Modeling and OLAPing social media: the case of Twitter

Maha Ben Kraiem[1] · Jamel Feki[1] · Kaïs Khrouf[1] · Franck Ravat[2] · Olivier Teste[2]

Abstract In the recent year, social networks have revolutionized the ways of interacting and exchanging information on the Internet. Millions of users interact frequently and share variety of digital content with each other. They express their feelings and opinions on every topic of interest. These opinions carry import value for personal, academic, and commercial applications, but the volume and the speed at which these are produced make it a challenging task for researchers and the underlying technologies to provide useful insights into such data. We attempt to extend the established online analytical processing (OLAP) technology to allow multidimensional analysis of social media data. In this paper, we pursue a goal of providing a generic multidimensional model dedicated to the OLAP of social media and specially Twitter. The proposed model reflects on some specifics such as recursive references between tweets, Empty dimension, and different types of hierarchies. It is implemented using NetBeans IDE platform. We present also some experimental results. We expect our proposed approach to be applicable for analyzing the data of other social networks as well.

Keywords Twitter · Tweets · Multidimensional model · OLAP

✉ Maha Ben Kraiem
Maha.Benkraiem@yahoo.com

Jamel Feki
Jamel.Feki@fsegs.rnu.tn

Kaïs Khrouf
Khrouf.Kais@isecs.rnu.tn

Franck Ravat
Franck.Ravat@irit.fr

Olivier Teste
Olivier.Teste@irit.fr

[1] MIR@CL, University of Sfax, Airport Road Km 4, P.O.Box. 1088, 3018 Sfax, Tunisia

[2] IRIT, University of Toulouse, 2, Rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 9, France

1 Introduction

The considerable development experienced by the technologies in recent decades has led to the emergence of relatively simple panoply of Internet applications based on open source software and services designed to improve online collaboration to the large public as social networking sites, blogs, wikis, video sharing sites, hosted services, and web applications. These new services and online applications are part of a broader technology trend known as Web 2.0, which puts the Web in value with more interaction and collaboration. Since their appearance, blogs and social networks have made users more active in participatory networks. Among these social networks, the best known is Twitter that has been developed in such a way that it is currently the third most visited website in the world.[1] Twitter is a microblogging platform. More precisely, it is a special information system sharing, allowing a user to either follow other users who post short messages called *tweets*, or even to distribute their own messages. The Twitter user (said Twitto) should post information of interest to his subscribers and try to join others (notably through retrieval). Over the recent years, Twitter has experienced explosive growth and has led to generation of

[1] Twitter Team, "Twitter turns six," 2012. [Online]. Available: http://blog.twitter.com/2012/03/twitter-turns-six.html.

tremendous volumes of user-related data. In fact, this social network provides a novel way to gather data in massive quantities in real time. These data can be analyzed in various ways and have given birth to a novel area of data analysis, namely Social Media Analysis Kaplan and Haenlein (2010). By making its public stream available through a set of APIs,[2] Twitter has triggered a wave of research initiatives aimed at analysis and knowledge discovery from the data about its users and their messaging activities. Most of the works provided in the literature of this domain are tailored toward solving specific requirements. These tasks include but are not limited to trend discovery, content enrichment, user profiling, topic-based clustering, sentiment analysis, etc.

However, online analytical processing (OLAP) offers an interactive online analysis of data in an environment capable of holding large volumes of data. In addition, it provides a simple and flexible modeling to perform different types of analyses based on a pre-defined multidimensional model and including pre-calculated data that accelerate the analytical processing. Nevertheless, so far, very few studies have focused on the multidimensional modeling of data issued from tweets (content, metadata, and activities of twittos). This modeling, if it includes the entire data extracted from tweets, could be a judicious opportunity to explore these tweets through an OLAP process and, then, helps solving a wide variety of analytical needs Chaudhuri and Dayal (1997).

That is why our main objective in this paper is first to model the tweets' data and, secondly, to develop an independent application and universal analytical platform that promotes multidimensional storage of contents of tweets in order to allow analyzing ad hoc requirements without being restricted to a set of predefined needs.

The remaining of this paper is organized as follows. In the next section we discuss related work. Our generic multidimensional model is presented in Sect. 3. Section 4 describes the logical model, as well as its elaboration rules. In Sect. 5, results and analyses for testing this multidimensional model on data extracted from tweets are presented.

2 Related work

Twitter, as a new data provider, has largely contributed to the appearance of new issues related to the modeling and manipulation of data. In this context, the analysis of textual content of tweets and their metadata is a promising research topic that has attracted the attention of many researchers and has given birth to novel analysis areas, such as *Social Media Analysis*. The work related to this area can be subdivided into two major categories: those addressing the analysis of the tweets and data mining, while the others are interested in the storage of data from tweets (multidimensional modeling). Hence, the works related to both categories can be eventually divided into three major sections: (1) analysis of textual content of tweets, (2) analysis of textual content and the metadata of the tweets, and (3) analysis of textual content, the metadata, and social aspect of the messages.

We start by studying works that aimed at analyzing social media in one or another specific scenario. A spectacular novel area of *tweet's content* analysis is the detection of events, sentiments, and trends in real time. In the work of Sakaki et al. (2013), the authors propose to analyze the content of tweets in order to detect in real time alarms during the appearance of earthquakes. Moreover, the authors of TwitterMonitor: Mathioudakis and Koudas (2010) developed a system for automatically extracting trends in data streams.

A rather similar approach is proposed in the works of Phelan et al. (2009) where the authors considered RSS ("Really Simple Syndication") as a source for the extraction of information included in tweets in order to detect the various needs of Twitter users. Thus, Cuvelier and Aufaure (2011) developed a tool called EVARIST that allows a user, relying on a set of keywords (defined by the user himself), to visualize the most associated terms of Twitter, hence forming the searing news (Buzz) on the chosen topic. This tool is based on a five-step approach: (i) retrieving tweets containing keywords, (ii) cleaning the tweets (removing stop words, punctuations,…), (iii) stating the table of context with the tweets as objects and the words as attributes, (iv) building the corresponding Galois lattice (A Galois Lattice allows to group, in an exhaustive way, objects in classes, called concepts, using their shared properties), and (v) visualization of results.

Bifet et al. (2011) proposed MOA-TweetReader, a new system to perform Twitter stream mining in real time. The input items of this system are the tweets obtained from the Twitter stream. These tweets are preprocessed and converted by a tf-idf filter to vectors of attributes. The second component of the system is a frequent item miner that stores the frequency of the most frequent terms. Finally, a change detector monitors changes in the frequencies of the items.

Other approaches have been proposed for the exploitation of the contents of tweets to detect people's mood changes throughout the day. Martínez and González (2013) propose a statistical study of sentiment produced in an urban environment by collecting tweets submitted in a certain timeframe. Each tweet was processed using its own sentiment classifier and assigned either a positive or a

[2] https://dev.twitter.com/.

negative label. Lampos et al. (2013) consider changes in mood through the use of emotional words in tweets. By counting the frequency of terms related to the mood in the content of messages, they define four emotions, namely fear, sadness, joy, and anger.

Other studies, with similar objectives, focused on the analysis of the content and metadata of tweets, sentiments, and trends in real time using only the tweet's content. A pioneering work on this field was carried out by Quercia et al. (2011); these authors use Twitter-predicted users' personality types based on their Twitter activity and profile. They identified each user's type by their followers and subscription amounts and scored their personality based on how active they appeared to be on Twitter. Personality data were collected from 355 Twitter users and then used to study the relationship between user type and their personality traits. The researches could effectively predict users' personality types from their public Twitter data. Few studies were interested in the use of facets for the exploitation of tweets, such as the works of Kumar et al. (2012) and Vosecky et al. (2013). Kumar et al. (2012) have proposed a navigation system by facet called NIFT ≪ Navigating Information Facets on Twitter ≫ based on three facets: the Geo Facet shows the location of tweets in a map; subject facet is a word cloud showing the different thematic exchanged by the tweets; and the time facet presents the number of tweets in a given date. In the same context, Vosecky et al. (2013) propose a model namely Multi-Faceted Topic Model (MFTM) for the extraction of rich latent topics from tweets and temporal trends associated. They define four facets: person, organization, location, term, and time, used to exploit the stream of tweets.

