Recherche et extraction d'information sur le réseau social Twitter

1st Ben ammar Iliass
UFR STN
Université Paris 8
Vincennes - Saint-Denis
iliass.ben-ammar@etud.univ-paris8.fr

2nd Ramakichenin Nalan
UFR STN
Université Paris 8
Vincennes - Saint-Denis
nalan.ramakichenin@gmail.com

***Abstract*—Ces derniers temps, Twitter est devenu l'un des moyens les plus populaires les plus appréciés pour communiquer des informations aux personnes du monde entier. Il existe plusieurs types d'informations, informations médiatiques, informations à usage commercial, informations du quotidien... Dans cet article nous allons voir les différents types d'informations, quels sont les données que l'on peut extraire du réseau social Twitter.**

***Index Terms*—Extraction, informations, Twitter, réseaux sociaux, Recherche, Media, API**

I. Introduction

De nos jours, des informations utiles sont intégré un peu partout sur internet comme dans une page html ou les réseaux sociaux par exemple.

Et si l'on pouvait récupérer et utilisé ces informations à diverses fins ? La croissance d'Internet nous a permis de constater que l'obtention d'informations ainsi que la diffusion devient de plus en plus accessible, ce qui signifie que les utilisateurs ont accès aux données depuis n'importe ou n'importe quand.

Les réseaux sociaux sont de nos jours présent et utilisé très fréquemment. Pour de la prise d'information ou pour relayer de l'actualité provenant des médias, certains réseaux sociaux sont même plus rapides que les médias, comme Twitter par exemple.

Cependant, ces informations peuvent tout aussi bien être des fake news, visant à diffamer des entreprises, des marques, des personnalités connue, des communautés... C'est pourquoi les réseaux sociaux ont un certain rôle a joué, d'une part, ils se doivent de laissé la liberté d'expressions, d'autre part, ils se doivent de supprimer les cas trop graves de fake news.

Aujourd'hui, près de 80% des marques sont présentes sur les réseaux sociaux. Elles utilisent ces moyens de communication pour gérer la relation et la communication avec leurs cibles. Leur contenu donc peuvent être analysés qui peut être ensuite utilisé dans un domaine d'activité comme la commercialisation.

Ce document est organisé comme suit : La section II présente l'énoncé du problème, la section III décrit les travaux connexes. La section IV décrit les différentes méthodes d'extraction d'informations possible, la section V d'écrit la méthode et la stratégie adopté, la section VI décrit les résultats de notre programme, enfin dans la section VII nous concluons l'article.

II. Problématique

L'extraction de données du Web est un problème important qui a été largement étudié à l'aide de différents outils et applications scientifiques.

De nombreuses approches ont été proposées et conçues pour résoudre des problèmes spécifiques et fonctionner dans divers domaines. Aujourd'hui, les entreprises, les marques, les influenceurs, médias fournissent des informations utiles sur les réseaux sociaux comme Facebook, Instagram, Twitter ou YouTube. Les petites organisations qui n'ont pas de site web personnel s'appuient sur les réseaux sociaux pour offrir leurs services. Les petites entreprises qui n'ont pas de sites Web personnels s'appuient sur les réseaux sociaux pour fournir des informations sur leurs services aux clients. Il y a également certains médias qui ont connu le jour grâce aux réseau sociaux.

C'est pourquoi nous visons à extraire des données provenant des petits média comme "mediavenir" "cerfia" qui sont des médias provenant de twitter suivis par des millions d'utilisateurs. Ces médias sont toujours à la pointe de l'actualité , généralement plus rapide et pertinent que certains gros média de la télévision ou autres
Une fois les données extraites nous voulons les incorporer de manière à les présenter sur un site intuitif et facile à interpréter et ainsi stocker toutes les informations dans la base de données.

Il se peut que les informations mises à disposition ne proviennent pas forcément de sources fiables. La véracité de celle-ci peuvent varier car il n'y pas toujours de vérification effectuée derrière.

III. Travaux connexe

A. Business Data Extraction from Social Networking

L'extraction de données peut s'avérer très utile pour l'extraction d'informations pertinente et concrète ces informations seront alors transformées et peuvent être utilisées à des fin diverses comme. Les informations peuvent être extraites de manière différente, chaque méthode sont des méthode d'extraction unique. Par exemple un système d'extraction d'analyse de texte. Cette méthode consiste à extraire des données par rapport à un texte donné.

