OLAPing Social Media: The case of Twitter

Nafees Ur Rehman, Andreas Weiler, Marc H. Scholl
University of Konstanz, Germany
Email: {nafees.rehman, andreas.weiler, marc.scholl}@uni-konstanz.de

Abstract—Social networks are platforms where millions of users interact frequently and share variety of digital content with each other. Users 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 to such data. We attempt to extend the established OLAP(On-line Analytical Processing) technology to allow multidimensional analysis of social media data by integrating text and opinion mining methods into the data warehousing system and by exploiting various knowledge discovery techniques to deal with semi-structured and unstructured data from social media.

The capabilities of OLAP are extended by semantic enrichment of the underlying dataset to discover new measures and dimensions for building data cubes and by supporting up-to-date analysis of the evolving as well as the historical social media data. The benefits of such an analysis platform are demonstrated by building a data warehouse for a social network of Twitter, dynamically enriching the underlying dataset and enabling multidimensional analysis.

I. INTRODUCTION

The increasing ubiquity, the humongous growth of web-based social networks and the frequent social activity result in generation of massive social network data. This data is typically semi-structured or unstructured, e.g., tweet, status update, message, comment, etc., and may contain opinions concerning brands, events, persons, or things, and it may also carry a sentiment value. Many analysts from research and business community are interested in gaining insights to such valued expressions for personal, academic and commercial applications.

These opinions – expressed as free-form text or other unstructured content – are turned into analyzable information by means of semantic analysis such as Sentiment Analysis, Entity Extraction, Keyword Extraction, Event Detection, and Topic Selection using variety of techniques like the ones mentioned in [1], [2], [3], [4], [5] and [6]. In order to enable meaningful insights to the data, the derived information has to be analyzed from useful perspectives with some of the latter being rather obvious, such as Time or User, and others being less obvious, such as characteristics derived from the original content using text and opinion mining. We leverage the capabilities of text mining to semantically enrich the underlying dataset to discover new measures and dimensions to offer more analytical aspects. 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. We present 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.

OLAP is central to business analytics across variety of data domains and supports decision making process in an efficient manner. However, this technology has its price, namely, that data elements and the relationships between them must be explicitly specified and the data must be structured into multidimensional data cubes. Data from social media in its original format, obviously, does not fulfill those constraints. To make OLAP compatible with the social media data, existing methods and techniques are coupled with novel ones as explained in the following sections.

This work is an extension of our previous work [7] where we adopted the extended Dimensional Fact Model (x-DFM)[8] to model dynamic, non-strict and fuzzy dimensional hierarchies in OLAP to discover classification hierarchies in the Twitter data. In this work, we focus on integrating extensive natural language processing capabilities in OLAP to perform multidimensional social media analysis.

A. Related Work

Data warehouses and OLAP are at the heart of decision support systems and have demonstrated competitive business advantages in a wide spectrum of application domains. Significant advances have been made to extend the data warehousing and OLAP technology from the relational and multidimensional databases to new emerging data in different application domains such as sequences[9], taxonomies[10], text[11], imprecise data[12], streams[13] and graphs[14].

Extreme popularity of online social networks (OSNs) enticed researchers from the industry and academia to analyze their growth, contents, impact, user behavior, etc. Twitter\(^1\) is one such network that has triggered a large number of research projects and application developments by allowing access to its public data streams through a set of APIs. We are particularly interested in Twitter because of its distinguishing feature of high-performance data streams. However, our findings and methods are easily applicable to other social networks.

There are many research studies and proposals to analyze Twitter data. In [5], the authors propose to analyze the content of the tweets (i.e., user messages posted on the Twitter platform) in real time to detect alarms during an earthquake.

