Large-Scale System for Social Media Data Warehousing: The Case of Twitter-Related Drug Abuse Events Integration

Jenhani Ferdaous, High Institute of Management of Tunis, Tunisia

Mohamed Salah Gouider, High Institute of Management of Tunis, Tunisia

ABSTRACT

Social media data has become an integral part in the business data and should be integrated into the decisional process for better decision making based on information which reflects better the true situation of business in any field. However, social media data are unstructured and generated in very high frequency which exceeds the capacity of the data warehouse. In this work, the authors propose to extend the data warehousing process with a staging area which is a large-scale system implementing an information extraction process using Storm and Hadoop frameworks to better manage their volume and frequency. Concerning structured information extraction, mainly events, the authors combine a set of techniques from NLP, linguistic rules, and machine learning. Finally, they propose the adequate data warehouse conceptual model for events modeling and integration with enterprise data warehouse using an intermediate table called Bridge table. For application and experiments, they focus on drug abuse events extraction from Twitter data and their modeling into the event data warehouse.

KEYWORDS

Big Data, Data Warehouse, Drug Abuse, Events, Hadoop, Hybrid Architecture, Linguistic Rules, Machine Learning, NLP, Social Media, Storm

INTRODUCTION

During the last two decades, enterprise' information systems are invaded with new kinds of data generated by the frequent and ubiquitous use of social media and mobile devices. Hence, social media became a rich source of business information. Thus, their analysis and integration into the data warehouse and the decisional process becomes a real business requirement in order to uncover hidden relationships, new insights and knowledge which will improve decision making and change all the business value chain.

Nevertheless, social media data, are characterized with their volume, variety and velocity, which is challenging the role of existing data warehousing systems. In fact, the data warehouse design based on the entity relationship model cannot be used to represent highly unstructured user-generated content. Moreover, the data warehouse developed mainly for batch processing cannot support streaming data which require real time collection and analysis. In addition, social media data volumes are continuously

DOI: 10.4018/IJDWM.290890

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

growing and exceed the data warehouse storage capacity. These issues appeal for new computational techniques and architectures to succeed social media content integration into the enterprise data warehouse for more realistic analytical results which reflects better the situation of the company in the business environment.

Recently, big data technologies and software solutions are going to replace the data warehouse. However, this is impossible regarding the importance of the latter technology in multidimensional data analytics and decision making and the foundation of the majority of existing decision support systems. Moreover, big data technologies are still under development. They don't have solid foundations and standard as well as normalized terminology. So, they cannot be used to replace the data warehouse. It presents a risk for enterprises, organizations and companies which just start to work with the data warehouse and just familiarized with its concepts and analytical capability, but aiming to make exploitation of big data sources and gain their profit.

In this work, the researchers propose to extend the data warehouse architecture with big data technologies namely Hadoop and Storm in order to enable the traditional data warehouse to support social media data volume and velocity. They develop a staging area to extract manageable structured information. Moreover, authors propose a conceptual model to represent information extracted from social media data and a bridge table to connect the new schema to the operational data warehouse. Later, new information from social media could be combined with operational data during analysis for new insights discovery and more interesting patterns in order to improve decision making. In experiments authors focus on twitter data processing, modelling and integration.

This paper is organized as follows; section two presents the state of the art of big data warehousing. Section three describes the proposed data warehouse architecture for social media data integration. Proposed techniques and algorithms for structured information extraction from social media text is presented in section four. In section five, authors describe the large scale implementation of the proposed solution. Section six is reserved for the data warehouse schema and conceptual modelling solution for successful data warehouse building. Finally, in section seven, the conclusion and open perspectives.

STATE OF THE ART

The complex nature of social media data is challenging the role of existing data warehousing tools and algorithms to integrate them into the enterprise decisional process. On the other side, their integration is becoming more and more required. Indeed, the foundational architecture of the data warehouse including ETL tools, integration techniques and conceptual models should be adapted to cope with big data volume, variety and velocity as well as succeed their integration.

In the recent literature, many approaches used existing big data ecosystems and their commercial software solutions to warehouse big data, so they totally replaced the traditional data warehouse. For instance, Chen (Chen, 2010) who have built a scalable data warehousing system on top of Hadoop. Data is collected, load into HDFS and processed in Map Reduce engine. Fact and dimension tables are represented in a non-relational model such as text, row-based binary array and columnar binary array. For data querying, an SQL-like language is proposed based on views created on top of the data warehouse schema enabling any user even not familiar with SQL to manipulate data. They enable also users to perform ad-hoc analysis by accessing directly raw data. In an other contribution (Thusoo et al., 2010), authors used Scribe to collect log data from web servers and load it in Hadoop cluster to be processed in map reduce. Then, the processed data are warehoused in Hive. The latter has a powerful system allowing users to formulate graphically their query thanks to its HiPal language. In these proposals, the data warehouse created is a new variety, schema-free; fact and dimension tables are files and the traditional definition of the data warehouse is totally abandoned.

