Contexte d'étude

Analyse de données Twitter

Twitter (`http://www.twitter.com`) est un réseau social de microblogging, qui permet à ses utilisateurs d'envoyer gratuitement des messages cours, appelés tweets. Twitter enregistre en 2022 870 millions de tweets par jour[1].

Le schéma 1 synthétise le fonctionnement de la plateforme.

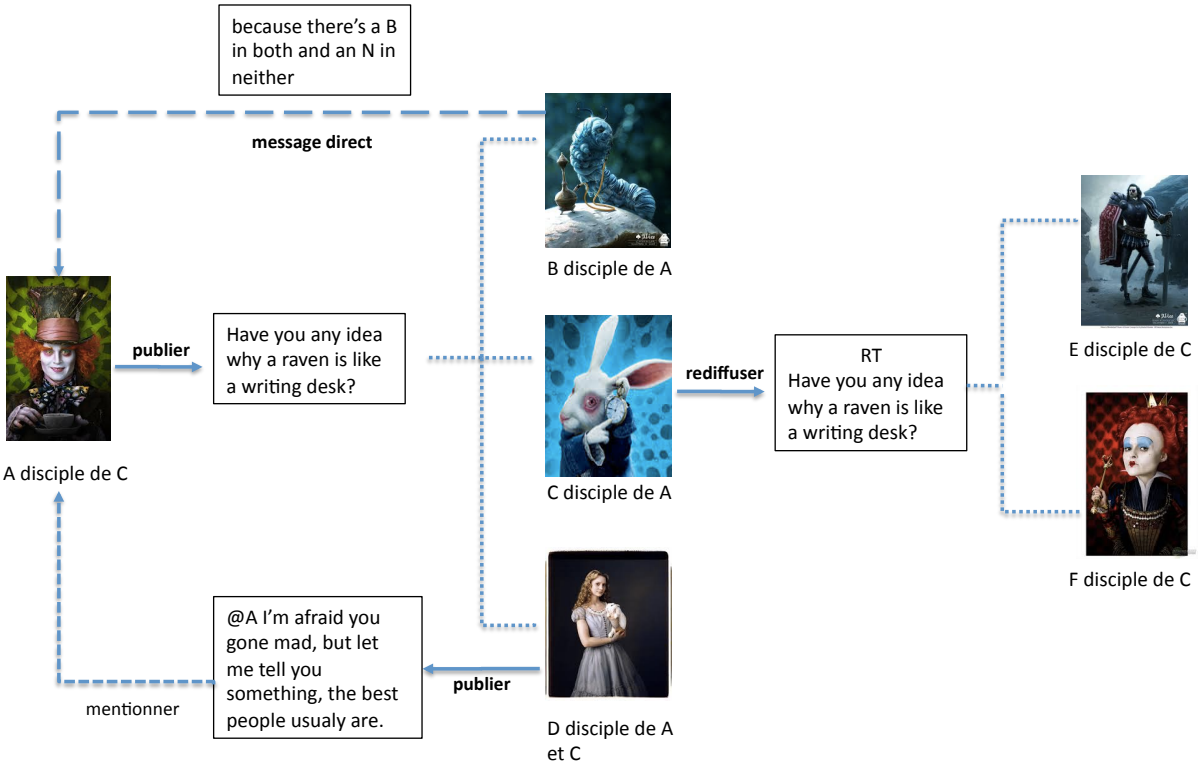

Figure 1: Fonctionnement de la plateforme de miccroblogging Twitter

Un microblogeur (un individu, une société ou encore un site d'information) peut s'abonner au flux d'autres microblogeurs (il devient un *follower*, c'est-à-dire un disciple de ces utilisateurs). Cet abonnement ne nécessite pas l'autorisation des utilisateurs concernés, et lui permet de suivre sur sa page d'accueil (sa *timeline*) toutes les publications (les *tweets*) des personnes qu'il suit. Cette timeline affiche les messages par ordre chronologique inverse de leur arrivée, c'est à dire que les plus récents sont affichés en premier. Lorsqu'un utilsateur publie un tweet, tous ses followers le voient donc, sauf si ce tweet est un message privé. Dans l'exemple de la Figure 1, le tweet "Have you any idea why a raven is like a writing desk?" est publié par l'utilisateur A. Ses disciples (*followers*) B, C et D le verront donc. Lorsqu'un tweet est rediffusé, on parle de *retweet*, et le message porte la mention RT.

