HAL
open science

Schéma multidimensionnel dédié pour l'OLAP des Tweets

Maha Ben Kraiem, Kaïs Khrouf, Jamel Feki, Franck Ravat, Olivier Teste

To cite this version:

Maha Ben Kraiem, Kaïs Khrouf, Jamel Feki, Franck Ravat, Olivier Teste. Schéma multidimensionnel dédié pour l'OLAP des Tweets. 9èmes Journées Francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2013), Jun 2013, Blois, France. pp.111-123. hal-04083755

HAL Id: hal-04083755

https://hal.science/hal-04083755

Submitted on 27 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive TOULOUSE Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/
Eprints ID : 12622

The contribution was presented at EDA 2013 :
http://eda2013.univ-tours.fr/

To cite this version : Ben Kraiem, Maha and Feki, Jamel and Khrouf, Kaïs and Ravat, Franck and Teste, Olivier *Schéma multidimensionnel dédié pour l'OLAP des Tweets.* (2013) In: 9èmes Journées Francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2013), 13 June 2013 - 14 June 2013 (Blois, France).

Any correspondance concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

Schéma multidimensionnel dédié pour l'OLAP des Tweets

Maha Ben Kraiem*,** Kaïs Khrouf*, Jamel Feki*, Franck Ravat**, Olivier Teste**

*Laboratoire Mir@cl, Université de Sfax
Route de l'Aérodrome km 4, B.P. 1088, 3018 Sfax, Tunisie
Maha.BenKraiem@yahoo.com, Khrouf.Kais@isecs.rnu.tn, Jamel.Feki@fsegs.rnu.tn
** IRIT, Université Paul Sabatier Toulouse III
118, route de narbonne, F-31062 Toulouse Cedex 9, France
{Franck.Ravat, Olivier.Teste}@irit.fr

Résumé. Les tweets permettent l'échange de faits et d'opinions entre les utilisateurs du réseau social "Twitter". Le nombre de tweets échangés ne cesse d'augmenter et constitue ainsi une nouvelle source importante d'informations. L'application des techniques OLAP "On-Line analytical Processing" sur ces gros volumes de tweets permet d'extraire de nouvelles informations et/ou connaissances concernant par exemple le comportement des usagers ou les sujets émergents. Cet article propose un schéma multidimensionnel générique dédié à l'OLAP des données dynamiques (tweets).

1 Introduction

Dans le monde interconnecté actuel, la vitesse de diffusion des informations amène à une formation des données tendant vers de plus en plus d'immédiateté. En effet, les grands réseaux sociaux permettent le partage et la diffusion de l'information de manière quasi-instantanée. Depuis leur apparition, les blogs et les réseaux sociaux ont rendu les internautes plus actifs au sein des réseaux participatifs. Ces derniers s'inscrivent dans une tendance technologique plus large, appelée Web 2.0, qui met le Web en valeur avec davantage d'interaction et de collaboration.

L'exploitation de ces nouvelles formes de publications suscite des applications décisionnelles exceptionnelles pour des décideurs non conventionnels. En effet, ces décideurs utilisent ces gros volumes de données comme nouvelles ressources documentaires pour y puiser de l'information qui n'était pas accessible par le biais des moyens classiques. Depuis son apparition en 2006, le site Web Twitter [1] s'est développé de telle manière qu'il est actuellement le dixième site le plus visité au monde [2] . Twitter est une plate-forme de *microblogging*. Il s'agit d'un système de partage d'informations via lequel un utilisateur a la possibilité, soit de suivre d'autres utilisateurs qui postent des messages courts, soit de diffuser ses propres messages à ses propres suiveurs ("followers").

1. http ://twitter.com.
2. http ://www.alexa.com/siteinfo/twitter.com.

En janvier 2010, le nombre de tweets échangés a atteint 1,2 milliards et plus de 40 millions de tweets sont échangés par jour en moyenne [3] . Il est à noter qu'un tweet ne doit pas dépasser 140 caractères et que les informations visibles et accessibles par les utilisateurs de ce réseau sont uniquement : l'émetteur du message, son pseudo, le texte en question, s'il est retweeté ou non.