However, in the big data warehouse called also the next generation data warehouse the limitations of traditional data warehouse architecture should be scaled back but not completely gone away

(Krishnan, 2013). In fact, new architectural and computational techniques should be considered to extend the traditional data warehouse capacity in support of big data in general and social media data in particular at any or all the levels of the data warehousing process.

ETL for Big Data

Some contributions handled big data issues at the ETL level trying to adapt traditional ETL tools to support Big Data. For instance, authors in (Boussaid et al., 2008), who proposed a multi-agent system at the ETL level for complex data extraction, structuring and loading. Authors handled data structuring issue, but, no any solution to manage data volume and velocity. Moreover, multi-agent systems are expensive solutions which lack scalability and availability. The latter approach considered internet content as complex data. However, complexity varies, and the extraction of hidden semantic is the most tedious task.

Rehman and his colleagues (2012) focused on warehousing streams of Twitter data. They considered the user and the tweet content itself main components, so facts. Moreover, these entities are enriched with semantic features extracted using NLP tools for entity detection, sentiment analysis, topic extraction, concept tagging and relation extraction. All the extracted details are represented as facts, dimensions, attributes and measures of the data warehouse. However, despite the successful results, the work does not handle the continuously growing volumes of data which exceeds the capacity of the data warehouse storage. Moreover, authors did not cope with the informal and noisy character of twitter data. So, how to integrate erroneous, informal and noisy data with enterprise business data?

Recently, map reduce paradigm is employed to abstract out the complexity of big data. For instance (Liu et al., 2011), where authors applied the ETLMR of web data in map reduce functions. Unstructured information about web pages is captured and stored into data file system then partitioned across mappers and reducers to build dimensions and facts. The proposed solution showed better performance and support data load in dimensional structure which is complex and verbose with Hive and Pig. However, the same research team presented in subsequent work CloudETL to parallelize dimensional ETL into Hive using Hadoop (Liu et al., 2013).

Many other approaches also used the powerful map reduce paradigm to automate ETL such as (Qu and Zhang, 2012) in which authors used Hadoop map reduce implementation for massive WAMS-Wide Area Measurement System log data processing. All the work of data cleansing and integration is performed in map reduce framework and data is read from HDFS.

The same idea developed in (Kumar et al., 2014) who used map reduce to extract, transform and load into the data warehouse log data files generated by application servers. Their approach ameliorated log data processing and achieved important performance according to parameters such as computation ratio, Network bandwidth and data locality factors.

In addition to the contribution of Das and Mohapatro (Das and Mohapatro, 2014), in which authors used Hadoop to automate ETL of unstructured external data. (Gupta and Rathou, 2013) who used text tagging and annotation techniques to extract entities from unstructured data such as scanned documents and emails based on domain ontology.

However, in recent map reduce based approaches did not focus on social media data, their semantic analysis for structured information extraction or topic identification and did not handle their streaming character.

Conceptual Modeling of Big Data Warehouse

Other contributions tried to generate adequate conceptual and multidimensional representations for the raw data extracted from internet and social media sources. Thus, manage big data at the conceptual level. For instance, in (Moalla et al., 2016), authors defined a set of rules for data warehouse schema design from extracted social media content. In fact, each object in Facebook and Twitter like posts, pictures, videos, etc. which has an important number of likes or shares will be considered as a Fact. Its measures are among others, number of shares, number of likes, number of followers. In (Yangui

et al., 2015; Yangui et al., 2017), and based on that a multidimensional model for a social data warehouse should be generated dynamically from the changing and evolving social networks structure, authors proposed DW4SN system. The system is composed of five modules which starts from a classical data warehouse schema and discover multidimensional concepts using a semi supervised clustering approach. Dynamic schema generation is an efficient method for continuously changing data modelling. But, analysing data over time will cause problems since some data does not exist at a given time.

Recently, the paradigm of information warehouse was introduced (Moulai and Drias, 2018) which focused on semantic extraction and modelling. It is the structure which stores data having meaning and significance such as text, image, video, etc., generated especially from social media data. The multidimensional model is composed of facts which are information, and dimensions are details representing the information. For experiments, they collected a set of tweets and applied association rules algorithm to extract relationships between terms and discover the most spoken topic.

Despite the powerful basic of their approaches to find a multidimensional structure which fits social media data particular type, they did not focus on the semantic of analysed text to the profit of a specific application domain. Moreover, they did not tackle volume and velocity problems of social media data. This latter challenge is handled by authors in (Santoso and Yulia, 2017). Their data warehouse system is based on Hadoop for better management of big data such as free text, RSS feeds, meta data, etc.