[1]Source: `http://www.internetlivestats.com/`, accédé le 20/06/2022.

Dans notre exemple, C rediffuse le tweet de A qui sera ensuite vu par ses followers E et F. Un tweet peut également mentionner/s'adresser à des utilisateurs particuliers, qui sont alors directement cités dans le tweet (*@mention*). C'est le cas ici respectivement de la publication de C qui mentionne A ("@A I'm afraid you gone mad, but let me tell you something, the best people usually are") et de celle de A qui est un message direct à B ("because there's a B in both and an N in neither").

La figure 2 donne un exemple de tweet.

Figure 2: Exemple de tweet

Ce tweet fait mention via le signe @ des utilisateurs Alice, theHatter, et theWhiteRabbit. Son texte est très court (il ne doit pas dépasser 280 caractères). Il est souvent (ce n'est pas le cas ici) exprimé dans un langage spécifique, à la mode SMS. Le tweet contient également deux *hashtag*s, dénoté par le signe #. Un hashtag indique un mot important pour l'auteur du tweet, qui peut ensuite servir lors d'une recherche directe dans la plateforme. Le tweet contient également une URL. Les liens sont souvent donnés avec une forme courte, générée par des services tels que `bit.ly` ou `tinyurl.com`, en raison du nombre limité de caractères autorisés dans un tweet. Lorsque l'URL fait référence à un tweet ou encore une image (c'est le cas ici), le contenu s'affiche directement sous le tweet.

Outre le contenu du tweet, des méta-données sont associées à chaque publication :

- l'auteur du tweet (ici CHESHIRECAT), auquel est associé un profil consultable par les utilisateurs de la plateforme,
- ses jours et heures de publication,
- le nombre de fois où le tweet a été aimé ou retweeté (respectivement 111 et 15 fois dans notre exemple),
- la géolocalisation associée, si l'auteur l'a activée dans le cas d'une publication sur téléphone mobile ou tablette,
- l'information éventuelle du fait que le tweet est une rediffusion (un *retweet*).

Les administrateurs de la plateforme Twittter se servent eux des données pour :

- permettre des recherches par mots-clés de tweets et utilisateurs,
- proposer des personnes à suivre aux utilisateurs,
- afficher les hashtags les plus populaires du moment.

Modélisation relationnelle de Twitter

Un Modèle Conceptuel des Données (MCD) possible pour modéliser la plateforme Twitter est présenté sur la figure 3 :

