Spyder peut également être utilisé comme bibliothèque d'extension PyQt, ce qui permet aux développeurs de s'appuyer sur ses fonctionnalités et d'intégrer ses composants, tels que la console interactive, dans leur propre logiciel PyQt [29].

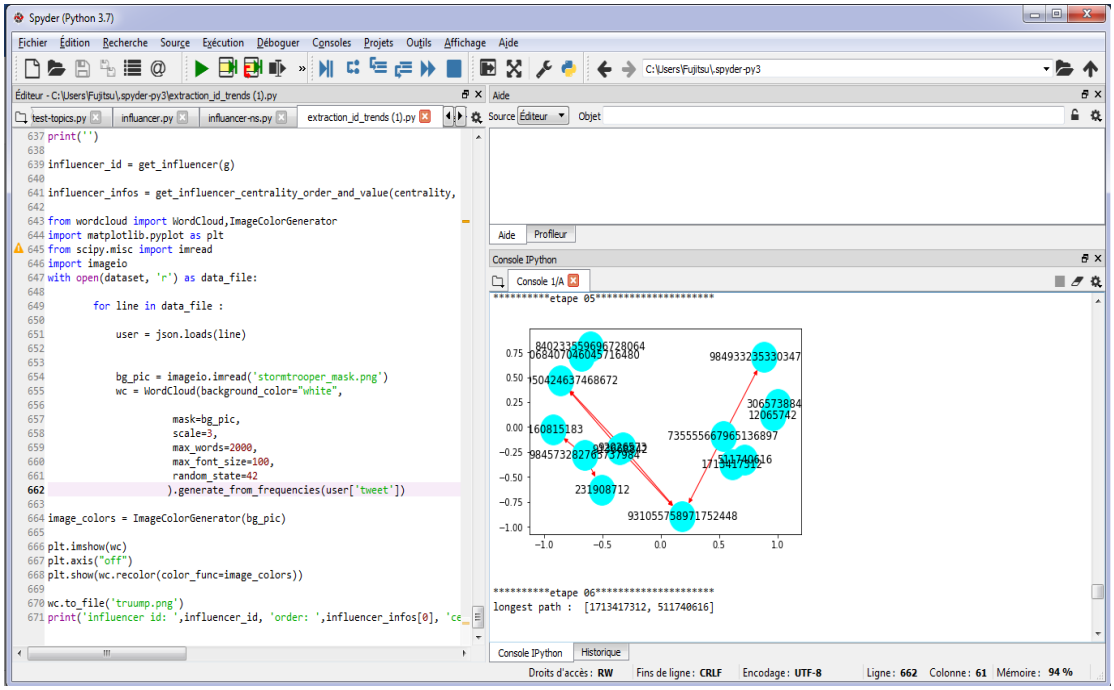


Figure 2. Représentation de spyder

1.5. Twitter API

Twitter offre une série d'API pour fournir un accès programmatique aux données de Twitter, y compris lire des tweets, accéder à des profils d'utilisateurs et publier du contenu au nom d'un utilisateur authentifié. Dans notre projet, l'utilisation de Twitter API est indispensable pour la récupération des listes des amis et suiveurs des utilisateurs dans les datasets utilisés dans nos expérimentations. Afin de permettre à accéder aux données de Twitter, il faut d'abord enregistrer une demande d'accès.

L'étape de l'enregistrement prendra quelques minutes. Après la connexion au compte Twitter, il faut pointer sur le manageur d'applications et créer une nouvelle application. Une fois l'application enregistrée, sous l'onglet *Clés et jetons d'accès*, on peut trouver les informations nécessaires pour authentifier la demande : la *Clé du Consommateur* et le *Secret Consommateur* (aussi appelé clé API et secret API, respectivement) est un paramètre de l'application. Le *jeton d'accès* et le *Secret du jeton d'accès* sont à la place des paramètres du compte utilisateur.