Des applications récentes ont été proposées pour analyser l'information contenue dans un grand nombre de tweets produits au cours du temps ; citons à titre d'exemple les applications de suivi de tendances ou de repérage de buzz Cuvelier et Aufaure (2011). Toutefois, très peu de travaux se sont intéressés à la manipulation multidimensionnelle des contenus de ces tweets. Par exemple, Bringay et al. (2011) a proposé un modèle multidimensionnel en étoile dédié pour l'analyse des tweets concernant la grippe. Cette insuffisance des travaux de l'état de l'art nous a motivé à aborder la problématique d'analyse des tweets. Plus particulièrement, l'objet de ce papier est de proposer un *modèle multidimensionnel générique réutilisable* (c'est-à-dire non dédié à un besoin analytique spécifique) et *adapté au caractère dynamique des données des tweets*. En réalité, les données issues des ces tweets possèdent des spécificités particulières, le modèle devrait pouvoir permettre de représenter, charger et manipuler ces données.

Pour cet article, nous avons choisi l'organisation suivante. La section 2 propose un état de l'art des travaux qui se sont intéressés aux tweets. La section 3 décrit la structure d'un tweet. Dans la section 4, nous présentons un modèle multidimensionnel en nous basant uniquement sur les éléments constituant la structure de tweets (représentation multidimensionnelle syntaxique). Ensuite, et dans la section 5, nous étendons le modèle proposé dans la section précédente par l'ajout de certaines dimensions sémantiques. Finalement, la section 6 conclut l'article en récapitulant nos contributions et en énumérant certaines de nos perspectives.

2 Etat de l'art

L'analyse des données textuelles issues des tweets est une problématique de recherche récente en pleine effervescence. Les travaux de la littérature concernent plusieurs domaines : enseignement, détection des tendances et des événements indésirables en temps réel... Cette section décrit des travaux relatifs à ces domaines.

Plusieurs approches ont été proposées dans le domaine de l'enseignement. Parmi ces travaux, nous citons Borau et al. (2009) qui utilisent Twitter avec des locuteurs natifs de Chinois de Shanghai Jiao Tong University pour enseigner l'anglais à distance. Ils analysent les tweets des élèves et montrent comment l'utilisation de Twitter forme des compétences communicatives et culturelles.

D'autres travaux se sont intéressés à la détection des événements en temps réel, tels que : Sakaki et al. (2010) et Mathioudakis et Koudas (2010).

Dans Sakaki et al. (2010), les auteurs proposent d'analyser le contenu des tweets pour détecter en temps réel des alarmes lors des apparitions de tremblements de terre. Ils supposent que chaque utilisateur de Twitter est un capteur qui est capable de détecter un événement cible et émet un rapport probabiliste en temps réel. Enfin, pour la détection des événements et l'estimation de localisation, les auteurs proposent d'utiliser deux modèles probabilistes : un modèle temporel et un modèle spatial.

3. http ://blog.twitter.com/2010/02/measuring-tweets.html.

Les auteurs de TwitterMonitor Mathioudakis et Koudas (2010) présentent un système pour extraire automatiquement les tendances dans le flot des streams. Le système proposé est basé sur un module TwitterListener qui reçoit en entrée un ensemble de 1.2 M tweets/Jour, via une API [4] de Twitter. L'ensemble de ces tweets est par la suite transféré à un module appelé *Bursty Keyword Detection*. En effet, un mot-clé est dit *Bursty* s'il est rencontré à un taux anormalement élevé dans le stream. Par exemple, le mot-clé "NBA" apparaît généralement dans 5 tweets par minute, et soudain passe à un taux de 100 tweets/min. Cet "éclatement" de la fréquence des mots-clés est généralement associé à un coup d'intérêt populaire pour un sujet particulier et est souvent motivé par des événements émergents. Ainsi, une augmentation soudaine de la fréquence des mots clés "NBA" peut être liée à un match important NBA en cours. TwitterMonitor traite ces mots comme des "points d'entrée" pour la détection des tendances. Les mots détectés et existants dans un nombre relativement important de tweets seront, par la suite, regroupés à travers un module de *Bursty Keyword Grouping*. Le résultat de ce module est un ensemble de tendances (groupe de mots-clés) qui seront finalement analysés à travers *Trend Analysis* suivant différents facteurs (région, Temps...).