Les données sont présentes partout sur internet et dans notre quotidien également. On peut les récupérer en utilisant ces système d'extraction de données de manière efficace.Par exemple extraire des données scientifiques, sociales ou encore des informations économiques et commerciales manuellement est une tâche qui n'est pas évidente. Grâce à cela les entreprises peuvent utiliser ces données à des fins commerciales comme analyser le comportement de leur client etc.

Les réseaux sociaux comme Twitter, Facebook, YouTube, Instagram, regorgent de données plus ou moins exploitables comme par exemple des données sur le business ou sur le commerce. Les données extraites peuvent être l'adresse, le numéro de téléphone etc. Sur cet article [1], les données ont été extraites via la technique du data crawling.

1) Qu'est-ce que le data crawling ?: Le data crawling consiste à creuser profondément dans les recoins du web pour récupérer des éléments sur lesquels on aurait pu passer à côté. Ce sont des bots (des robots) qui fouillent le web pour trouver tout ce qui est pertinent par rapport à ce que l'on cherche.

Les bots agissent sur un algorithme pour suivre des instructions. Le data crawling fonctionne un peu comme le moteur de recherche Google. Le processus suit des liens vers de nombreuses pages différentes.Ils ne se contentent pas de parcourir les pages, ils collectent toutes les informations pertinentes en les indexant au cours du processus, ils recherchent également tous les liens vers les pages pertinentes du processus. Ils peuvent extraire des informations en double d'un article de blog qui ont peut-être été copié-collé car ils ne connaissent pas la différence.

Le data crawling à été utilisé pour collecter les ID Facebook et d'autres informations comme un numéro de téléphone une adresse à partir des ID facebook. La stratégie adopté est divisé en deux blocs, un pour obtenir un les ID Facebook et l'autre pour le web content et le json.

B. CredFinder: La crédibilité des tweets en temps réel

Twitter est l'un des réseaux sociaux les plus privilégiés pour diffuser de l'information aux personnes autour du globe. Cependant, le principal défi auquel sont confrontés les utilisateurs est de savoir comment vérifier la véracité des informations publiées sur ce réseau.

Dans cet article [2], un système d'évaluation de la crédibilité du contenu nommé "CredFinder" sera présenté , Ce système est capable de mesurer la fiabilité des informations par l'analyse des utilisateurs et l'analyse du contenu. Il est également capable de fournir un score de crédibilité pour chaque tweets. Par conséquent, il offre aux utilisateurs la possibilité de juger la crédibilité de l'information plus rapidement.

Souvent, les personnes malveillantes utilisent Twitter comme moyen de répandre des rumeurs visant à diffamer des marques ou même en cas de politique.Ces informations peuvent être présentées sous la forme de images qui ont été modifiées pour s'adapter à la stratégie d'attaque ciblée. Les informations de cette nature sont généralement difficiles à vérifier, ce qui peut conduire des utilisateurs naïfs à propager des informations qui n'ont aucune crédibilité et, dans certains cas, la presse écrite pourrait finir par être impliquée dans un tel scénario. Par manque de temps afin de vérifier correctement les sources.

Les chercheurs ont déjà établi que le réseau social Twitter pouvait être utilisé en cas d'urgence compte tenu de sa capacité extrêmement rapide à réduire le temps nécessaire à la communication des informations .La difficulté est donc d'évaluer la crédibilité des informations publiées, ce qui est l'objectif du travail présenté dans cet article [2].

1) Qu'est-ce que le système CredFinder ?: Le système CredFinder est un outil d'extraction de données qui évalue la crédibilité des tweets en temps réel. Le système CredFinder peut classer les tweets en fonction de leur fiabilité. Ainsi, il calcule un score de crédibilité pour chaque tweets.Le processus de calcul du score de crédibilité implique une communication entre trois parties : (I) une extension de navigateur Web, (II) le serveur système et (III) et Twitter.

Le système fonctionne en 2 parties , le côté client et le côté serveur.

2) Côté client: L'utilisateur doit télécharger une extension de navigateur web nommée "CredFinder". Cette extension collecte donc les ID de tweet que les utilisateurs font par rapport à un sujet ou un événement spécifiques. CredFinder génère alors une liste avec tous les tweet qui sont en correspondance avec le sujet souhaité avec une note en crédibilité, ici des étoiles. Une étoile sur cinq signifie que le tweet est d'une fiabilité accrue.

3) Côté Serveur: La première étape côté serveur consiste à générer des "clés uniques" (token key). Ensuite, 2 paramètres

(TID, TokenKey) sont envoyés à une fonction de validation, qui vérifie la clé.