\(^1\)http://twitter.com
The authors of TwitterMonitor[15] present a system to automatically extract trends in the stream of tweets. A similar work is presented in [16]. Recommendation systems for Twitter messages are presented by Chen et al. [17] and Phelan et al. [18]. Chen et al. studied content recommendation on Twitter to better direct user attention. Phelan et al. also considered RSS feeds as another source for information extraction to discover Twitter messages best matching the user’s needs. Michelson and Macskassy [19] discover main topics of interest of Twitter users from the entities mentioned in their tweets. Hecht et al. [20] analyze unstructured information in the user profile’s location field for location-based user categorization. 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 by using the data warehousing and OLAP technologies. There are very few studies interested in the use of OLAP to integrate textual data into data warehouses. For example, Bringay et al. [21] proposed methods from Information Retrieval(IR) that can potentially extract analytical concepts for a textual cube providing contextual aggregation in OLAP. However, they used a static medical taxonomy to relate words in a hierarchy. Zhao et al. proposed Graph Cube in their work[14] that takes into account both attribute aggregation and structure summarization of the networks. Graph Cube goes beyond the traditional data cube model involved solely with numeric value based group-bys, thus resulting in a more insightful and structure-enriched aggregate network. Liu et al. in [22] 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 at sentiment analysis and visualization. However, our approach exploits knowledge discovery – a set of text mining algorithms to extract named entities, events and other semantic knowledge – along with the associated sentiments and model them as distinct dimensions hence providing meaningful insights to the cube data. While other contributions focus on mining or enhancing the contents of tweets, improving the frontend or generating meaningful recommendations, we exploit the advantages of the established OLAP technology coupled with text and opinion mining to enable aggregation-centric analysis of the meta-data about the Twitter users and their messaging activity from the discovered perspectives.

II. TWITTER

Twitter is a popular social network with microblogging service for real-time information exchange. Twitter offers a set of APIs for retrieving the data about its users and their communication. Extreme popularity of the Twitter and the availability of its public stream have resulted in the multiplication of Twitter-related research initiatives as over viewed in the Related Work. Our analysis platform can accommodate data from other social networks with little effort but for this paper we only considered the dataset obtained from the public data stream from Twitter.

A. The Data Model for Twitter

To understand what type of knowledge can be discovered from this data it is important to investigate the underlying data model. In a nutshell, it encompasses users, their messages (tweets), and the relationships between and within those two classes. 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. The third component is the timeline, which describes the evolution, or the ordering, of user and tweet objects. Using the terminology of the Twitter Developer Documentation [23], the data model consists of three object classes: status objects, user objects and timelines.

Though not tailored towards OLAP, the offered data perspective can be adapted for multidimensional aggregation. One data record in the stream encompasses a single tweet event stored as the message itself (content and metadata) along with a detailed description of the authoring user’s profile in terms of various activity counters. The provided dataset 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). 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. The dataset delivered by the Twitter Streaming API is semi-structured using JavaScript Object Notation (JSON) as its output format. Each tweet is streamed as an object containing 67 data fields with high degree of heterogeneity. A tweet record encompasses the message itself along with detailed metadata on the users profile and geographic location. 10% of the total public stream provided by the Streaming API covers more than one million tweets per hour, which is a heavy load even for a high performance data warehouse system. Our solution to coping with such a massive data stream is to convert the streamed objects into XML and buffer them in a native XML database BaseX[24] developed within our working group. The following XML snippet shows an excerpt of a tweet object:

```xml
<tweet>
  <text>
    If you havent read about Mario Balotelli yet, you MUST before todays #EURO2012 final: http://t.co/2aFDjnsD
  </text>
  <truncated>true</truncated>
  <date>2012-01-07 18:36:05.000</date>
  <source>web</source>
  <retweeted>true</retweeted>
  <user>
    <name>Marcel***</name>
    <date>2011-08-01 06:06:34:12.000</date>
    <utc-offset>-18000</utc-offset>
    <language>en</language>
    <geo-enabled>False</geo-enabled>
    <statuses_count>1521</statuses_count>
    <followers_count>121</followers_count>
  </user>
  ...
</tweet>
```
Data in a data warehouse is structured according to the aggregation-centric multidimensional data model, which uses numeric measures as its analysis objects [25]. A fact consists of one or multiple measures along with their descriptive properties referred to as dimensions. Values in a dimension can be structured into a hierarchy of granularity levels to enable drill-down and roll-up operations.

