Cas d'étude

Analyse de données Twitter

1 Cas d'étude

Ce TD se base sur des données issues de la plateforme de microblogging Twitter (`http://www.twitter.com`). Twitter enregistre en 2016 620 millions de tweets par jour[1].

Le schéma 1, synthétise le fonctionnement de la plateforme.

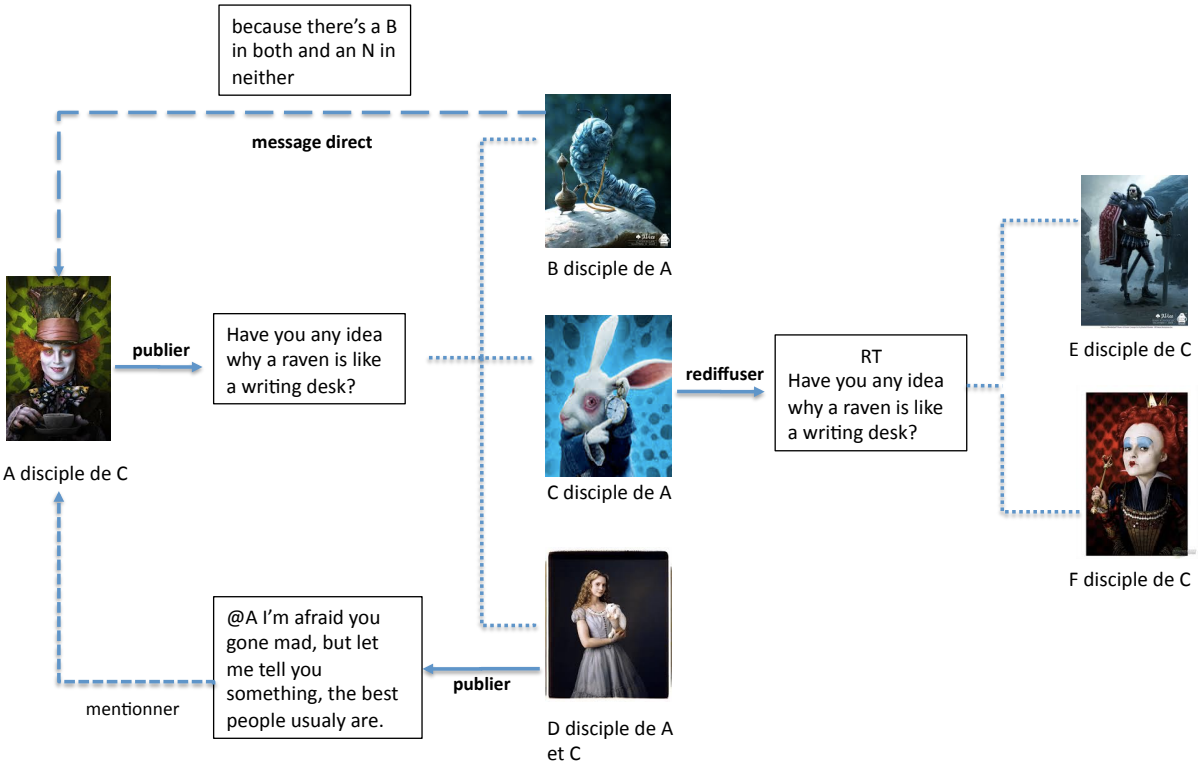

Figure 1: Fonctionnement de la plateforme de miccroblogging Twitter

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. Lorsqu'un tweet est rediffusé, on parle de *retweet*, et le message

[1]Source: `http://www.internetlivestats.com/`, accédé le 04/04/2016.

porte la mention RT. Un tweet peut également mentionner/s'adresser à des utilisateurs particuliers, qui sont alors directement cités dans le tweet (*@mention*).

La figure 2 montre 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.

2 Monde relationnel

Un Modèle Conceptuel des Données possible est présenté sur la figure 3 :

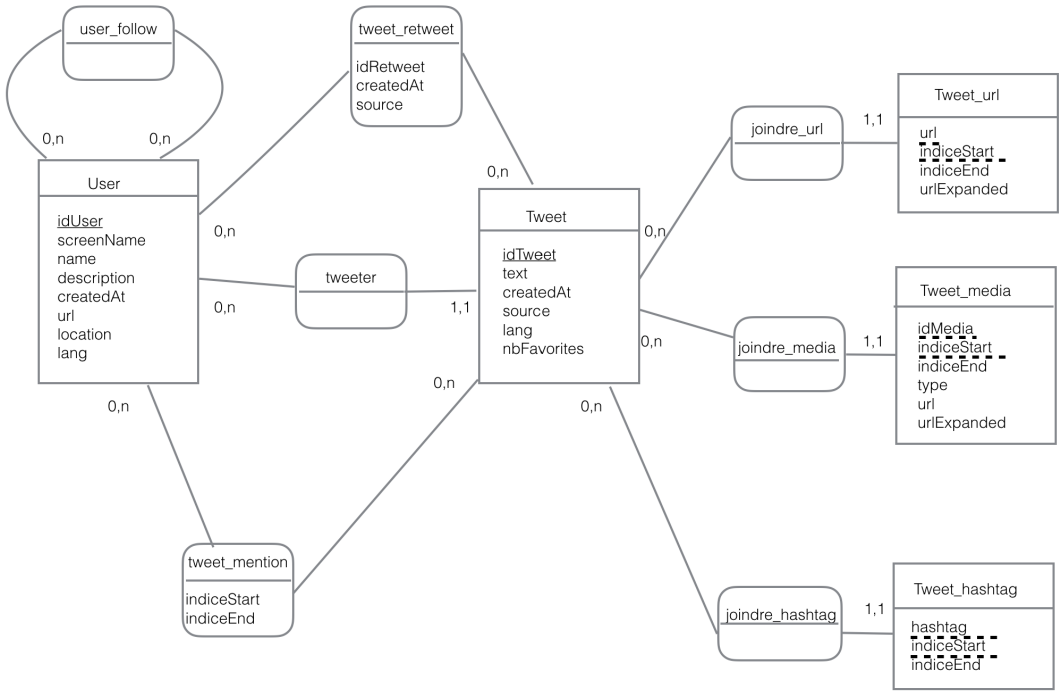

Figure 3: Modèle conceptuel des données Twitter

Comprenez et commentez ce schéma.

Ce schéma est indépendant des applications possibles, et toutes sortes d'applications sont envisageables (recherche de tweets, de followers, followees, recommandation de hashtags ou d'utilisateurs, affichage de hashtags populaires...).

Le schéma relationnel associé est le 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)
```