Research contributions related to the Twitter analysis mostly focus on improving the search and navigation in a huge flow of messages as well as on discovering valuable information about contents, metadata, and social aspect of the tweets. We are more interested in the latter types of works. Guille and Favre (2015) developed MABED, a mention-anomaly-based method for event detection in Twitter. MABED takes the social aspect of tweets into account by leveraging the creation frequency of mentions that users insert in tweets to engage discussion and estimates the period of time during which each event is discussed on Twitter. In 2007, Java et al. presented their observations of the microblogging phenomena by studying the topological and geographical properties of Twitter's social network. They came up with a few categories for Twitter usage, such as daily chatter, conversations, information, and URL sharing or reporting news.

However, to the best of our knowledge, most existing studies mainly focus on a specific analysis of tweets, while we pursue to enable decision makers and analysts to perform descriptive and predictive analytics using the data warehousing and OLAP technologies. Nevertheless, we identified few studies that have focused on the use of multidimensional tweets (*Content and metadata*). Among these works, the one of Bringay et al. (2011) defined a multidimensional star model for analyzing a large number of tweets. However, the proposed model was dedicated to a particular trend. In order to do this, the authors have proposed an adapted measure, called "TF-IDF$_{\text{adaptive}}$," which identifies the most significant words according to level hierarchies of the cube (the location dimension). Nevertheless, their case study deals with a specific area: the evolution of diseases, referring to the thesaurus MeSH (Medical Subject Headings) by adding to their multidimensional model a dimension called MotMesh (MeshWords).

Rehman et al. (2012) proposed a system for warehousing streams from Twitter. Their system lies on an architecture consisting of five layers: (i) the data source layer represented by the available Twitter APIs, (ii) the ETL layer (Extract, Transform, and Load) for the extraction of data from tweets and processing in a suitable format for the target database, (iii) the data warehouse layer for the storage of data issued from tweets, (iv) the analysis layer dedicated for OLAP analyses of the tweets, and (v) the presentation layer of analysis results. Then, the authors present an extension of their work Rehman et al. (2013) where they focus on integrating extensive natural language processing capabilities in OLAP to perform multidimensional social media analysis. A rather similar approach is proposed in the works of Mansmann et al. (2014). The authors propose to extract multidimensional data cubes for OLAP from semi-structured datasets and to extend the resulting model by including dynamic categories and hierarchies discovered from the data through Data Mining methods and other computations. Liu et al. (2013) presented a text cube to analyze and model human, social, and cultural behavior (HSCB) from the Twitter stream in a textual database. They introduced a text cube approach mainly focused on sentiment analysis and visualization.

Other studies have simply focused on the automatic extraction of information when available, in order to supply a hierarchy and then associate a tweet to a specific geographic location in order to facilitate multidimensional analysis. Among these works, we cite the approach proposed by Bouillot et al. (2012) where the authors analyze in a first step the content of tweets in order to retrieve the relevant terms that might correspond to a specified location. This step is performed using appropriate specific models. Then, the authors retrieve location information from the metadata tweets and try to identify the geographic location from the location information extracted, and, if such information is not available, they use the time zone to estimate the location.

Further to this study, we may conclude that most of these works ensure a special treatment of tweets but do not offer tools for the decision maker to manipulate the information contained in the combined metadata associated with their tweets.

In addition, we notice that very few studies have examined the use of cubes for tweets and the exploitation of their multidimensional potential. Hence, our aim is to provide a multidimensional model dedicated to the *content*, *metadata*, and *social aspect* of tweets that is generic (i.e., independent of the special needs pre-defined a priori) and taking into account the structural specificity and possibly semantic data.

3 OLAP modeling of social media

3.1 Concepts

Conceptual modeling provides a level of abstraction independently of technical aspects and focusing on decision-making needs. The multidimensional modeling consists in defining the subject to be analyzed as a point in a multidimensional space Kimball (1996). In fact, data are organized in such a way to bring out the subject of analysis represented by the concept called *fact*, composed of *measures* corresponding to the additive information of the analyzed activity as well as the *dimensions* of this activity representing analysis axes. A dimension is composed of attributes expressing the characteristics according to which the measures of the fact are analyzed (i.e., activity). The attributes of a dimension can be organized into hierarchies, from the finer to the most general granularity. Relying on the fact and dimension concepts, it is possible to build different multidimensional models; the most popular one is the star model. A star model is composed of one central fact surrounded by dimensions, whereas the constellation model consists in defining a set of facts that share common dimensions.

The provided dataset from social media already displays some favorable characteristics for data warehousing, such as being *temporal* (by including the time dimension), *non-volatile* (no modifications of existing entries), and *measure centric* (maintaining accumulative counters). A straightforward mapping of this set of attributes to a multidimensional perspective results in the identification of cubes for storing the contents and the metadata of the messages and for storing the statistical measurements provided with each record, respectively. However, the multidimensional data model and implementations come with a set of further constraints, such as homogeneity, atomicity, summarizability, avoidance of NULL values, etc., which are not met by the input dataset. Moreover, the specificity of social aspect (the Tweet/Tweet-responses) requires reviewing the principles used in the implementation of the OLAP cubes in order to reflect their characteristics.

For multidimensional modeling of social media, we retain the constellation schema and then suggest some extensions in order to reflect the specificities of such data.

Conventionally, a constellation is composed of facts inter-connected, by common dimensions.

- A **constellation C** is defined by (F_i; D_i; *StarC*) where
- $F_i = \{F_1, \ldots, F_n\}$ is a non-empty set of $n \geq 1$ facts,
- $D_i = \{D_1, \ldots, D_m\}$ is a set of $m \geq 1$ dimensions, and
- *StarC*: $F_i \rightarrow 2_i^D$ associates each fact to the set of dimensions, according to which it can be analyzed.

3.1.1 Fact and its components

We have extended the classic concept of fact to add a ***reflexive relationship*** (denoted R) between ***fact instances*** that allows connecting an instance of the fact to one or several instances of the same fact:

Definition 1 $\forall i \in [1, \ldots, n]$, a **fact** $\boldsymbol{F_i}$ is defined by ($NameF_i$; M_i; INS_i; R_i) where

- $NameF_i$ is the name identifying the fact F_i in the constellation,
- $M_i = \{m_{i1}, \ldots, m_{ik}\}$ is a set of k measures of F_i,
- $\mathrm{INS}_i = \{\mathrm{ins}_{i1}, \ldots, \mathrm{ins}_{il}\}$ is the set of l instances of fact F_i, and
- R_i: $\mathrm{INS}_i \rightarrow \mathrm{INS}_i$, as $R(\mathrm{INS_i}) \subseteq \mathrm{INS}_i$.

Note 1 According to this definition, we define a new ***Fact–Fact relationship*** with a minimum multiplicity of "**0**" at the end of the reflexive fact. This multiplicity allows the existence of fact instances that are not related with any other fact instance.

Definition 2 $\forall j \in [1, \ldots, k]$, a **measure** $\boldsymbol{m_{ij}}$ is defined by ($Name_{ij}$; t_{ij}; f_{ij}) where

- $Name_{ij}$ is the name of the measure,
- t_{ij} is the type of the measure, and
- f_{ij} is a set of aggregation functions, compatible with the summarizability property (i.e., additivity) of the measure, $f_{ij} \subseteq \{\mathrm{SUM}, \mathrm{AVG}, \mathrm{MAX}, \ldots\}$.