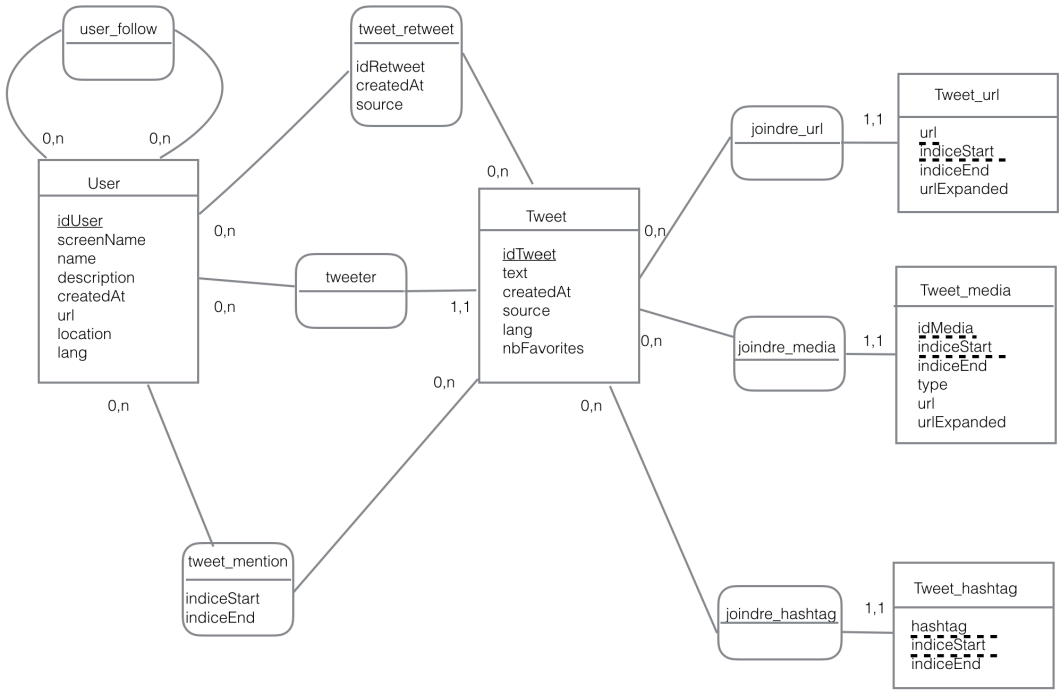

Figure 3: Modèle conceptuel des données Twitter

Ce schéma est indépendant des applications possibles, et toutes sortes d'applications sont envisageables (recherche de tweets, de *followers*, *followees*[2], recommandation de hashtags ou d'utilisateurs, affichage de hashtags populaires...). La signification des différents attributs est la suivante :

Classe d'entité / d'association	**Attribut**	**Signification**
`User`	`idUser`	Identifiant de l'utilisateur
	`screenName`	Pseudo de l'utilisateur
	`name`	Nom complet de l'utilisateur
	`description`	Texte descriptif de l'utilisateur (fourni par l'utilisateur)
	`createtAt`	Date de création du compte utilisateur
	`url`	Site web de l'utilisateur (peut être vide)
	`location`	Localisation de l'utilisateur (peut être vide)
	`lang`	Langue de création du compte, préférée par l'utilisateur
`Tweet`	`idTweet`	Identifiant du tweet
	`text`	Texte du tweet (limité à 280 caractères)
	`createdAt`	*Timestamp* du tweet (date de publication)
	`source`	Appareil sur lequel le tweet a été posté (smartphone, tablette, ordinateur)
	`lang`	Langue détectée du tweet
	`nbFavorites`	Nombre de fois où le tweet a été *liké* (aimé)
`Tweet_url`	`url`	Url (réduite - *tiny*) contenue dans le tweet
	`indice_start`	Indice du début de l'url dans le texte du tweet
	`indice_end`	Indice de fin de l'url dans le texte du tweet
	`url_expanded`	Version complète de l'URL
`Tweet_media`	`idMedia`	Identifiant relatif du media contenu dans le tweet
	`indice_start`	Indice du début de l'url du media dans le texte du tweet

[2]Personne suivies par l'utilisateur.

	`indice_end`	Indice de fin de l'url du media dans le texte du tweet
	`type`	Type du media (image, video, news, etc.)
	`url`	Url (réduite - *tiny*) du media
	`url_expanded`	Url complète du media
`Tweet_hashtag`	`hashtag`	texte du hashtag dans le tweet
	`indice_start`	Indice de début du hashtag dans le texte du tweet
	`indice_end`	Indice de fin du hashtag dans le texte du tweet
`tweet_retweet`	`idRetweet`	Identifiant de retweet du tweet
	`createdAt`	*Timestamp* du retweet
	`source`	Appareil sur lequel le tweet a été retweeté (smartphone, tablette, ordinateur)
`tweet_mention`	`indice_start`	Indice de début de la mention dans le texte du tweet
	`indice_end`	Indice de fin de la mention dans le texte du tweet

Quelques remarques concernant la modélisation:

- Les hahstags, media et url sont présents dans le texte du tweet, mais sont modélisés dans des classes d'entités séparées pour être facilement identifiés.
- Dans les classes d'entités `Tweet_url`, `Tweet_media` et `Tweet_hashtag`, les identifiants sont relatifs. Ils dépendent de l'`idTweet` correspondant.

On obtient enfin le schéma relationnel suivant:

```
User (idUser, screenName, name, description, createdAt, url, location, lang)
Tweet (idTweet, text, createdAt, urlEnd, source, lang, nbFavorites, #idUser)
Tweet_Url (#idTweet url, indiceStart, indiceEnd, urlExpanded)
Tweet_Media(#idTweet, idMedia, indiceStart, indiceEnd, type, url, urlExpanded, urlMedia)
Tweet_Hashtag(#idTweet, hahstag, indiceStart, indiceEnd)
Tweet_Mention(#idTweet, #idUser, indiceStart, indiceEnd)
User_Follow(#iduser, #idUserFollow)
Tweet_retweet(#idTweet, #idUser, idRetweet, createdAt, urlEnd, source)
```