Si la clé est invalide alors un message est envoyé à l'utilisateur pour lui demander de générer une autre clé, en d'autre termes l'ID du tweet est utilisé pour aller chercher le tweet original.

Le système CredFinder a montré au cours des tests qu'il a réalisé d'excellentes performances en termes de temps de réponse. Cependant, comme le système n'en est qu'à ses débuts, l'analyse est limitée à seulement 70 utilisateurs à ce jour. Les notes sont généralement bonnes. Mais le système doit tout de même être testé par plusieurs utilisateurs.

C. Information Extraction from Social network for Agro-produce Marketing

Les réseaux sociaux jouent un rôle important dans la vie des gens en fournissant une plate-forme qui permet aux utilisateurs de partager des idées, activités, des événements et des intérêts.La plupart des contenus sur le web se présentent sous la forme de texte non structuré, qui peut être structuré à l'aide de techniques d'extraction de données.

Dans cet article [3] il sera question du potentiel marketing des plateforme de réseaux sociaux dans le domaine de la commercialisation des produits agroalimentaires. Les agriculteurs peuvent poster des informations à propos de leur produits sur une plateforme de réseau social comme twitter. Les commerçants peuvent être intéressés par un produit en particulier en publiant un tweet. Ces tweets peuvent donc être analysés et des suggestions peuvent être générées pour les agriculteurs et les commerçants.

L'objectif de cet article est d'exploiter le potentiel marketing des sites des réseaux sociaux au profit de la société. Il y a un utilisateur central et les autres utilisateurs qui sont abonnés à cet utilisateur central. Les Tweets ont une structure prédéfinie : "@CentralUser AgroSell", Type de produit, sous-type de produit, Prix unitaire minimum, Quantité, Date etc

Les tweets sont récupérés sur le réseau social Twitter.Le processus de récupération des données utilise les API Twitter avec une architecture REST (Representational State Transfer). Les API renvoi les tweets en format JSON, XML, RSS et ATOM. Chaque tweet est associé à un utilisateur.

IV. Les différentes méthodes d'extraction

Il existe plusieurs technique et méthodes pour extraire des données, nous allons en aborder quelques une :

A. Qu'est-ce que le data scraping ?

Le data scraping est une technique par laquelle un programme informatique extrait des données depuis une source lisible par un être humain et produite par un autre programme informatique.
Le data scraping permet d'extraire des données et de les structurer depuis une source prévue à l'origine pour être lue par un humain et où l'information n'est donc pas structurée, pas documentée, et pas optimisée pour être extraite facilement. En général, le data scraping est utilisé pour interfacer un programme avec un programme plus ancien qui ne propose pas d'API ou pour extraire des données depuis une source tierce, parfois sans son accord.

B. Qu'est-ce que le data extractor ?

Le data extractor est un outil pour extraire les données.Le processus commence par l'indexation. Il explore les données des sites Web en faisant du data crawling. Après avoir creusé en profondeur les sites webs, il nous renvoie les liens des sites disponibles. Dans ces pages Web, beaucoup de données indésirables et utiles ont été mixées, nous devons donc extraire les données utiles et les convertir les données dans un format souhaité. L'extracteur de données extrait les données nécessaires dans les sites Web qu'il à explorées. Pour les séparer et utiliser les informations qu'on désire.

C. Qu'est-ce que le web wrapper ?

Le web wrapper est une procédure d'extraction de données structuré provenant de sources de données non structurées (ou semi-structurées) qui sont considérés comme des "wrapper". Il s'agit d'un processus qui implémente un ou plusieurs classes d'algorithmes, qui trouvent des données en fonction des exigences de l'utilisateur, puis en les extrayant depuis une source de données non structuré et les convertir en données structurées.

V. La stratégie et la méthode adopté

Pour commencer, nous avons choisi d'extraire les données à l'aide de l'API Twitter, comme dans l'article [3]. Car l'api est bien fournis et documentés, elle nous donne accès à la plupart des informations disponibles sur le site de Twitter.

Ensuite, il nous a fallu créer un côté BackEnd car nous avons remarqué qu'une communication directe entre le côté FrontEnd et l'api n'était pas possible du fait des erreurs de CORS.

A. Qu'est-ce que le CORS ?

Le Cross-Origin Response Sharing (CORS) est un mécanisme de navigateur qui permet un accès contrôlé aux ressources situées en dehors d'un domaine donné. Cependant, il offre également un potentiel pour les attaques inter-domaines [4].