A notre connaissance, les travaux existants ne recourent pas à la technologie multidimensionnelle afin de représenter et analyser des cubes de tweets. Seuls les auteurs de Bringay et al. (2011) définissent un modèle multidimensionnel en étoile permettant d'analyser un nombre important de tweets pour une tendance particulière. Ils ont proposé une mesure adaptée, appelée "TF-IDF adaptatif", qui permet d'identifier les mots les plus significatifs selon le niveau des hiérarchies du cube (la dimension localisation). Néanmoins, l'illustration proposée concerne un domaine spécifique : évolution des maladies en utilisant le thesaurus MeSH (Medical Subject Headings).

La plupart des travaux existants propose un traitement particulier des tweets et n'offre pas d'outils généraux permettant au décideur, en fonction de ses besoins, de pouvoir manipuler l'information contenue dans les tweets combinée aux méta-données leur étant associées. De plus, nous avons noté que très peu de travaux se sont intéressés à l'utilisation de cubes pour les tweets et l'exploitation de leurs caractéristiques multidimensionnelles. Le caractère dynamique des tweets nécessite de revoir les principes employés dans le développement des cubes OLAP afin de tenir compte de certaines caractéristiques des données du tweet ; ces données changent instantanément et méritent, par conséquent, de les charger en temps réel. Pour des besoins de clarté, la section suivante sera dédiée à la présentation de la structure des tweets. Ensuite nous proposons un modèle multidimensionnel générique dédié à l'analyse en ligne (OLAP) d'informations portant aussi bien sur le contenu des tweets que les méta-données relatives à la structure de tweets (utilisateur, localisation...) voire à la sémantique contenue dans les tweets (tendance, mots-clés...).

3 Présentation des tweets

Un tweet compte au plus 140 caractères. Par contre, le code généré par un tweet fait une dizaine de lignes. En fait, un tweet est une structure contenant de nombreuses informations qui pourront être exploitées lors des analyses décisionnelles. Dans un tweet, nous pouvons trouver des champs obligatoires, comme, l'utilisateur originaire du tweet ou l'heure d'émission du

4. Exemple http ://apiwiki.twitter.com/Twitter-API-Documentation.

tweet, mais également d'autres champs dédiés à certaines fonctionnalités permettant de savoir si le tweet est tronqué, s'il est utilisé par les services SMS, son lieu d'émission ou le nombre de suiveurs du propriétaire émetteur du tweet. Ainsi, un tweet n'est pas simplement un texte mais il peut être assimilé à une structure complexe comprenant une information codée et un ensemble de métadonnées associées.

La structure complète d'un tweet est présentée dans l'annexe . Toutes les informations d'un tweet, y compris celles invisibles, peuvent être décomposées en trois parties :

- La partie ***Tweet*** contenant son identifiant, le texte émis, sa date de création, le nombre de fois que le tweet a été retweeté (retransmis) et l'application numérique d'où provient le tweet (Web, par exemple). S'il s'agit d'un tweet réponse, il contient alors en plus l'identifiant du tweet répondu, l'identifiant et le nom de l'utilisateur répondu.
- La partie ***User*** (*c'est-à-dire* propriétaire du compte Twitter) est décrite par un ensemble d'informations concernant l'utilisateur (son identifiant, son nom, le nom Twitter adopté -pseudo- et une URL), d'autres informations concernent le compte (date de création, description, lieu associé, fuseau horaire, l'offset en seconde et la langue choisie) et des informations concernant le profil (Image d'arrière plan choisie par l'utilisateur pour sa page Twitter, couleurs des caractères...).
  Cette partie User contient aussi des statistiques qui peuvent être modifiées pour chaque nouveau tweet émis (nombre de tweets favoris, nombre de tweets produits, nombre d'utilisateurs que cet utilisateur suit et nombre d'utilisateurs amis avec cet utilisateur). Notons que ces valeurs statistiques sont fortement changeantes dans le temps puisqu'elles varient, en particulier, à la suite de chaque envoi/réponse à un tweet. Leur entreposage doit nécessairement tenir compte de cette spécificité : leur changement rapide. De plus, un compte Twitter indique si l'utilisateur :
  - restreint ses messages seulement à ses amis
  - a permis ou non le marquage géographique de ces tweets
  - jouit d'un compte certifiant son identité
  - a autorisé à un autre utilisateur d'écrire en son nom ; il faut indiquer l'identité de la personne autorisée s'il y a lieu
- La partie ***Place*** caractérise l'identification de la localité associée au tweet, l'URL pour une description détaillée du lieu, nom abrégé du lieu, nom complet du lieu, type du lieu ("City ou Neighborhood"), code pays ainsi que son nom abrégé, son nom complet avec les longitudes et latitudes des intersections des côtés périmètres de ce pays ("Bounding Box").