We have identified three levels of granularity analyzing Twitter data, namely, a) User, b) Tweet and c) Tweet-content. The schema of the User cube has measures `Friendscount`, `FollowersCount`, `StatusCount` and `ListedCount`, which correspond to the analysis at User level. `retweetcount` and `favouriteCount` in the Tweet cube correspond to Tweet level, while `NoOfHashtags`, `NoOfEntities` and `NoOfEvents` correspond to Tweet-content level.

III. SOCIAL MEDIA DATASET ENRICHMENT

The social media dataset can be enriched and extended in many ways offering a whole new set of analytical aspects to business analysts. The shaded elements in Figure 2 represent the initial Data Warehouse schema, which is to be extended later to reflect the enrichments in the dataset. Following is the description of various data enrichment methods.

A. Straightforward Mapping

In the Twitter data objects, there are several data fields that can be directly assigned as fact/measures, dimensions or dimensional hierarchies. For example, `NoOfHashtags`, `LenOfTweet`, `RetweetCount` and `FavoriteCount` are perfect candidates for measures and are therefore modeled in the Tweet Cube as shown in Figure 2. Similarly, `FriendCount`, `FollowersCount`, `StatusCount` and `ListedCount` are derived from the Tweet object and modeled as measures in the User Cube as shown in Figure 2. In addition to the measures, there are candidates for dimensions, e.g., `User`, `Time`, `Location`, and `Source` are dimensions and are modeled as shown in the schema in Figure 2. Based on the data in these dimensions, multiple hierarchies are extracted as aggregation paths for drill-down and roll-up operations. One such aggregation hierarchy (`Source = ⊥SourceDeviceSource ≤ AppType ≤ App ≤ T_{tweet}`) is derived from the `Source` field in the original dataset. Source field contains information on the device from where a tweet is sent. The hierarchy presented in Figure 1 depicts various levels and instances of `source `device hierarchy.

B. Knowledge Discovery Based Derivation

Extracting meaning and value from unstructured text and integrating it within the OLAP context is the main focus of this paper. We used a set of text and opinion mining algorithms along with sentiment analysis to support exploratory and predictive analysis of the social media.

Modern data warehouse tools offer additional linguistic features to help analysts deal with the textual data. For example, the `TermExtractor` component in the Microsoft SQL Server Integration Services can be used to extract various terms from text fields. However, there are third-party tools and services that are more effective. Two of such services are AlchemyAPI[1] and OpenCalais[2]. They offer far more features than those available in the in-database components like Sentiment Analysis, Entity Extraction, Keyword Extraction, Topic Extraction, Concept Tagging, and Relation Extraction. These services offer various flavors of operations like internet-based & on-premise APIs. We used the internet APIs for semantic enrichment of our dataset. The User’s profile contains a `Description` field of 160 characters where the user fills in details to describe him/herself. We scanned this field using the above two APIs and modeled the knowledge extracted as `ProfileTopicDIM`, `ProfileTagDIM` and `ProfileSentimentDIM` dimensions for user profiling. A similar procedure is run for the Tweet-Text field, whereas `TweetTopicDIM`, `TweetTagDIM` and `TweetSentimentDIM` dimensions maintain the corresponding knowledge as shown in Figure 2. This enables analysis from hidden and useful perspectives.

C. Data Mining

An important part of our work relates to the use of data mining to perform predictive analytics. Data mining models are developed and trained to see if a tweet is spam or a particular tweet will go viral or how many followers will a user gain or loose, or whether two users will become friends. The model that we discuss here is a tweet popularity classifier. It is a decision tree based classification and prediction model that we developed based on the dataset presented in Figure 3. Tweet are ranked based on the following scoring formula:

$$RankingScore = RetweetCount \times 80 + FavoriteCount \times 20$$

Tweets from the input dataset are categorized based on the rules given in Table I. The result classifier takes the form of rule-based decision tree to predict the `popularity` category of an unseen tweet. The hierarchical structure of the predictive classifier is a good candidate for rule-based aggregation hierarchy and is useful for descriptive or exploratory analysis as well. Figure 2 also shows a similar `UserPopularityModel`. Based on the similar concept, a variety of other ad-hoc data mining models can be developed and deployed.
IV. Dynamic Dimensions and Hierarchies