Consumer key	xvz1evFS4wEEPTGEFPHBog
Consumer secret	L8qq9PZyRg6ieKGEKhZolGC0vJWLw8iEJ88DRdyOg
RFC 1738 encoded consumer key (does not change)	xvz1evFS4wEEPTGEFPHBog
RFC 1738 encoded consumer secret (does not change)	L8qq9PZyRg6ieKGEKhZolGC0vJWLw8iEJ88DRdyOg
Bearer token credentials	xvz1evFS4wEEPTGEFPHBog:L8qq9PZyRg6ieKGEKhZolGC0vJWLw8iEJ88DRdyOg

Figure 3. Exemple de clés de twitter-API

Le niveau d'accès à ces paramètres définit ce que l'application peut faire pendant l'interaction avec Twitter au nom d'un utilisateur. L'option en lecture seule est l'option la plus conservatrice, car l'option ne sera pas autorisée à publier quoi que ce soit ou à interagir avec d'autres utilisateurs par l'intermédiaire d'une application directe.

1.6. PKE : Module d'extraction des phrases-clés en Python

PKE est une boîte à outils d'extraction de phrases-clés basée sur python et open source. Il fournit un pipeline d'extraction de phrases-clés de bout en bout dans lequel chaque composant peut être facilement modifié ou étendu pour développer de nouveaux modèles. PKE permet également de comparer facilement les modèles d'extraction de phrases-clés les plus récents et est livré avec des modèles supervisés formés sur le jeu de données SemEval-2010. La version utilisée dans ce projet est celui de 1.8.

1.7. NetworkX

Est une bibliothèque Python qui vous permet de créer, manipuler et étudier la structure, la dynamique et les fonctions des réseaux complexes et d'instancier des graphiques composés de nœuds et de ponts, ou de liens. Grâce à cette bibliothèque, la manipulation de ces graphiques est simplifiée.

Avec NetworkX, vous pouvez charger et stocker des réseaux dans des formats de données standard et non standard, générer de nombreux types de réseaux aléatoires et traditionnels, analyser la structure du réseau, construire des modèles de réseau, concevoir de nouveaux algorithmes de réseau, concevoir des réseaux, et bien plus [36]. La version utilisée dans ce projet est celui de 2.1.

1.8. Tweepy

Tweepy est un paquet logiciel open-source, hébergé sur GitHub et permet à Python de communiquer avec la plate-forme Twitter et d'utiliser son API. La version utilisée dans ce projet est celui de 3.7.

1.9. Le format Json

JSON est un format de fichier pratique pour stocker des données structurées qui peuvent être traitées un enregistrement à la fois. Il fonctionne bien avec les outils de

traitement de texte. C'est un format idéal pour les fichiers journaux. C'est aussi un format flexible pour transmettre des messages entre les processus de coopération.

2. Résultats préliminaires obtenus

Dans cette section nous présentons et discutons les résultats obtenus dans notre étude. Tableau 3 décrit l'ensemble des données utilisées après l'élimination des informations inutiles, comme les tweets des utilisateurs isolés (sans amis et suiveurs en dataset), liste des topics les plus fréquentés.

Dataset	**Nbre. Tweets**	**Nbre. Utilisateurs**	**Nbre. Utilisateurs après prétraitement**	**Topics**	**Plus long chemin**	**Influenceur**
Twitter_API-Billux	317	297	101	{'python', 'reasons', 'big data', 'common questions', 'data science', 'highl'}	[1713417312, 511740616]	1713417312
Twitter samples	20 000	4711	705	{'David Cameron', 'miliband', 'sturgeon', 'clegg', 'farage, tory', 'tories', 'ukip',		