4 Modèle multidimensionnel de base

Les tweets échangés sont associés à des méta-informations incluses dans les messages sous la forme de balises ("tags") ayant une signification ou bien non-incluses dans les messages (comme la date, la géo-localisation...). Rappelons que l'objet de ce travail est de proposer un modèle multidimensionnel dédié pour les traitements analytiques en ligne (OLAP) ou pour d'autres types d'analyses des tweets. Par ailleurs, nous visons à ce que ce modèle soit générique, c'est-à-dire comportant toutes les données issues d'un tweet et susceptibles d'être des concepts multidimensionnels de telle sorte que le modèle pourra couvrir le maximum de besoins analytiques. Il s'agit alors d'une stratégie de conception orientée données plutôt que re-

quêtes/processus d'affaires. Ainsi, le schéma que nous proposons est non dédié à un ensemble de besoins prédéterminés.

Pour ce faire, nous avons examiné minutieusement toutes les données extraites des tweets afin de juger celles qui pourraient être utiles pour des analyses OLAP. Actuellement, le résultat de cet examen a écarté les données suivantes :

- Données concernant le profil utilisateur (Image d'arrière plan choisie par l'utilisateur pour sa page Twitter, couleurs des caractères...)
- Liens entre un tweet et la personne répondue ; on se contente de savoir si le tweet a été répondu par d'autres utilisateurs ou non.
- La liste des contributeurs d'un tweet (c'est-à-dire personne ayant reçu le privilège d'écrire au nom du propriétaire d'un compte) ; par contre nous sommes restreints à l'indicateur booléen ("Contibutors_enabled") pour savoir s'il y a des contributeurs ou non.

Pour la modélisation multidimensionnelle des tweets, nous retenons le schéma en constellation Kimball (1996) de faits. La constellation est une généralisation de la modélisation en étoile, elle est constituée de plusieurs faits et de plusieurs dimensions éventuellement partagées.

Tenant compte de la spécificité des données dynamiques dans un tweet (Nombre important de tweets échangés, données qui varient très rapidement dans le temps tel que le nombre de tweets favoris pour un utilisateur), nous distinguons deux types de faits et deux types de dimensions.

Un fait peut être :

- Un *fait conventionnel* : c'est la clef de voûte du modèle dimensionnel où sont stockés les indicateurs de performances de l'activité analysée ; ces indicateurs (ou mesures) décrivent une observation donnée sur les tweets.
- Un *fait temps réel* : il s'agit d'un fait rapidement changeant, c'est-à-dire que la fréquence de mise à jour des attributs qui alimentent ses indicateurs est très élevée. Ainsi, les mesures d'un fait temps réel dépendent étroitement du temps de chargement. Il est alors conseillé de charger ses valeurs à l'instant même de leur analyse.

Une dimension peut être :

- Une *dimension conventionnelle* : c'est un axe d'analyse habituel du fait. Ses attributs sont extraits à partir de la structure générique d'un tweet.
- Une *dimension sémantique* : qui traite de la sémantique du contenu des tweets. Cette sémantique peut être exprimée au travers des tendances des tweets ou des mots-clés pertinents extraits du contenu des tweets.

Notre étude des tweets nous a permis de dégager quatre faits : deux faits conventionnels et deux faits temps réel.

Les deux faits conventionnels sont F1_Tweet et F2_User ; ils s'intéressent respectivement à l'analyse des tweets et à l'analyse des utilisateurs de ce réseau social.