In order to take into account the specificities of data extracted from social media, we distinguish three types of measures: *numerical* measures, *textual* measures, and measure composed of *list of elements*.

- A numerical measure has numerical values.
- A textual measure is a measure which content is a string (one or several words).

Table 1 Measure types and their aggregate functions

Type of measure	Aggregate functions allowed
Numerical	Arithmetic functions (SUM, AVG, MIN, MAX,…), COUNT
Textual	TOP_KW (Ravat et al. 2008), COUNT
List	AVG_KW(Ravat et al. 2008), COUNT

- A measure *list of elements is* composed of a list of keywords, representing the most significant words of a tweet: hashtags in our case (a hashtag is a word or an unspaced phrase pre-fixed with the hash symbol (#) indicating the subject assigned to the message).

The OLAP environment offers many aggregate functions, depending on the type of measure. Some of these functions are still convenient to these measures. Table 1 shows the possible aggregate functions by measure type.

3.1.2 Dimension and its components

Definition 3 $\forall i \in [1,\ldots, m]$, **a dimension D_i** is defined by ($NameD_i$; A_i; $A'i$; H_i, TD_i) where

- $NameD_i$ is the name identifying the dimension in the constellation,
- $A_i = \{a_{i1},\ldots, a_{iz}\}$ is the set of z dimension attributes (parameters and weak attributes) extracted from *raw data*,
- $A_i^r = \{a'_{i1},\ldots, a'_{iq}\}$ is the set of q dimension attributes extracted from *calculated data*,
- $H_i = \{h_1,\ldots, h_{ip}\}$ is the set of p hierarchies showing the arrangement of the attributes of D, and
- TD_i is the type of dimension.

Definition 4 $\forall j \in [1,\ldots, z]$, an **attribute a_{ij}** is defined by ($Name_{ij}$; DOM_{ij}; VAL_{ij}) where

- $Name_{ij}$ is the name of the attribute,
- DOM_{ij} is the domain of the attribute (String, Number,…), and
- VAL_{ij} is a Boolean to indicate that the attribute is not always valued.

Note 2 According to this definition, we distinguish two types of dimensions:

- A *Non-Empty dimension* composed of different attributes and hierarchies.
- An *Empty Dimension*: it is a dimension with ***Null*** for all the attributes.

Taking into account this type of dimension (Empty Dimension), we define an ***Incomplete Fact–Dimension Relationship*** where the minimum multiplicity at the end of the dimension is "***0***." In fact, an association between a fact and a dimension is complete if, for every fact instance, there exists a dimension instance which is related to that fact instance; otherwise the association is incomplete.

Definition 5 $\forall j \in [1,\ldots, p]$, A **hierarchy h_{ij}** is defined by ($Nameh_{ij}$; P_{ij}; $WEAK_{ij}$) where

- $Nameh_{ij}$ is the name which identifies the hierarchy in the constellation,
- $P_{ij} = \{p_{i1},\ldots, p_{iy}\}$ is the set of parameters of the hierarchy, and
- $WEAK_{ij}$: $P_{ij} \rightarrow 2^W$ associates each parameter to a possible empty subset of weak attributes of the dimension of h_{ij}.

Note 3 By examining data extracted from social media, we notice some nulled attributes, empty set. For this reason, we define two types of hierarchies:

- *Simple Hierarchy*: a hierarchy where all the members exist and have a value (Not Null); it allows for obtaining views of data with different granularity, i.e., summarized or detailed through roll-up and drill-down operations, respectively.
- *Non-covering hierarchy:* a hierarchy where some members may have a *Null value*.

3.2 Graphical formalism

Associated with all the concepts, we define the graphical formalism showing the extended multidimensional model. Figure 1 shows our graphical formalism. This modeling has provided solutions to the specifics of data extracted from social media where many types of dimensions and hierarchies are defined.

3.3 Application to tweets

In order to extract facts and dimensions from the semi-structured and unstructured social media data, it is important to first understand the underlying data model of Twitter. A tweet is a short message which contains less than 140 characters. On the opposite, the generated code for a tweet is a dozen-line size. In fact, a tweet is a data structure containing several information (user data and metadata) that could be used in decision analyses. This structure is composed of mandatory fields visible to twittos (i.e., Twitter users) such as the author of the tweet or the

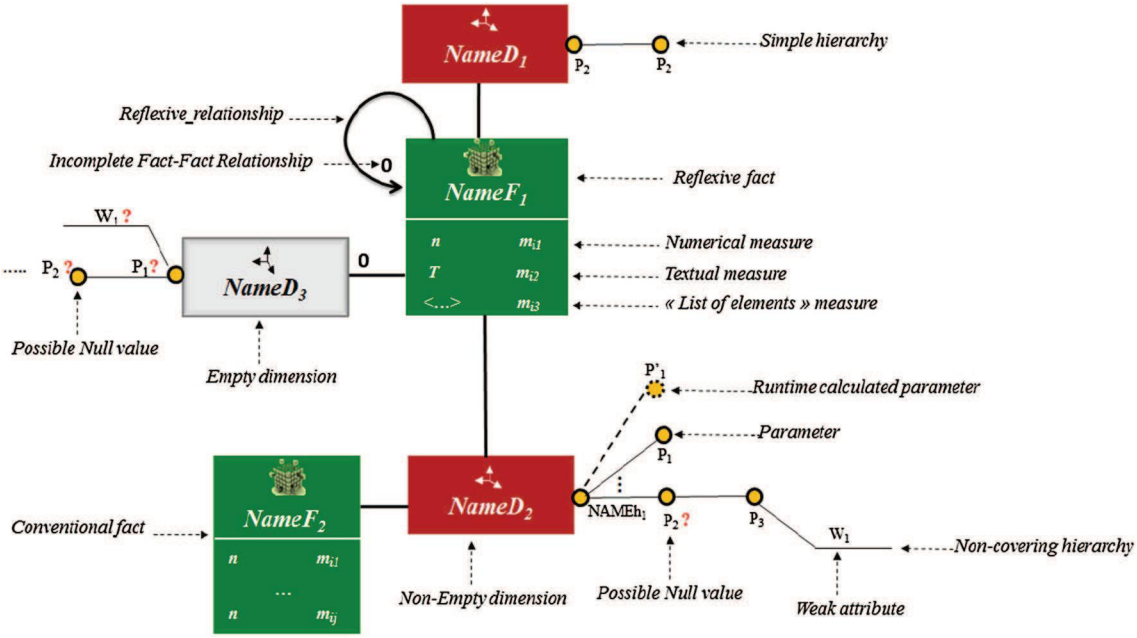


Fig. 1 Graphical formalism of multidimensional constellation schema

tweet's creation date; it also contains other hidden fields. Users can be friends or followers of other users, be referenced (i.e., tagged) in tweets, be authors of tweets, or retweet other users' messages (Appendix 1).

Given a stream of tweets, our main objective is the definition of multidimensional model dedicated to the OLAP of tweets. Moreover, we aim to ensure that this model is generic, i.e., containing all the data from a tweet (*Textual content*, *metadata*, and *social aspect*). Hence, we examined all data of the tweets in order to judge those that could be potentially useful for OLAP analysis. Following this review, we excluded the following data which we considered a bit useful or even useless:

- Data describing the user profile (e.g., user profile's photo, background image chosen by the user for his own Twitter page, colors for characters and bars,…).
- List of contributors of a tweet (i.e., a collection of brief user objects (usually only one) indicating users who contributed to the authorship of the tweet, on behalf of the official tweet author); however, we are restricted to the Boolean indicator ("Contributors-Enabled") to point out whether this account has enabled contributors.