B. Pourquoi bloquer le CORS ?

Le blocage du CORS permet d'empêcher la falsification des requêtes intersites. Imaginons que vous vous connectiez à twitter.com et que votre navigateur stocke le jeton d'authentification afin que vous soyez connecté automatiquement à l'avenir. Et pour chaque requête vers l'origine twitter.com, ces en-têtes auth-token seront présents. Imaginez maintenant un scénario dans lequel vous cliquez sur une pop-up ouvrant un site mal veillant. Ce site aura également la capacité d'effectuer des requêtes vers twitter.com. Dans ce cas, le site malveillant pourrait envoyer des requêtes en votre nom et vous pirater. Pour éviter cela, l'erreur CORS a été introduite.

C. Le Backend

La partie Backend est l'endroit ou va se jouer le plus gros du projet, c'est la partie invisible d'un site Web, toute l'algorithmie, la technique va se jouer dans cette partie. Dans ce Backend nous allons y mettre toutes les requêtes vers l'api Twitter. Nous allons aussi y mettre les requêtes de stockage vers la base de donnée et y définir les différentes routes que notre côté Frontend va récupérer.

D. L'Api Twitter

L'api Twitter va donc nous donner une clef privée qui va nous permettre à nous développeurs d'envoyer des requêtes en notre nom et récupérer les informations disponibles sur l'api.

E. Le Frontend

Le Frontend quant à lui va s'occuper d'afficher correctement les différents tweets, étant donné que l'on veut faire un site disponible au grand publique, il se doit d'être esthétique, ergonomique et facile à prendre en main.

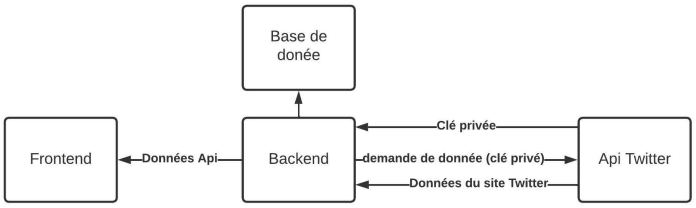

Fig. 1. Schéma représentatif du projet.

F. Résumer de l'ensemble de notre strategie

Pour résumer, L'api nous fournit une clé d'authentification, ensuite on va avoir le Frontend qui va discuter avec le Backend pour dire on a besoin de telles informations...
Le backend va utiliser la clé fournit par l'api et faire une demande de données.
L'api va lire la clé et voir que nous sommes éligibles à l'accès de ces informations et nous donner les informations.
Le Backend va renvoyer les informations perçues par l'api au Frontend.
À l'avenir, nous souhaitons que ces informations soient stockées dans une base de donnée en passant par le Backend.

VI. Les résultats de notre programme

Nous avons réussi à regrouper les différents grands blocs comme le Frontend, BackEnd, l'Api et mettre en place la plateforme pour afficher les derniers tweets de deux médias connus sur twitter 'CerfiaFR' et 'MediaAvenir'.

Fig. 2. Page de présentation du site.

Nous avons également mis en place un formulaire pour que les utilisateurs puissent nous faire des propositions de médias à mettre sur notre site.

Fig. 3. Formulaire de propositions.

VII. Conclusion

Pour conclure, nous avons créé un programme permettant d'afficher correctement les derniers tweets de médias comme 'CerfiaFR' et 'MediaAvenir' en utilisant un Backend qui va récupérer les informations à partir de l'api Twitter. De plus, nous avons mis à disposition un formulaire de propositions de médias qui va nous envoyer un mail dès lors que l'on va recevoir une proposition. À l'avenir, nous allons mettre des boutons de médias à disposition et faire en sortes qu'ils affichent les tweets de ces boutons et faire stocker toutes ces informations en base de données.

Remerciements

Nous souhaitions, en fin de cet article, remercier Mme Seddiki pour son accompagnement lors de la réalisation de cette partie du projet, de nous avoir aiguillé et de nous avoir permis d'avancer considérablement dans notre projet.

References

[1] A. Khan and B. Ratha, "Business Data Extraction from Social Networking" Utkal University, 2016.
[2] M. Alrubaian, M. Al-qurishi,M. Al-rakhami, M. Hassan, and A. Alamri, "CredFinder: a Real-time Tweets Credibility Assessing System" King Saud University, 2016.
[3] A. Khan and B. Ratha, "Information Extraction from Social network for Agro-produce Marketing"Gandhinagar, Gujarat, 2012.
[4] A. Ranganathan,"cross-site xmlhttprequest with CORS"Hack Mozilla