A dimension is a one-to-many characteristic of a fact/measure and can be of arbitrary complexity, from a single data field to a large collection of related attributes, from uniform granularity to a hierarchical structure with multiple alternative and parallel hierarchies. Dimensions are the aspects and hierarchies are the aggregation paths to the measures of a cube. In order to speed up response time to queries, data of a cube is precomputed and materialized. This works fine as long as the underlying dataset and the relationships in the hierarchies do not change. However, data do change in the real world and should be reflected in the subsequent analysis. This introduces a new challenge of dynamic or changing dimensions. In the following sections we address the issues arising from the change in the data.
A. Change Detection

The Extraction, Transformation and Loading (ETL) process is repeated each time new data is uploaded in the data warehouse. This data may be totally new and may contain a set of updated values for certain pre-existing attributes across the data warehouse. It is a business decision to choose attributes that must be tested for any change in the existing data. We can choose to discard the existing data by replacing it with the new data or we may choose to keep all – historic and current – states or versions of the data.

Based on the proposals of Slowly Changing Dimensions presented in [26], we are interested in tracking changes in the Name and ScreenName attributes as Type I change i.e., we are overwriting the existing data with the new data. We further want to keep history for the Popularity attribute, therefore, we treat this attribute as Type IV change. In order to perform efficient comparison of the new and existing data, we generated checksum values for the set of Name and ScreenName attribute as it will be more efficient to compare numbers rather than strings. Any change in either of the attributes will result in the overwriting with the new data in both attributes. Popularity of users can change over time triggering a cascading effect in our OLAP settings. Please recall that we treat the Popularity prediction classifier as an aggregation-hierarchy, therefore, any change in the membership must also take effect here.

B. Maintenance of Dynamic Dimensions

In the staging area of the data warehouse, new data is staged in a dimension as shown in Table III where three attributes IsNew, IsType1 and IsType4 are added to keep track of the type of change. In case of purely new data, a new record is appended into the User dimension as shown in Table IV. Tweet record with the UserKey 5463521 is purely new in this example, and is appended to the User Dimension as reflected in Table IV. The tweet record for UserKey 7567534 induced Type I change where new data replaces the old data. As we can see, the Popularity category of the record with UserKey 4325643 changed from popular to famous and this is a Type IV change resulting in storing a copy of old data in the User History dimension as shown in Table V while the current state is reflected in the User dimension shown in Table IV.

C. Dynamic Hierarchy

Dimension hierarchy is a central concept in OLAP as it specifies valid aggregation paths for exploring the facts in a data cube at different levels of detail and in a hierarchical fashion from a more abstract view of coarsely-grained aggregates to a more precise view obtained through a sequence of drill-down and slice & dice operations. In OLAP tools, dimensions are presented in the form of hierarchical data navigation structures [8]. Figure 3 depicts a dynamic hierarchy where the popularity of tweets change over a period of time causing update in the rolls-up-to relation of its members. Because the aggregation path is defined by set of rules it can also be implemented as a three-level hierarchy by combining all rules of aggregation paths to a corresponding level as shown in Table 4.

V. DEMONSTRATION

Twitter has become a reflection of all real-world events. Let it be the Arab uprising, any natural disaster, political elections, movie or song launch or sport events, it gets reciprocated into a huge social activity on Twitter. Twitter users talk and react to such events. We had a variety of events to choose from for demonstrating our work. We opted to consider The 2012 UEFA European Football Championship, commonly referred to as Euro 2012, for this experiment. The final match was played between Spain and Italy. Apart from setting record on Twitter, Euro 2012 set the record for both the highest aggregate attendance (1,440,896) and the highest average attendance per game (46,481) under the 16-team format since 1996.

A. Dataset

We consider the dataset obtained for the 2012 European Football Championship final played between Spain and Italy on July 1, 2012. This game sat a new sports related record on Twitter where 15,000 tweets per second (TPS) were sent across Twitter platform and a total of 16.5 millions tweets were sent during the course of the game. We have about half million tweets in our dataset collected for three hours starting from the beginning of the game. We buffered the streaming tweets in BaseX [27] and filtered relevant tweet based on the trending topics listed in table VI.