To perform a complete analysis on data extracted from tweets and the activity of twittos, we propose a constellation composed of two facts: *Activity-Twittos* and a reflexive fact called *Activity-Tweet*.

- Activity-Twittos fact: it corresponds to observations on user accounts and allows the analysis of the user activity over time. For this fact, we define the four numerical measures as follows:
  - *Fav_C*: Number of favorite tweets a user may have;
  - *Sta_C*: Number of tweets of a user;
  - *Fre_C*: Number of followings; and
  - *Follow_C*: Number of followers.
- Activity-Tweet fact: it is a reflexive fact. It models links between a tweet and the person concerned by the answer (answered person) and then allows participants and other readers to easily follow the exchange of tweets (cf. Fig. 2). Being reflexive, it links instances of the same entity. It is composed of measures of type *Textual* (the 140 tweet's characters), *List of elements* (Hashtags), and *Numeric* (Retweet-c) characterizing the number of times a tweet was retweeted and indicating the degree of importance of exchanged tweets.

While loading dimension records that we have set for the modeling of tweets, we can identify three types of dimensions:

Case 1: Raw data Dimension It is a dimension composed of the metadata explicitly available: raw data of the original set of tweets (Appendix).

- PLACE dimension: An Empty dimension that allows the identification of the user (if enabled by the user during the configuration of his/her tweet account), the name, the geographical address, and phone number (coordinates), in addition to other information about the place associated with tweets.

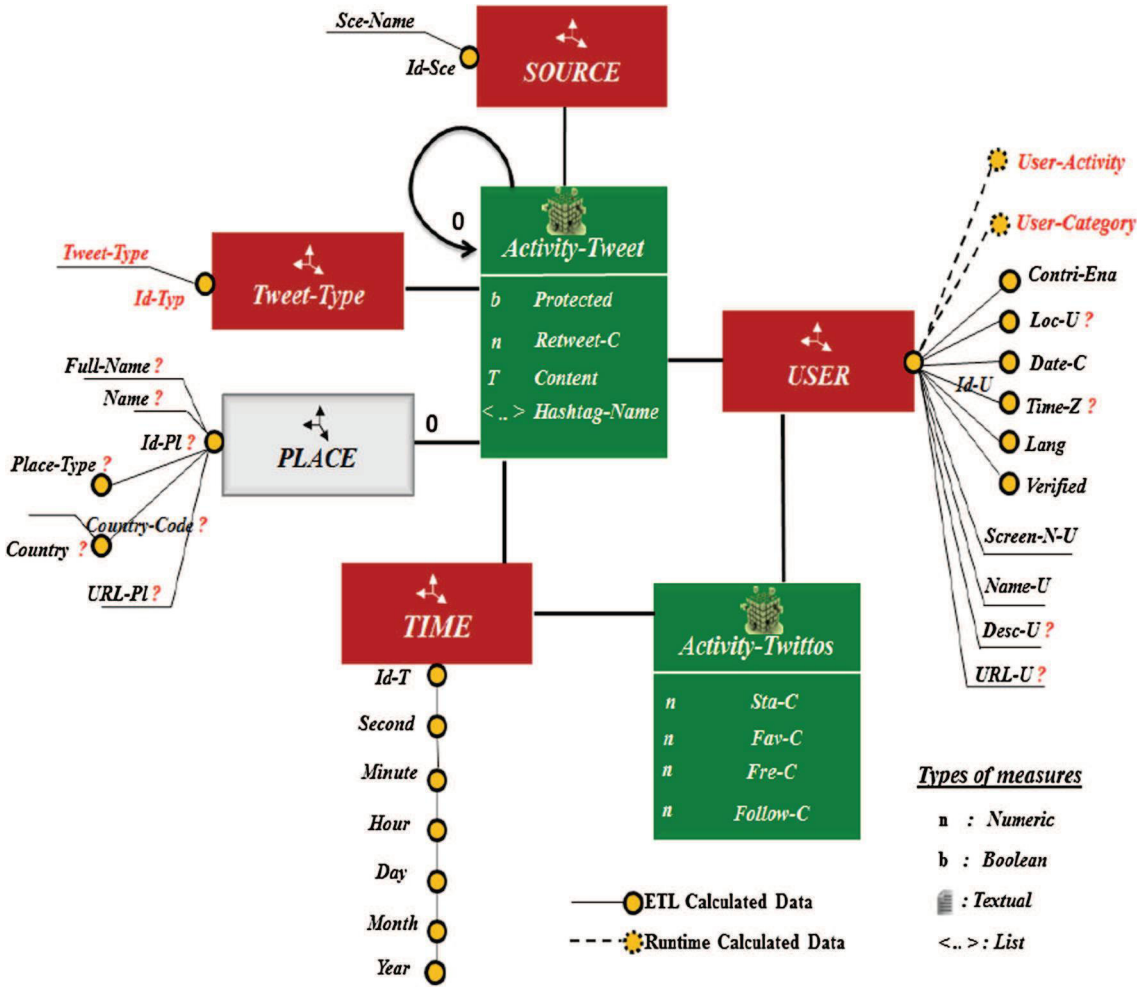


Fig. 2 Multidimensional constellation schema dedicated for the OLAP of tweets

- TIME dimension: It has parameters organized from the finest level (Second) to the most general one (Year). For the Activity-tweet fact, this dimension timestamps the creation of tweets.
- SOURCE dimension: Utility used to post the tweet, as an HTML-formatted string. Tweets from the Twitter website have a source value of web.

Case 2: Raw and Calculated data Dimension So far, we restricted ourselves to the metadata explicitly available: *raw data* of the original set of tweets, for constructing the cube. However, some dimensions can be enriched by additional elements. These new elements are *pre-calculated data* derivable from existing raw data by applying some functions. This dimension is:

- USER dimension: it is composed of elements coming from the user data and the metadata of a tweet and pre-calculated data. This dimension has an identifier, four weak attributes (name, screen-name, description, and URL), and eight parameters (language, verified,…) including six parameters extracted from raw data and two pre-calculated attributes (parameters) *User Category* and *User Activity.*

– *User Category* is based on the user's category in terms of the number of that user's followers and friends (Java et al. 2007). There exist different methods dealing with the classification of the Twittos, but most of them agree on the prevailing role of the number of followers and friends. We define formula (1) in order to classify users into three categories.

$$\text{User Category} = \text{FollowerCount} / \text{FriendCount} \quad (1)$$

with

– *FollowerCount*: Number of followers for a user, extracted from his most recent tweet;
– *FriendCount*: Number of users a user is following, extracted from his most recent tweet.

Table 2 User category

Twittos category	Category description
Information seeker	Person who posts rarely (*User Category* < 0.8), but follows other users regularly
Information sharing	This user posts tweets frequently (*User Category* > 1) and has a large number of followers due to the valuable contents of his tweets
Friendship relationship	Equivalence between friends and subscribers ($0.8 \leq$ *User Category* ≤ 1)

These three categories are explained in Table 2.

- *User Activity* reflecting the frequency of tweeting relative to the period elapsed since the creation date of the user's account. We note that other studies in the literature have adopted the number of retweet to assess the activity of users. However, with the number of retweet, we lose a part of the user activity (his own tweets). We define formula (2) to determine the activity of a user:

$$\text{User Activity} = \text{StatusCount}/\ \text{TimeDif} \quad (2)$$

with

  - *StatusCount*: Most recent number of tweets for a user;
  - *TimeDif: The period elapsed since the creation* date of the user's account, computed from other data fields.

This category should classify each user into one of the following four clusters: **"Old and active," "New and active," "Old and passive,"** and **"New and passive"** for those users who registered long ago or recently and who tweet more or less frequently.

Calculated elements may evolve over time along with the evolution of the dataset (User Category, User Activity); in this case, these elements are calculated at the time of analysis. Once an element has been added to the dimension and populated with values, it can be used in OLAP queries.