				‘snp’, ‘libdem’}		

Tableau 3. Les résultats obtenus

3. Test d’ algorithme sur le Dataset « Twitter_API-Billux » :

Etape 1. Collecter des datatsets de Tweeter disponible sur le net

- On a téléchargé le dataset avec le twitter API en format json pour manipuler et extraire les données nécessaires

```
# Variables that contains the user credentials to access Twitter API

OAUTH_TOKEN = '959154344569589760-HHLP86YEfIT3UpSkAQ23k6kPTVaqmk8'

OAUTH_TOKEN_SECRET = '81Sz3x4ECgzqVEdL6QRqP3wrUQek6ZyMNNsMVEFhdC7Ob'

CONSUMER_KEY = 'rj19IFUzugE9UzfzDb9c8ZTi0'

CONSUMER_SECRET = 'Vc9TgRa6G31NVRwABhXewuhnQgm9tGTKCtJ4JueFGDzurMio0X'


# Setup tweepy to authenticate with Twitter credentials:

#

auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)

auth.set_access_token(OAUTH_TOKEN, OAUTH_TOKEN_SECRET)

#

# Create the api to connect to twitter with your creadentials

api = tweepy.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=True, compression=

auth = twitter.oauth.OAuth(OAUTH_TOKEN, OAUTH_TOKEN_SECRET,

                           CONSUMER_KEY, CONSUMER_SECRET)


twitter_api = twitter.Twitter(auth=auth)
```

Figure 4. Connection à twitter API pour téléchargement du dataset

Etape 2. Trier les tweets par temps de de création si la liste des tweets en dataset n’est pas triée

- On a trouvé les tweets déjà triés dans cette dataset.

Etape 3. Utiliser Twitter API pour récupérer la liste des amis et suiveurs de chaque utilisateur ayant un Tweet dans le data set. La liste doit ensuite être filtrée pour ne garder que des utilisateurs qui sont déjà dans le dataset.

- Nous avons trouvé ce résultat qu'est présenté dans cette figure ci-dessous :

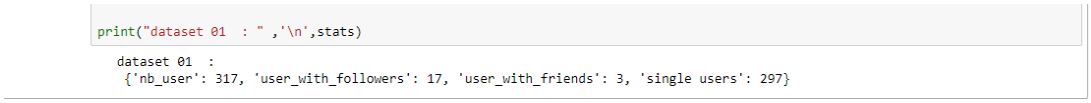


Figure 5. Les détails de dataset **Twitter_API-Billux**

Etape 4. Faire extraire de la liste des topics discutés dans chaque tweet en data set

Voilà la fonction qu'on a proposé pour l'extraction du topic

```
pos = {'NOUN', 'PROPN', 'ADJ'}

with open('u'+dataset, 'w') as updated_file:

    with open(dataset, 'r') as data_file:

        nb = 0

        for line in data_file :

            nb+=1

            user = json.loads(line)

            if user['id'] in users:

                extractor = pke.unsupervised.TopicRank()

                extractor.load_document(input=user['tweet'].lower(), language='en', norm

                extractor.candidate_selection(pos)

                extractor.candidate_weighting()

                user['topics'] = extract_keys(extractor.topics)

                if nb > 1:

                    updated_file.write('\n')

                json.dump(user, updated_file)

        data_file.close()

updated_file.close()
```

Figure 6. La fonction proposée pour l'extraction des topics

Après l'extraction des topics en regardons que la liste des topics et ajouter dans notre dataset.

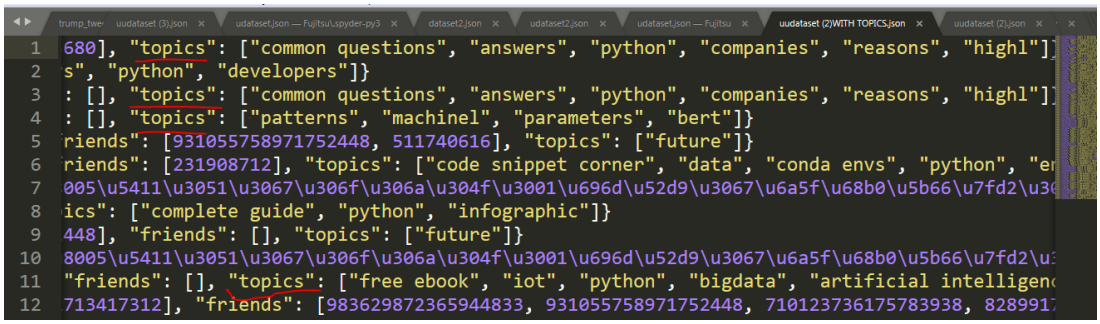


Figure 7. L'extraction des topics

Etape 5 et 6. Pour chaque utilisateur, nous parcourons la liste de ses amis et suiveurs partageant au moins un des topics discutés dans son tweet. Si c'est le cas, un arc entre l'utilisateur et son ami partageant des topics de leurs tweets est ajouté à un graphe indiquant que le tweet a été propagé par son ami. Cette étape est répétée pour tous les utilisateurs dans le datatset. Cela nous permet de construire un graphe orienté indiquant le flux de propagation de chaque tweet dans le réseau.