- *F1_Tweet* est un fait particulier dit sans mesure ("factless fact") puisqu'il modélise les tweets en tant qu'événements n'ayant pas d'indicateur pour l'évaluer. Il est relié aux quatre dimensions *TWEET*, *USER*, *TIME* et *PLACE*. Il est utile pour répondre à des requêtes de comptage comme par exemple : Nombre de tweets par mois et par pays, nombre de tweets par langue et par année, liste des tweets d'un utilisateur pour une période de temps, etc.

– *F2_User* correspond à une observation donnée sur les comptes utilisateurs (par exemple : Nombre de compte par langue et par mois, nombre de compte par année et par localité...). Ce fait est relié à deux dimensions : *USER* et *TIME*.
  Pour ce qui est des faits temps réel, leurs mesures seront chargées au moment de l'analyse parce que les valeurs des attributs qui alimentent ces mesures varient instantanément avec les échanges du tweet entre Tweeters. A titre d'exemple, l'attribut "favourites_count" indiquant le nombre de tweets favoris pour un utilisateur dans la source diffère d'un instant t à l'instant $t + \delta t$.

Nous distinguons graphiquement un fait temps réel en le faisant précéder son nom dans le schéma multidimensionnel par le symbole de montre.

– Le fait *temps réel* appelé *F3_Pertinence* contient une mesure temps réel *Retweet_C* indiquant le nombre de fois que le tweet a été retweeté ; cette mesure permet d'indiquer le degré d'importance du tweet échangé.
– Le fait *temps réel* appelé *F4_EvolutionStat* est construit sur la partie ***User*** de la structure générique. Il contient quatre mesures :
  – Fav_C : Nombre de tweets favoris.
  – Sta_C : Nombre de tweets produits.
  – Fre_C : Nombre d'amis qu'il suit.
  – Follow_C : Nombre d'amis qui le suivent.
  Ce fait est relié aux deux dimensions *USER*, L*OADING_TIME* ; la présence de cette dernière dimension justifie que les mesures de ce fait ont des valeurs différentes à chaque chargement des tweets.

L'ensemble des dimensions de la modélisation des tweets sont les suivantes :

– *TWEET* : cette dimension est constituée des éléments de la partie tweet de la structure générique. Elle est composée d'un identifiant du texte du tweet ; ce texte est formé par un ensemble de mots et limité à 140 caractères, une source indiquant l'application d'où provient le tweet et Fav-Tw indiquant s'il s'agit d'un tweet favoris ou non.
– *USER* : composée des éléments de la partie User de la structure générique d'un tweet. Elle est composée d'un identifiant, de huit paramètres (fuseau horaire, l'offset, langue...) et de quatre attributs faibles (Nom, pseudo, description et URL).
– *PLACE* : permet l'identification, si l'utilisateur l'a permis lors du paramétrage de son compte, du nom, des coordonnées géographiques et d'autres informations complémentaires à propos du lieu qu'il a associé aux tweets.
– *TIME* : possède des paramètres allant du niveau le plus fin minute (*Min*) au niveau le plus générique année (*Year*) : Au niveau du fait *F1_Tweet*, cette dimension joue le rôle de la date de création des tweets alors que pour le fait *F2_User* elle joue le rôle de la date de création du compte Twitter.
– *LOADING_TIME* : exprime la date de chargement des données d'un fait temps réel, ces données sont fortement dynamiques.

La figure 1 présente le modèle multidimensionnel proposé pour l'OLAP de tweets.