Case 3: Calculated data Dimension

- *Tweet-Type*[3] dimension. This should classify each tweet into one of the following four types: "***Normal-Tweet"*** (every message comprising less than 140 characters posted on Twitter); "***Mention"*** (tweets containing the Twitter username of another user, prefixed by the "@" symbol, as follows: Hi @NeonGolden! Are u ok?); **"*Responses*"** (tweet beginning with the username of another user and is in response to one of his Tweets as follows: @ NeonGolden How can you think that this movie was bad?); and finally **"*Retweet*"** (a tweet starting with the symbol RT).

Figure 2 depicts the multidimensional model for tweets extended with some specifics and enriched with these extensions highlighted in red color. The cardinality 0 of a reflexive fact means that a tweet is not necessarily an answer to another tweet. The second specificity is relative to the possibility of having tweets without any associated locality (absence of the PLACE dimension). This aspect is taken into account by our model. Indeed, we defined a relationship of type 1:0 between the fact Activity-Tweet and the PLACE dimension. This occurs when the twittos did not allow, during the configuration of his Twitter account, the identification of the place which he associated with tweets.

4 Logical modeling: R-OLAP

Once the conceptual model is defined, the logical model can be derived by applying a set of transformation rules Vassiliadis (1999). In this section, we present the main transformation rules of a constellation into R-OLAP logical model. Although there are various types of R-OLAP models, we decided to detail the transformation rules for the denormalized R-OLAP model. This model is the most used because less joins are needed during queries execution.

We transform the proposed model into R-OLAP logical model according to the following set of three rules:

- Each dimension *D* is transformed into a relational table which columns are parameters and weak attributes of all hierarchies of *D*. The primary key of the table is the attribute of the finest level of granularity of hierarchies of *D*.
  For *Empty Dimension*, our solution is to generate a surrogate key with Null for all the other attributes.

- Each fact *F* is transformed into a relational table of the same name which columns are measures of *F* and foreign keys referencing the dimensions connected to *F*. For a reflexive fact, the primary key contains an additional attribute (Id-Activity-Twt) and a foreign key (Id-Activity-Twt-P) which can contain either a *null value*, or *only values from an existing tweet* (Id-Activity-Twt). In other words, when a foreign key value is used, it must reference a valid and existing primary key in the parent table. The reflexive relationship is supported by the referential integrity constraint, which, when satisfied, requires every value of one attribute (column) of a table (the fact Activity-Tweet in our case) to exist as a value of another attribute in the same

[3] https://support.twitter.com/articles/119138-types-of-tweets-and-where-they-appear.

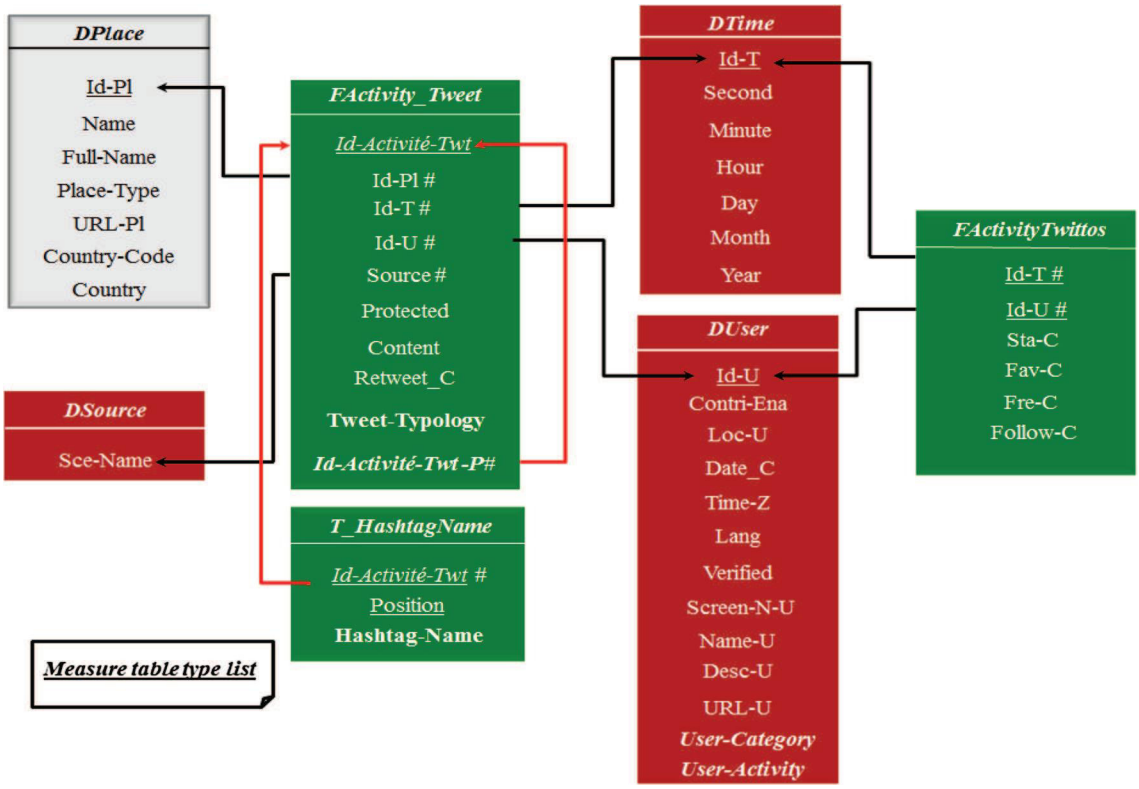


Fig. 3 R-OLAP logical model for constellation of Fig. 2

table. For a non-reflexive fact, the primary key could be either the concatenation of its foreign keys or a surrogate key.

- Each measure of type *List of elements* is transformed into a relational table called T-*MeasureName* containing the primary key of the corresponding fact. The primary key of a *T-MeasureName* table is the concatenation of the primary key of the fact with an additional attribute (position of a hashtag in our case).

Example The measure *Hashtag-Name* of type *List of element* (Fact Activity-Tweet) becomes a relational table named *T-HashtagName,* where Id-Activity-Twt and Position of a Hasthag in the tweet are primary keys of the table.

Figure 3 depicts the R-OLAP model resulted from the transformation process of the multidimensional constellation diagram (cf. Fig. 3).

5 Experimental results

In order to evaluate our approach, we have elaborated a software prototype called *OLAP4Tweet,* developed using JAVA and ORACLE 10 g database, since it offers a stable database environment. Figure 4 depicts the architecture of *OLAP4Tweet,* which is composed of four modules namely:

- Twitter platform as a data source,
- Multidimensional schema design,
- Extract, transform, load (ETL), and
- Querying.

Next, we explain each of these modules.

The *data source* is represented by the available Twitter APIs kalucki (2010) for data streaming. In fact, The Streaming API provides real-time access to tweets in sampled and filtered form. This API is HTTP based; GET, POST, and DELETE requests can be used to access the data. In Twitter terminology, *individual messages* describe the "status" of a user. The Streaming API allows near real-time access to subsets of public status descriptions, including replies and mentions created by public accounts. The dataset delivered by the Twitter Streaming API is a semi-structured data file conforming to JSON (JavaScript Object Notation) output format. Each tweet is streamed as an object containing 67 data fields.

The *multidimensional schema design* module is based on three steps. It aims to propose a multidimensional model dedicated to conventional online analytical processing and, in addition, should allow more elaborate treatments of tweets.

The *ETL* module takes care of capturing the original data stream, bringing it into a format compliant with the target database and feeding automatically the various components of the multidimensional model (fact, dimensions, and parameters) issued from the tweets using Hibernate software and Oracle 10 g.