La fonction qu'on a utilisé pour la création du graph est la suivante :

```
470 def create_graph_from_data(dataset) :
471
472     g = nx.DiGraph()
473
474     ids = get_ids_from_final_dataset(dataset)
475
476 #    print(ids)
477
478     for id in ids :
479
480         user = get_user(dataset,id)
481
482         topics =  user['topics']
483
484         followers = list(set(user['followers'] + user['friends']))
485
486         for follower in followers :
487
488             ufollower = get_user(dataset,follower)
489
490 #            print(follower, ufollower)
491
492             ftopics = ufollower['topics']
493
494
495             if any(t in ftopics for t in topics) :
496
497                 if not g.has_edge(id, follower) and not g.has_edge(follower, id) :
498
499                     g.add_edge(id, follower)
500
```

Figure 8. La fonction de la création de graphe

Les résultats obtenus après cette fonction on a un graphe qui représente la propagation des topics

```
**********etape 05**********
```

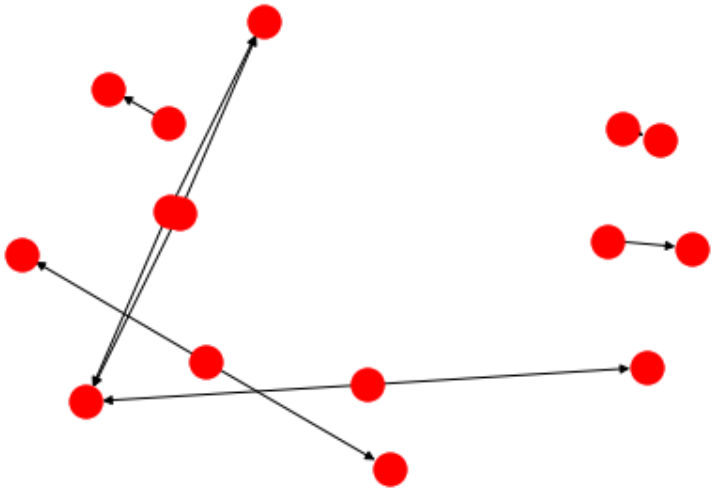


Figure 9. La construction du graph

Etape 7. Trouver la ou les plus longs chemins dans le graphe. La figure 17 montre le plus grand chemin dans le graph obtenu par l'étape 5 et 6. Le plus long chemin est présenté sous forme d'une liste montrons la séquence des utilisateurs dans le chemin. Notons que si plusieurs chemins de mêmes longueurs sont présents, l'algorithme choisi le premier dans le graph.

```
**********etape 06**********
longest path :  [1713417312, 511740616]
```

Figure 10. Le plus long chemin

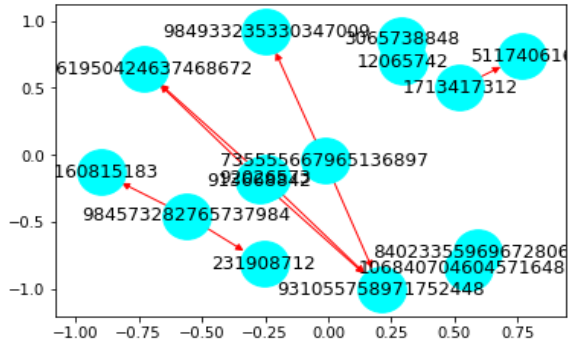
Etape 8. Les influenceuses sont les identificateurs des nœuds en tête de chaque. L'utilisateur influenceurs est celui en premier position dans le plus long chemin obtenu par l'étape 7. Le calcul de degré de centralité de cet utilisateur montre que le degré de centralité n'est pas un facteur décisif pour la détection des influenceurs.

```
**********etape 07**********

influencer id:  1713417312 order:  11 centrality value:  0.23980056885558415
```