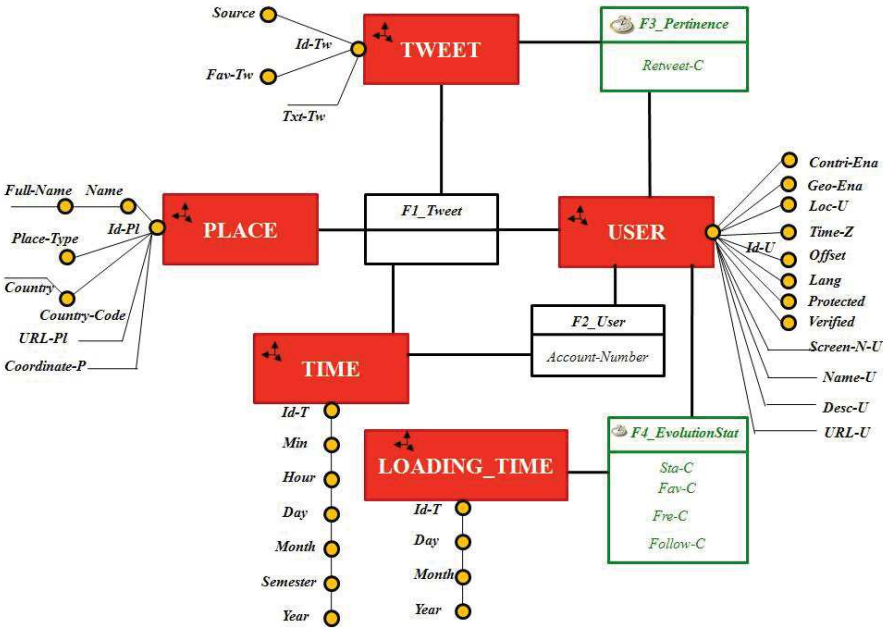

FIG. 1 – *Schéma multidimensionnel en constellation.*

5 Modèle multidimensionnel étendu

Le modèle multidimensionnel proposé dans la section précédente s'est intéressé uniquement au contenu des tweets (le texte envoyé, des informations concernant l'émetteur du message, sa localité...). Cependant, le nombre de tweets échangés est énorme et les sujets abordés sont très variés. Il serait alors fort intéressant d'intégrer des dimensions qui traitent de la sémantique du contenu de tweets. En effet, le grand réseau social Twitter, en permettant le partage, et donc, la diffusion de l'information de manière quasi-instantanée, accélère aussi la formation des tendances ("Trending topic") concernant l'actualité. Cela fait de Twitter un bon observatoire des tendances. Twitter génère automatiquement par un algorithme une liste de tendances appelée *Tendances Mondiales*, contenant une liste des 10 sujets les plus populaires à un instant donné. Cette liste de tendances permet aux utilisateurs de découvrir les nouvelles les plus fraîches en temps réel, ainsi que les sujets émergents les plus discutés.

C'est pour tenir compte de ces tendances que nous proposons d'étendre le modèle précédent par deux nouvelles dimensions permettant la contextualisation des tweets. Ces deux dimensions sont :

– *TREND* : elle correspond à des mots, des hashtags (mots précédés du symbole #) utilisés pour catégoriser les tweets ou des phrases qui ont été retweetées de multiples fois durant une période concernant un sujet émergent.

– *KEYWORD* : elle contient l'ensemble des mots constituants un tweet après suppression des mots vides, ponctuations, etc. Ces mots sont organisés sous forme hiérarchique ; l'utilisation d'une ontologie pourrait faciliter cette opération comme dans Bringay et al. (2011).

Les analyses possibles avec ces deux dimensions seront des requêtes de comptage telles que le nombre de tweets pour une tendance donnée par pays et par date.

La figure 2 montre le schéma multidimensionnel des tweets enrichi par cette contextualisation.