Finally the *Querying module*. Once the multidimensional model is generated and loaded with data, the decision maker can perform OLAP analyses on tweets using the OLAP tool offered by the implementation platform (e.g., Oracle Discoverer).

In order to assess our proposed approach for modeling and OLAPing of tweets, we have loaded a data set containing 72,000 tweets collected by crawling 2 days of public tweets (6–7/11/2014). These tweets are written in different languages (cf. Fig. 5). It should be noticed that, among these tweets, only 1010 (1, 40 % of tweets) were associated with a place and 12,085 are tweet responses.

Once these tweets were loaded into our generic multidimensional model of Fig. 2, the number of instances for each multidimensional component is given in Table 3.

Once the model has been populated with values, OLAP queries could be executed. We now demonstrate the power of applying OLAP when solving specific Twitter-related analysis tasks.

To evaluate our approach, several experiments were conducted. We proceed by presenting two cases of Analysis: *Mono-Dimension Analysis* and *Multi-Dimension Analysis.*

Case 1: Mono-Dimension Analysis We start by studying the new type of fact *reflexive fact*. This type allows connecting an instance of the fact table to one or several instances of the same table. We present below an example of query using the reflexive fact Activity-Tweet.

```
SELECT              LEVEL, IDACTIV, CONTENTTWEET, SCREEN_N_U, IDACTIVREPONSE
FROM                ACTIVITY_TWEET A, DUSER D
WHERE               A.ID_U = D.ID_U
CONNECT BY PRIOR    IDACTIV = 530673967825031168
START WITH          IDACTIVREPONSE IS NULL
ORDER BY            IDACTIV
```

LEVEL	IDACTIV	CONTENTTWEET	SCREEN_N_USER	IDACTIVREPONSE
1	530673967825031168	A5.)Mac & Cheese Muffins #FoodieFriday	@rashmirnaik	NULL
2	530674721235677184	@ rashmirnaik Ans 5 - "Mac & Cheese Muffins" #FoodieFriday 18	@infoodnetwork	530673967825031168
3	530674826109673473	@infoodnetwork @ rashmirnaik I will be there	@CocaCola	530674721235677184

Then, we have executed another query showing the number of tweets per language. For language recognition, we have used language detection APIs, the one offered by JSON. When detected, the language information can be used for analysis and aggregation (Table 4).

Figure 5 gives the distribution of the 72,000 loaded tweets per tweet language.

We are now interested in the variation of the number of users by User Category. Figure 6 plots the number of all users derived from the dataset and shows the distribution of each user category.

Conversational posts (Java et al. (2007)) provide the building blocks of the social interaction between users which leads to the development of community, creation of interpersonal relationships, and the perception of reciprocity between Twitter users and their followers. These conversations are based on *Tweet-Response, Mentions, and Retweet* (*Tweet-Type*). Figure 7 shows all detected entity types. Each tweet was scanned to associate it with a Type.

Case 2 Multi-Dimension Analysis

Social media in general and Twitter in particular have changed the way people socialize and share content on the Internet. Twitter continues to grow at a record pace. For this reason, we study the evolution of Twitter accounts created per year (the creation date for the Twitter account) and language (Fact: Activity-twittos). We notice that, since Twitter was launched, this service rapidly gained worldwide popularity, in a way that the service quickly became popular, and also we have identified that most users are from the United States. In addition, Twitter launched in Japan, almost exactly 3 years ago, is largely in response to the popularity of Japanese.

Then, we study afterward the distribution of users analyzed by Source and Date (The UTC date time that the user account was created on Twitter: Dimension User). *We notice that we have chosen the most relevant source for this analysis.* The results presented in Table 5 lead to the following observations: The number of users' accounts is more and more important from 2007 to 2012 especially for the Web source.

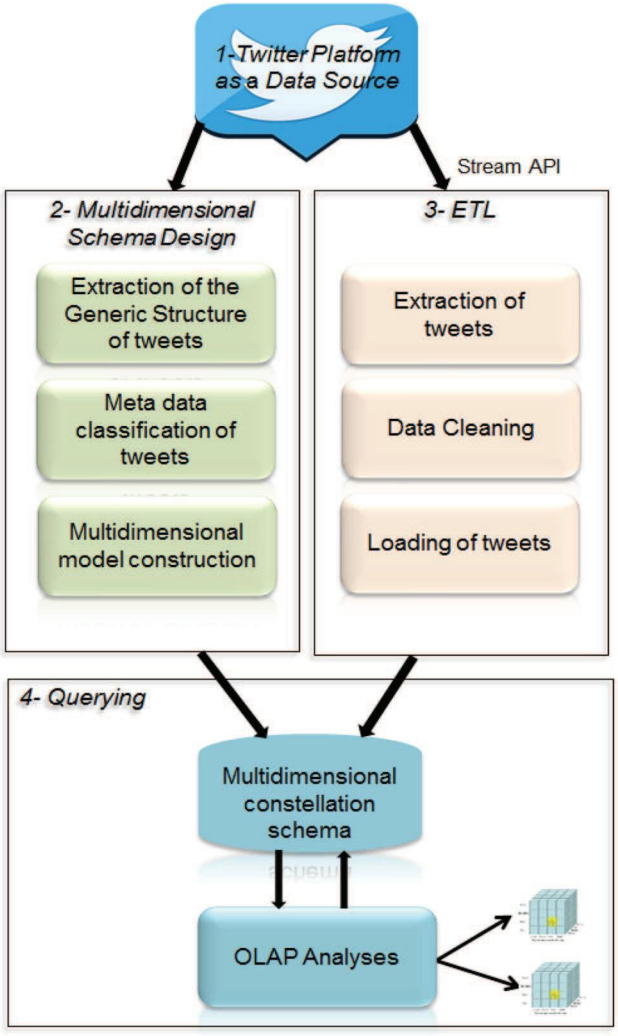


Fig. 4 Architecture of the *TweetOLAP* prototype

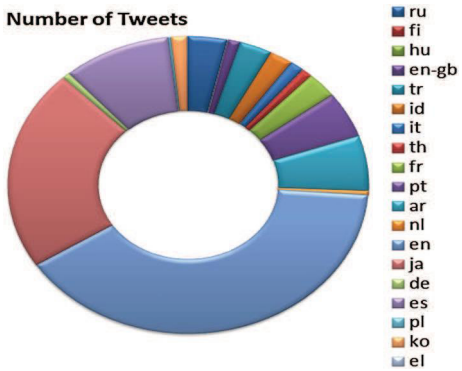


Fig. 5 Distribution of tweets per language

Table 3 Description of the dataset

Table (dimensions and facts)	Number of instances
DUser	70298
DPlace	1010
DTime	72000
DSource	3063
FActivity_Tweet	72000
FActivity_Twittos	72000
FActivity_Tweet_H	27268

Table 4 The growth rate of users' accounts per language and year (%)

Language	Year					
	2008	2009	2010	2011	2012	2013
English	0.36	3.47	−0.25	3.41	1.94	−1.51
Japanese	0.040	0.54	1.34	−0.07	0.89	2.23
Spanish	0.02	0.33	0.82	0.55	1.10	0.23
Arabic	–	0.22	−0.19	0.19	0.84	0.52
French	0.005	0.065	0.07	0.13	0.42	0.08
Italian	0.004	0.022	0.03	0.08	0.14	0.03
Portuguese	0.004	0.216	0.16	0.38	-0.03	0.55
Russian	–	0.01	0.07	0.14	0.21	0.19
Turkish	–	0.05	0.06	0.04	0.3	0.18
…	…	…	…	…	…	…