Figure 11. Trouver l'influenceur

Figure 15 montre le graph résultant de l'application de notre algorithme sure la première datatset.




```
longest path :  [1713417312, 511740616]
centrality :  [(0.3117407395122594, 931055758971752448), (0.28776068262670096, 1061950424637468672),
(0.2637806257411426, 1068407046045716480), (0.2637806257411426, 984933235330347009), (0.2637806257411426,
3065738848), (0.2637806257411426, 1160815183), (0.2637806257411426, 511740616), (0.2637806257411426,
231908712), (0.23980056885558415, 984573282765737984), (0.23980056885558415, 840233559696728064),
(0.23980056885558415, 735555667965136897), (0.23980056885558415, 1713417312), (0.23980056885558415,
913668842), (0.23980056885558415, 92026573), (0.23980056885558415, 12065742)]
influencer id:  1713417312 order:  11 centrality value:  0.23980056885558415
```

Figure 12. Résultat final de notre algorithme

4. Conclusion

Dans ce chapitre, nous avons présenté notre contribution à la détection des influenceurs en Twitter. Notre proposition consiste à déterminer l'influenceur par l'étude de propagation des sujets discutés dans les tweets pour une période de temps donné. Pour cela, une extraction des sujets discutés dans les tweets a été élaborée, ensuite un graphe est construit qui détermine la propagation des sujets entre les amis et suiveurs dans un sous réseaux de Twitter. L'influenceur est celui en tête de plus long chemin du graph obtenu.

Conclusion Générale

Par ce projet, nous avons contribué à l'identification des influenceurs en Twitter. Notre proposition prend en considération l'aspect sémantique (propagation des sujets en discussions) des réseaux. Pour cet objectif, nous avons proposé un algorithme basé sur le taux de rediffusion des tweets (retweets) capturant le flux d'informations dans le réseau et non seulement l'aspect structurelle des réseaux sociaux. Les résultats sont préliminaires et sont étroitement liés aux facteurs suivants :

1. La disponibilité des datasets des tweets pour une période spécifique.
2. Les limites de Twitter API pour la récupération des listes des amis et suiveurs de chaque utilisateur.
3. L'algorithme utilisé pour l'extraction des sujets discutés en tweets.

Ces trois éléments ont une influence directe sur les résultats obtenus par notre étude. Profitant du fait que ce domaine de recherche n'en est qu'à ses débuts, nous planifions d'étendre cette étude par l'utilisation d'autres algorithmes et techniques efficaces d'extraction des topics depuis l'association des sujets des tweets aux sujet de DBPedia.

Bibliographies

[1] Wasserman S, Faust K (1994), *Social network analysis: methods and applications*. Cambridge University Press, Cambridge, UK.

[2] Mohammed Zuhair Al-Taie, Seifedine Kadry (2017), *Python for Graph and Network Analysis*, Springer Publishing Company.

[3] E. Bakshy, J.M. Hofman, W.A. Mason, and D.J. Watts (2011), *Everyone's an influencer: quantifying influence on Twitter*, Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM'11), Hong Kong, China, pp. 65–74, ACM.

[4] L.A. Adamic et E. Adar (2003), *Friends and neighbors on the web*., Social Networks 25 (3), pp. 211–230.

[5] N. Agarwal, H. Liu, L. Tang, et P.S. Yu (2008), *Identifying the influential bloggers in a community*, Proceedings of the International Conference on Web Search and Web Data Mining, pp. 207–218, ACM.

[6] Reza Zafarani, Mohammed Ali Abbasi, Huan Liu (2014), *Social media mining: and Introduction*, Cambridge University Press, Cambridge, UK.

[7] Marco Bonzanini (2016), *Mastering Social Media Mining with Python*, Packt Publishing, Birmingham, UK.