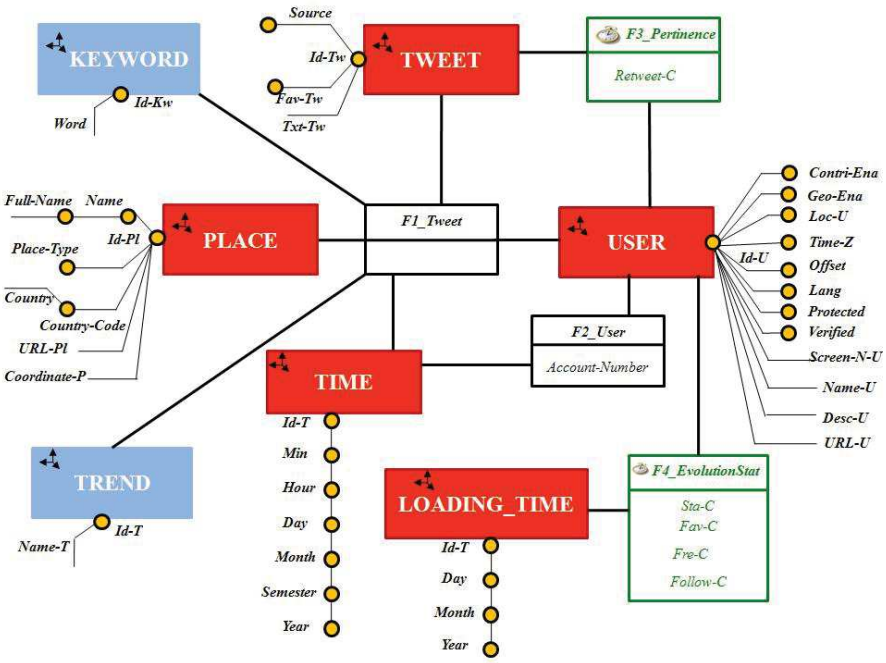

FIG. 2 – *Schéma multidimensionnel en constellation étendu.*

6 Conclusion

Vu l'intérêt grandissant accordé par la communauté scientifique aux travaux de recherche sur les tweets ainsi que l'importance des informations qui peuvent être extraites suite à l'analyse des tweets (message échangé et métadonnées), nous avons jugé important de mener une réflexion sur une modélisation permettant de faciliter les analyses de ces tweets et en particulier selon la technologie OLAP. C'est dans ce contexte, nous avons proposé dans cet article

un modèle multidimensionnel dédié à l'analyse en ligne (OLAP) des tweets échangés et leurs données dynamiques. Nous avons veillé à ce que ce modèle soit *générique*, c'est-à-dire non dédié à un ensemble de besoins prédéterminés, ce qui lui confère un potentiel analytique assez large.

Par ailleurs, nous avons veillé à ce que ce modèle tienne compte des spécificités des données issues des tweets : données fortement dynamiques. Pour cela, nous avons procédé à une extension du modèle initial par la proposition d'un nouveau type de fait, à savoir *le fait temps réel*. Un fait temps réel contient des mesures dont le chargement doit absolument être différé au moment de l'analyse et ceci afin de chercher les valeurs les plus récentes de ces données quasiment instantanément changeante. (e.g., nombre de fois qu'un tweet a été retweeté).

Egalement, nous avons enrichi le modèle multidimensionnel générique des tweets pour étendre son potentiel analytique au contenu. Dans ce sens, nous avons introduit deux dimensions sémantiques : Tendance (*TREND*) et Mot-clé (*KEYWORD*).

Comparé aux modèles de la littérature, notre modèle proposé se distingue de celui de Bringay et al. (2011) par : i) son *aspect générique* puisqu'il inclut toutes les données contenues dans un tweet et qui sont représentées sous forme multidimensionnelle ; ii) l'introduction du concept de *fait Temps réel* permettant de tenir compte de certaines données dans les tweets qui varient très rapidement dans le temps ; il incorpore également certains aspects sémantiques.

Actuellement, nous sommes en train de développer un prototype logiciel de faisabilité de nos propos ; son état est encore embryonnaire. Pour le long terme, plusieurs perspectives à ce travail sont envisageables. Il est important de réaliser des expérimentations d'analyse OLAP sur un nombre important de tweets. Nous comptons aussi proposer de nouveaux opérateurs OLAP qui tiennent compte des spécificités des données dynamiques. Nous souhaitons également exploiter des techniques de fouille de textes "Text Mining" pour extraire de la connaissance à partir de tweets et renforcer davantage la sémantique dans le schéma générique proposé dans cet article.

Références

Borau, K., C. Ullrich, J. Feng, et R.Shen (2009). *Microblogging for language learning: Using Twitter to train Communicative and cultural competence*. Berlin Heidelberg: M.Spaniol et al.(eds.), 8th International conference on Web based Learning-ICWL.

Bringay, S. A. L., P. Poncelet, M. Roche, et M. Teisseire (2011). *Analyse des gazouillis en line*. In 7ème journées francophones sur les entrepôts de données et analyse en ligne, France.