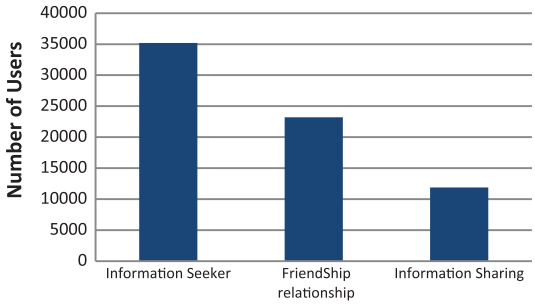


Fig. 6 Distribution of users per user category

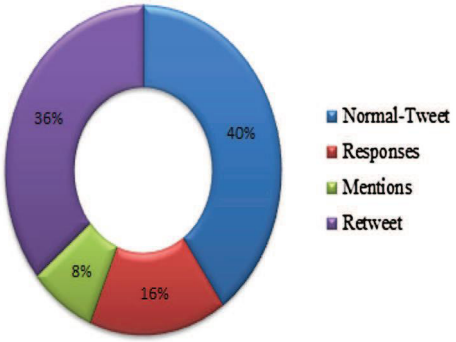


Fig. 7 Number of tweets per tweets-type

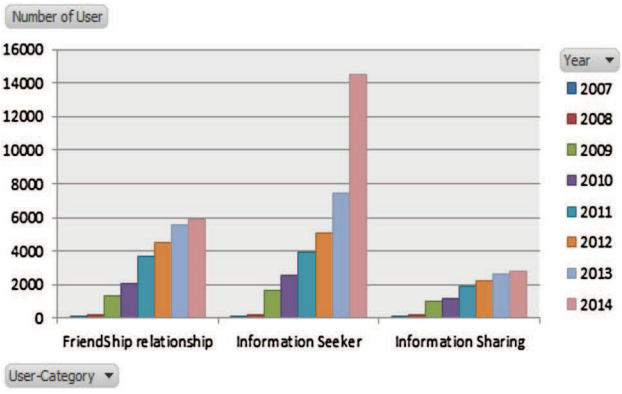


Fig. 8 Number of users per user category and year

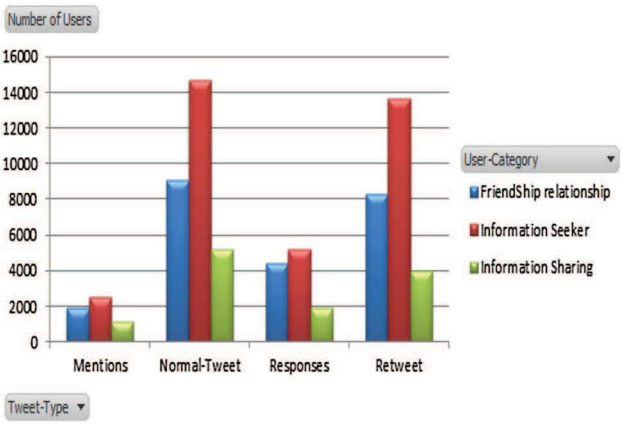


Fig. 9 Number of users per tweets-type and user category

We are now interested in the variation of the number of tweets produced by language and source. We notice that, since the appearance of the phone service provider (Twitter for iPhone, Twitter for Android,…), the number of tweets produced using Twitter for iPhone is more important using English language (Table 6).

We have conducted another study about the number of users by User Category (Dimension User Tweet) and Date (Year). These results are depicted in Fig. 8. These results lead to the following observations: The number of users' accounts is more important for the category *Information Seeker*. In fact, most of the users post rarely but follow other users regularly due to the valuable nature of their updates (*Information Sharing*).

The last query studied the variation of the number of users by Tweet-Type and User Category. Figure 9 plots the list of all User Category derived from the dataset and shows the distribution of each Type of tweet.

We have drawn the following conclusions from these experimental results. Actually, the more data volume is important the more the execution time increases. This is due to the fact that the software tool used for querying the logical multidimensional model is not initially designed to support huge data volumes. In order to alleviate this difficulty, we expect using recent technological tools dedicated to the Big Data paradigm. This will lead us to re-implement the logical model under the Hadoop platform using the Map and Reduce functions. This will speed up the processing, reduce the response time, and ensure scalability.

Table 5 The growth rate of users per source and year (%)

Source	Year					
	2008	2009	2010	2011	2012	2013
Web	1.19	2.55	4.12	−0.7	1.65	7.1
Twitter for iPhone	0.19	5.19	2.71	−1.32	2.78	4.14
Twitter for BlackBerry®	0.03	1.07	1.46	−0.35	1.86	−2.03
Twitter for Android	0.18	1.83	−0.22	4.59	4.72	−2.44
Mobile Web (M2)	0.007	0.09	0.17	0.29	0.15	−0.60
Twitter for iPad	0.029	0.35	0.1	0.29	0.27	−0.8
Facebook	0.003	0.32	0.16	0.03	0.04	0.3
…	…	…	…	…	…	…

Table 6 Number of tweets per language and source

Language	Source			
	Twitter for Android	Twitter for BlackBerry®	Twitter for iPhone	Web
English	5907	3149	12068	6665
Spanish	1520	2108	4869	789
Portuguese	543	20	335	3998
Arabic	378	544	411	187
French	143	86	213	246
Japanese	732	2	920	256