[8] Zejun Sun, Bin Wang, Jinfang Sheng, Yixiang Hu, Yihan Wang, et Junming Shao (2017), *Identifying Influential Nodes in Complex Networks Based on WFCA,* IEEE Access, vol. 5, pp. 3777-3789.

[9] D. G. Kourie, S. Obiedkov, B. W. Watson, and D. van der Merwe (2009), *An incremental algorithm to construct a lattice of set intersections*, Science of Computer Programming 74 (3), pp. 128–142.

[10] L. Zou, Z. Zhang, and J. Long (2015), A fast incremental algorithm for constructing concept lattices, Expert Syst. Appl. 42 (9), pp. 4474–4481.

[11] E. Bakshy, J.M. Hofman, W.A. Mason, and D.J. Watts (2011), Everyone's an influencer: quantifying influence on Twitter, Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM'11), Hong Kong, China, pp. 65–74, ACM.

[12] J. Outrata and V. Vychodil (2012), Fast algorithm for computing fixpoints of Galois connections induced by object-attribute relational data, Information Sciences 185 (1), pp. 114–127.

[13] Shasha Wang, Yuxian Du, Yong Deng (2016), A new measure of identifying influential nodes: Efficiency centrality, Communications in Nonlinear Science and Numerical Simulation, vol. 4, doi: 10.1016/j.cnsns.2016.11.008

[14] D. J., Watts, D. J. and P.S. Dodds (2007), Influentials, Networks, and Public Opinion Formation. Journal of Consumer Research, vol. 34, pp. 441-458.

[15] ROMERO, D.M., GALUBA, W., ASUR, S. & HUBERMAN, B.A. (2011). *Influence and passivity in social media. In Proceedings of the 20th International Conference Companion on World Wide Web*.

[16] J. M., Kleinberg *(1998), PageRank as a function of the damping factor.* In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms.

[17] P. Boldi, M. Santini, S. Vigna (2005). *PageRank as a function of the damping factor.* In Proceedings of the 14th International Conference on World Wide Web. pp. 557-566

[18] P. E. Brown, J. Feng (2011). Measuring User Influence on Twitter Using Modified K-Shell Decomposition, Fifth International AAAI Conference on Weblogs and Social Media, pp. 18-23.

[19] M. Barthélemy (2004). *Betweenness centrality in large complex networks, vol. 38, Issue 2, pp 163–168*

[20] Kazuya Okamoto, Wei Chen, Xiang-Yang Li (2008). *Ranking of Closeness Centrality for Large-Scale Social Networks. FAW 2008: Frontiers in Algorithmic pp 186-195.*

[21] L. L¨u, Y.-C. Zhang, C. H. Yeung, T. Zhou (2011), *Leaders in social networks, the delicious case, PloS One 6, e*21202.

[22] J. Weng, E. P. Lim, J. Jiang, Q. He (2010). *Finding topic-sensitive influential twitters. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*

[23] B. Florian, pke: an open source python-based keyphrase extraction toolkit, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pp. 69-73, 2016.

Webgraphie

[24] classemate https://www.classmates.com

[25] facebook https://www.facebook.com

[26] youtube https://www.youtube.com

[27] twitter https://twitter.com

[28] facebook https://www.journaldunet.com/ebusiness/le-net/1125265-nombre-d-utilisateurs-de-facebook-dans-le-monde/

[29] twitter https://www.journaldunet.com/ebusiness/le-net/1159246-nombre-d-utilisateurs-de-twitter-dans-le-monde/

[30] flickr https://www.flickr.com

[31] https://fr.linkedin.com

[32] Twitter website, https://developer.twitter.com. 23 JAN 2019.

[33] Snap website, https://snap.stanford.edu/data/#twitter. 28 DEC 2018.

[34] Python website: https://pypi.python.org/pypi /

[35] Anaconda website: https://www.anaconda.com/distribution/

[36] Networkx website: http://networkx.github.io

[37] resaux social site web http://socialonline.over-blog.com/2016/01/les-reseaux-sociaux-et-son-histoire.html