B. Semantic Enrichment

The original dataset consists of about 67 different data fields. Out of these fields, only two fields i.e., User Description and Tweet, are free-form text fields where users can fill in anything they like in a textual form. User Description has a maximum length of 160 characters. However, some users do not fill in anything or those who do seldom make any changes

\[\text{http://www.uefa.com/uefaeuro/index.html}\]
\[\text{http://www.euro2012.twitter.com}\]
to it. But a tweet must contain some content with a maximum length of 140 characters. It can also contain user names and URLs of external websites, photos and videos. These two fields are focus of semantic analysis as these contain valued user expressions and opinions. These fields are processed by AlchemyAPI[1] & OpenCalais[2] for semantic enrichment, the earlier is used for NER and sentiment analysis while the later is used for topic extraction and concept tagging for uniformity of results.

These APIs enforce a daily request rate-limit, therefore instead of one both APIs were used to enrich more tweets in less time. Text contents from User Description and Tweet were submitted for semantic enrichment. User Description is semantically analyzed only once for any user in our dataset as the field value does not change much. However each tweet is potentially a new user message and hence all tweets are semantically analyzed with the exceptions of re-tweeted tweets where the change is reflected in Retweet Count field.

Table VII shows the distribution of results for the Sentiment Analysis performed on a dataset of 428735 tweets relevant to the event under consideration. Entity Detection performed on this dataset returned results offer insights of the sports lovers who engaged in social interaction during the course of the game. The entity detection model[2] that we used identified entity types as many as 36 despite the fact that the context is restricted to only 140 characters. Figure 5 plots the top 10 entities detected for entity type Person and Country while figure 6 (a) shows all the detected Entity Types. Each tweet was scanned to associate it with a Topic from a set of supported topic set [2] to provide aggregation and enable insightful analysis. Figure 6 (b) plots the list of all topics derived from the dataset and shows the distribution of each topic discussed.

The occurrence of macro and micro events also gets reciprocated on social networks and potentially contains important information. Analysts can largely benefit from the set of semantic enrichment methods and can leverage the information extracted using Entity & Event Detection to offer more and
C. Semantic enrichment across social engagement

Social engagement represents the user’s activity directly triggered by a social action of another user. The Twitter terminology for social engagements includes Favorite, Retweet and Reply-To. A tweet may trigger none or any combination of these engagements. Figure 8(a) plots the sum of social engagement for Favorite-Count and Retweet-Count across team orientation of Twitter users. Figure 8(b) plots similar statistics with the addition of sentiments across each team. This graph shows tweets which received such social actions from Twitter users across the sentiment. A Retweeted message is shared directly with all followers of the user and therefore contributes to trending or popularity of the same message and it’s content. We employed Favorite-Count and Retweet-Count as measures in our OLAP cube along with other derived measures and modeled topic, entities, events etc as dimensions around it. Doing so, allows the discovery of local and global popular topics, personalities, things, events etc by exploring the cube along the given dimensions. Figure 8(a) is a small reflection of such an exploration depicting popularity of the teams. Tweets for which team support could not be derived are also represented in this graph. Figure 8(b) plots similar statistics along sentiments and enables analysts to see whether sentiments of the tweet contributed to popularity.

VI. CONCLUSION

This work proposes an analysis platform for the huge data generated from social activities online. It advocated the use of the mature data warehousing technology coupled with data mining to enable efficient and multidimensional analysis of social network data from the newly discovered perspectives. These discovered perspectives are derived out of the underlying dataset using a variety of conventional and knowledge discovery methods. The dataset is semantically enriched by extraction and identification of entities, events, language, sentiment, topics etc from the user messages. More-
over, DM methods are used to discover hidden characteristics and relationships among data. This discovered knowledge is translated into various data warehouse objects to enrich the underlying data model which in turn, allows analysis from these new perspectives. The work also demonstrated how the evolving characteristics of social network data are reflected in the system and the analysis considers not only the historic but the current state of data. A detailed demonstration of such an analysis is presented in this work using the data obtained from the publicly available Stream API of Twitter.

VII. ACKNOWLEDGMENT

We thank our colleague Dr. Svetlana Mansmann for her support. This work is partially supported by DFG Research Training Group GK-1042 "Explorative Analysis and Visualization of Large Information Spaces", Germany.

REFERENCES