Cuvelier, E. et M. A. Aufaure (2011). EVARIST *: Un outil de monitoring de buzz et de l'e-reputation sur Twitter*. In 9ème atelier de visualisation et d'extraction des connaissances, EGC 2011.

Kimball, R. (1996). *The data Warehouse toolkit : practical techniques for building dimensional data warehouse*. John wiley & sons, ISBN 0-471-15337-0.

Mathioudakis, M. et N. Koudas (2010). *TwitterMonitor : trend detection over the Twitter stream*. In proceedings of 2010 international conference on management of data, (SIGMOD)'2010.

Sakaki, T., M. Okazaki, et Y. Matsuo (2010). *Earthquake Shakes* T*witter Users : Real-time Event Detection by Social Sensors*. In proceedings of 19th world wide web conference, WWW 2010.

Annexe

	Description	*Ligne de* code	Nom *court proposé*
Tweet	ID unique d'un Tweet	« id »	Id-Twt
	Text du Tweet	« texte »	Txt-Twt
	Date de création du Tweet	« created_at »	Dat-C-T
	S'il s'agit d'un Tweet réponse, ID de l'utilisateur répondu	« in_reply_to_user_id »	Id-R
	Le nom Twitter de l'utilisateur répondu	« in_reply_to_screen_name »	Screen-N-R
	ID du Tweet répondu	« in_reply_to_status_id »	Id-Tw-R
	S'il s'agit d'un Tweet favoris ou non	« favorited »	Fav-Tw
	Nombre de fois que le Tweet a été retweeté	« retweet_count »	Retweet-C
	L'application numérique d'où provient ce Tweet	« source »	Source
User	ID unique de l'utilisateur	« id »	Id-U
	Nom Twitter adopté par l'utilisateur	« Screen_name »	Screen-N-U
	Nom de l'utilisateur	« Name »	Name-U
	Description du compte utilisateur	« description »	Desc-U
	URL de l'utilisateur	« url »	URL-U
	Lieu que l'utilisateur a associé à son profil	« location »	Loc-U
	Date de création du compte Twitter	« created_at »	Dat-C-C
	Si l'utilisateur a permis à un autre utilisateur d'écrire en son nom	« contributors_enabled »	Contr-Ena
	Nombre de Tweets favoris de l'utilisateur	« favourites_count »	Fav-C
	Nombre de Tweets produits par cet utilisateur	« statuses_count »	Sta-C

	Nombre d'utilisateurs auxquels cet utilisateur suit	« friends_count »	Fre-C
	Fuseau horaire local associé au compte utilisateur	« time-zone »	Time-Z
	Offset en seconde	« uto_offset »	Offset
	Langue choisie par l'utilisateur	« lang »	Lang
	L'utilisateur restreint ses messages à seulement ses amis	« protected »	Protected
	Si l'utilisateur a permis ou non le marquage géographique de ces tweets	« geo_enabled »	Geo-Ena
	Nombre d'utilisateurs amis avec cet utilisateur	« followers_count »	Follow-C
	Si l'utilisateur jouit d'un compte certifiant son identité	« verified »	Verified
	Identité, s'il y a lieu, de l'utilisateur qui a rédigé le Tweet au nom de l'utilisateur officiel	« contributors »	Contri-Id
Place	Identification du lieu associé au Tweet	« id »	Id-Pl
	URL pour une description détaillée du lieu	« url »	URL-Pl
	Nom du lieu	« name »	Name
	Nom complet du lieu	« full_name »	Full-Name
	Type du lieu (City ou neighborhood)	« place_type »	Place-Type
	Code du pays correspondant à ce lieu	« country_code »	Country-Code
	Nom complet de ce pays	« country »	Country
	Longitudes et latitudes des intersections des côtés périmètres de ce pays (Bounding Box)	« coordinates »	Coordinate-P (générique)

Summary

Tweets enable exchange of facts and opinions from users of the social network *Twitter*. The number of tweets exchanged is increasing and thus constitutes an important new source of information. Application of OLAP technology (On-Line analytical Processing) on these large volumes of tweets aims to retrieve new information and/or to extract knowledge such as user behavior, emerging issues, trends, etc. This paper proposes a generic multidimensional model dedicated to the dynamic data encapsulated in tweets while taking into account the specificities of this type of data.