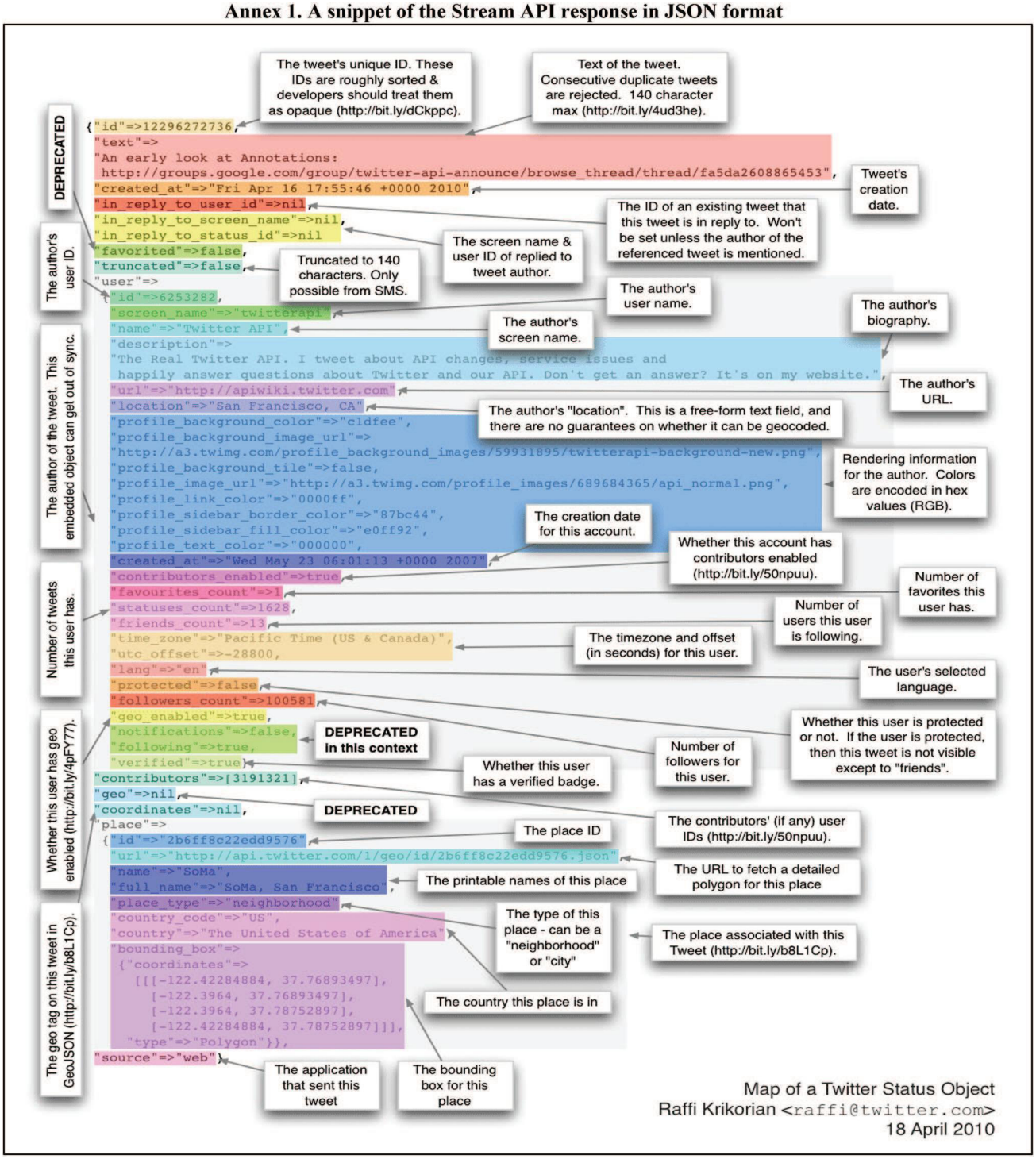
6 Conclusion

In this paper, we applied the data warehousing technology to enable comprehensive analysis of massive data volumes generated by the Twitter social network. We have proposed a multidimensional model dedicated to the OLAP of the data exchanged through tweets. This model mainly consists of a set of facts and dimensions. We have ensured that this model is generic, that is not limited to a set of pre-determined analytical requirements, which gives it a broad analytical potential and capacity to respond to ad hoc needs. Besides, we also took into account the specificities of data from tweets: links between tweets and tweets' answers. For that purpose, we have extended the concept of fact by the proposal of a new type of fact named *reflexive fact*. This type allows connecting an instance of the fact table to one or several instances of the same table. This relationship will guarantee that every tweet response added to the table corresponds to an existing tweet, and then analyses of linked tweets are possible. We also suggested the concept of *Empty Dimension*, different types of measures and hierarchies. Furthermore, in order to validate our proposals, we have developed a software prototype *TweetOLAP* and through extensive experimentation (*Mono-Dimension Analysis* and *Multi-Dimension Analysis)* we prove how the resulting data warehousing can be used for solving a variety of analytical tasks. We currently continue to perform other OLAP experiments on a larger number of tweets.

Several perspectives for this work are possible. It is important to note that social network data entries (e.g., user profile fields, message status, etc.) evolve over time and the occurring changes must be reflected in the corresponding analysis. For this reason, it would be interesting to define an approach enabling OLAP to keep up with volatile data using the concepts of slowly changing dimensions to enable analysis of both the recent state of data and any of its previous states. It would be also interesting to define new OLAP operators Ravat et al. (2008) that take into consideration the specificities of this new multidimensional model, as reflexive fact and dynamic data. These operators will allow facilitating the interpretation of the results of the multidimensional analyses on the tweets and their metadata. We also expect to exploit the "Text Mining" techniques in order to extract knowledge from tweets and strengthen more semantics in the generic model proposed in this work.

Appendix: A snippet of the stream API response in JSON format

Annex 1. A snippet of the Stream API response in JSON format



References

Bifet A, Holmes G, Pfahringer B, Gavaldà R (2011) Detecting sentiment change in Twitter streaming data. In: 2nd workshop on applications of pattern analysis, JMLR: workshop and conference proceedings 17, pp 5–11

Bouillot F, Poncelet P, Roche M (2012) How and why exploit tweet's location information? In: Proceedings of the AGILE'2012 international conference on geographic information science, Avignon, 24–27 Apr, ISBN: 978-90-816960-0-5

Bringay S, Béchet N, Bouillot F, Poncelet P, Roche M, Teisseire M (2011) Towards an on-line analysis of Tweets processing. In: 22nd international conference on database and expert systems applications, DEXA, Toulouse

Chaudhuri S, Dayal U (1997) Data warehousing and OLAP for decision support. In: DOOD, pp 33–34

Cuvelier E, Aufaure MA (2011) A buzz and e-reputation monitoring tool for twitter based on galois lattices. Concept Struct Discov Knowl 6828:91–103

Guille A, Favre C (2015) Event detection, tracking and visualization in twitter: a mention-anomaly-based approach. In: Social network analysis and mining no. SNAM-D-14-00102R1

Java A, Song X, Finin T, Tseng B (2007) Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD workshop on web mining and social network analysis, ACM, pp 56–65

Kalucki J (2010) Twitter streaming API. http://apiwiki.twitter.com/Streaming-API-Documentation

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 53(1):61

Kimball R (1996) The Data Warehouse Toolkit. Wiley, New York. ISBN 978-0-471-15337-5

Kraiem MB, Feki J, Khrouf K, Ravat F, Teste O (2014) OLAP of the Tweets: from modeling toward exploitation. In: 8th international conference on research challenges in information science (IEEE RCIS'2014), Marrakesh, pp 45–55, 28–30 May 2014, ISBN #978-1-4799-2393-9

Kumar S, Morstatter F, Marshall G, Liu H, Nambiar U (2012) Navigating information facets on twitter (NIF-T). In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, 12–16 Aug, Beijing

Lampos V, Lansdall-Welfare T, Araya R, Cristianini N (2013) Analysing mood patterns in the United Kingdom through Twitter content, arXiv preprint arXiv:1304.5507

Liu X, Tang K, Hancock J, Han J, Song M, Xu R, Pokorny B (2013) A text cube approach to human, social and cultural behavior in the twitter stream. In: Proceedings of the 6th international conference on social computing, behavioral-cultural modeling and prediction

Mansmann S, Rehman N, Weiler A, Scholl MH (2014) Discovering OLAP dimensions in semi-structured data. Inf Syst (2014 close proximity). http://dx.doi.org/10.1016/j.is.2013.09.002i

Martínez V, González VM (2013) Sentiment characterization of an urban environment via Twitter. In: Urzaiz G, Ochoa SF, Bravo J, Chen LL, Oliveira J (eds) Ubiquitous computing and ambient intelligence. Context-awareness and context-driven interaction, Springer, Berlin, pp 394–397

Mathioudakis M, Koudas N (2010) Twittermonitor: trend detection over the twitter stream. In Proceedings of international conference on management of data, SIGMOD 2010

Phelan O, McCarthy K, Smyth B (2009) Using twitter to recommend real-time topical news. In: Proceedings of the third ACM conference on recommender systems. pp 385–388

Quercia D, Kosinski M, Stillwell D, Crowcroft J (2011) Our twitter profiles, our selves: Predicting personality with twitter. In: IEEE international conference on social computing, pp 180–185

Ravat F, Teste O, Tournier R, Zurfluh G (2008) Algebraic and graphic languages for OLAP manipulations. Int J Data Warehous Min 4(1):17–46

Rehman N, Mansmann S, Weiler A, Scholl M.H (2012) Building a data warehouse for twitter stream exploration. In: ACM fifteenth international workshop on data warehousing and OLAP, DOLAP

Rehman N, Weiler A, Scholl MH (2013) OLAPing social media: the case of twitter. In: IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2013)

Sakaki T, Okazaki M, Matsuo Y (2013) Earthquake shakes twitter users: real-time event detection by social sensors. In: IEEE computer society, vol 25, Issue 4, Apr 2013

Twitter Team (2012) Twitter turns six [Online]. http://blog.twitter.com/2012/03/twitter-turns-six.html

Vassiliadis P (1999) A survey of logical models for OLAP databases. ACM SIGMOD Record 28(4):64–69

Vosecky J, Jiang D, Leung KWT, Ng W (2013). Dynamic multi-faceted topic discovery in twitter. In: Proceedings of the 22nd ACM international conference on information & knowledge management CIKM'13, pp 879–884

ApacheHadoop. http://hadoop.apache.org/