Social media data complexity in general and twitter data for instance involves their streaming character, continuously growing volume, noisy and informal content which are main issues face to their integration into the enterprise data warehouse to be combined with business data for more interesting insights. However, whatever the level in state of the art approaches, no any work tackled all these issues at once. Moreover, for more meaningful analysis, social data analysis and management should be highly domain-related.

Large Scale Data Warehousing Architecture

Social media data integration into the data warehouse is a big challenge regarding their unstructured character, volume and velocity. Indeed, social media data are generated in very high frequency which explains its continuously growing volume. Moreover, it is informal, free form as it is user-generated content. It is plenty with abbreviations, misspellings and slang terms. The extraction of meaningful and structured information to be injected into the data warehouse is a tedious task.

In this work, the authors aim to solve these issues and achieve our main objective. They cope with social media data particularities at all the levels of data warehousing process from extraction to load. At the ETL level, they extend the data warehouse architecture with a large scale system in which they handle volume, velocity and unstructured character. The researchers implement in a parallel and distributed framework a set of computational algorithms for structured information extraction from unstructured, streaming and voluminous twitter streams, to be easily integrated with business data.

At the conceptual level, the authors propose a multidimensional model which fits better the content extracted and connect the new structure to the enterprise data warehouse with a bridge table for periodic information loading. In the figure below an overview of the proposed system.

STRUCTURED INFORMATION EXTRACTION

In this work, the researchers handle three main issues; volume, velocity and unstructured data character. This section is reserved to structured information extraction for their smooth integration later. In fact, structured information has many forms; named entities, relations and events. The latter is the most complete form of information involving the named entities and semantic relationships among them in addition to circumstances information like time and place.

Figure 1. Large scale data warehousing architecture

Event Data Model

An event is defined in (Poibeau et al., 2013) as who did what to whom, when, where, through what method and why. It involves extraction of several entities and relationships between them. They are called the four basic 'Ws' (What, Who, Where, When). Based on this definition, the authors propose a formal description in order to facilitate further conceptual modeling. The following is a formal representation of the event:

$E = \langle S, L, D, A, I \rangle$ where:

S: is the semantic dimension of the event. It is the 'what' evoked in the definition above and designed with a trigger which is the word expressing the semantic of an event.

L: is the spatial dimension. The location where the event took place.

D: is the temporal dimension. The date when the event took place.

A: is the agentive dimension. The set of persons 'who' participated to an event.

I: the Instrumental dimension. Each event has objects by which it is accomplished answering the How question. For example in the drug abuse epidemiology, the instrument is the drug (DRUG). In a violence event or a crime, the instrument is the weapon used.

This specification is mandatory for identification common and domain-dependent entities.

Named Entity Recognition

The authors propose a hybrid approach in which they combine NLP techniques, linguistic rules and machine learning algorithms to succeed events extraction.

First step is about social media text cleaning like Twitter which is a noisy text, plenty with special characters, duplicates, redundancy, URLs and URIs, meaningless punctuations, abbreviations, external links, etc. thus, the straightforward application of NLP tools on this data as they are generated, will lead to erroneous results. Therefore, first step toward efficient structured information extraction and an indispensable work is data cleaning to facilitate analysis and increase performance. The

cleaning module is charged of: removing repeated letters to reduce the word to its canonical form (mariiiijuaaaaa to marijuana); removing hashtags #,

@; removing URLs; removing undesirable punctuations and characters (?; !,; ; *; (; \$,]; [.&).

Cleaned text is moved to the NLP pipeline for semantically-rich named entities identification like names of persons, dates, places, etc. which are easy to extract straightforward from text using Stanford CoreNLP. However, this latter cannot extract domain-related entities like names of drugs. Therefore, authors extend its pipeline with an annotator based on dictionary lookup. The dictionary is fed with seed elements from domain gazetteers and ontologies, and the process of its update is automated later in the solution. However, social media text is plenty with abbreviations and colloquial words which cannot much the dictionary content. Therefore, the authors propose to use a fuzzy matching technique to overcome this issue and recognize all possible positive instances.

The researchers tested the developed solution on 1000 000 Tweets collected with Tweets crawler. The accuracy of named identification system is affected with the threshold distance value chosen for fuzzy matching. Thus, authors made many tests with little variation of the threshold value, then, calculate the number of positive entities, false positives and false negatives. Positive entities are positive objects marked as positive, negative entities are negative objects marked as negative, false positives are negative objects marked as positives and false negatives are positive objects marked as negative. The obtained results are depicted in the table below:

Table 1. Statistics Variation for different threshold distance values

	#P	#FP	#FN
0.89	1874	1012	136
0.9	1886	612	180
0.92	1783	336	223
0.94	1765	272	318

The increase of the threshold value decreases the number of false positives but the number of false negatives increases. Moreover, misspelled terms, new slang terms, abbreviations and n-gram tokens which are positive entities are lost. In addition, the number of positives decreases when the similarity measure is higher. The objective is not to be close to the gazetteer since natural language texts frequently ungrammatical and plenty with errors. Actually, the most important is a higher recall value and not higher precision, so, authors carried the experiments using 0.89. For the collected

Table 2. Statistics collected for each Class Type with the threshold distance value 0.89

Entity type	#P	#FP	#FN
DRUG	50838	12632	1360
PSY-EFF	1606	3801	169
PHY-EFF	2250	2594	154
MED-COND	5450	2639	157
UNIT	1036	435	10
ROI	3706	3738	185

corpus of data, and focusing on drug abuse named entities extraction namely drug, psychological effect, physical effect, medical condition, unit, route of intake.

Our solution is more efficient in detecting drugs, medical conditions and units with minimum false positives as well as false negatives and with precision, recall and f-measure, 71%, 96% and 82% respectively.

Events Extraction

For events extraction, the authors look for semantic relationships among entities. They develop a hybrid approach composed of two layers in which they combine rule-based approach with a learning technique in order to obtain efficient results. In fact, in the first layer they develop a rule-based technique for pattern matching. The solution consists on matching the set of output entities with predefined lexico-syntactic patterns. Tested on 1000 000 tweets the solution gave 53% precision, 74% recall and 51% F-measure.

The performance of the rule-based approach individually performs poorly since it generates a large number of false positives and false negatives. In fact, due to the informality of social text, the authors define the majority if not all the arguments of linguistic rules as optional in order to capture all possible instances of the defined event which bias the results. So, many relevant events are missed, and invalid events are tagged which affects badly the performance of the system, bias the results and leads to an increasing number of false positives and false negatives. Therefore, they propose to build a classification model capable to extract meaningful events and distinguish real positive instances from false positives in order to increase the precision of the system. Writers use rule-based module powerful output such as its automatically annotated data set and its set of features with a classification algorithm to give accurate results and improve events extraction results. A classification process requires an efficient set of classification features and a sufficient annotated corpus for training.

The features used are discriminative which can distinguish effectively and accurately different events. The majority of features used in the learning process are inspired from ODIN specification. In addition to other features added to face social data particularity like shortness. The problem of negations is also treated. Indeed, the linguistic patterns formulated can capture explicit negations in sentences like “I don’t use any substance”. However, the sentence “like I am under alcohol” is certainly matched to drug abuse patterns (DA tag). However, it is filtered by the classifier thanks to the word window feature type which is very efficient in improving the event detection results. The authors propose to use the word before subject (WBS) like in the above example and the word after subject or before trigger (WBT). They trained four classification algorithms namely J48 for Decision Tree, Logistic Regression, Libsvm for Support Vector Machine and Naive Bayes implemented in Weka. The collected data is divided into 60% for training and 40% for test.

Table 3. Precision, Recall, F-measure of different classification techniques for Drug Abuse Event

	Drug Abuse Event		
	P	R	F
Decision Tree	0.93	0.93	0.93
Logistic Regression	0.91	0.9	0.9
SVM (libsvm)	0.93	0.93	0.93
Naive Bayes	0.84	0.83	0.83

Figure 2. Storm framework for streaming social text clustering

LARGE SCALE IMPLEMENTATION

Social media data is a variety of big data characterized with their volume and velocity. In order to succeed the information extraction from continuously growing volumes of data, the authors propose a large scale system implementation based on big data frameworks.

Storm for Data Collection and Preprocessing

The extraction of coherent information from Twitter is challenged by tweets shortness. In fact, tweets are limited to 140 characters before 2017 then increased to 280 characters. But, it is still not enough to express in one message. So, semantically-related details about given information could be scattered within many text segments. This makes the individual analysis of tweet segments meaningless and could lead to erroneous results. For example, to gather a complete a coherent event, the object (what?), the place (where?), the participants (who?), the time (when?), etc. are necessary information. These details may be not collected from a unique tweet, and dispersed within many tweets posted at different time instants. Moreover, social media data in general and twitter text in particular are generated in a streaming way.

In order to solve above issues, the authors propose to use a data stream processing framework called Storm1 for data collection and pre-processing. Storm is the leader of real time processing of large volumes of streaming data. It is a distributed and parallel real-time big data processing system developed to process vast amount of data in a fault-tolerant and horizontal scalable method. It is capable to process tens of thousands messages per second on cluster and guarantees that all messages will be processed tuple by tuple through the topology at least one time. Storm is connected to Twitter streaming API and for streaming data collection. The proposed Storm cluster architecture is presented in the figure below:

In fact, the authors implement at the heart of Storm an hierarchical bottom-up agglomerative approach in which they assume each tweet as a cluster, then gather all semantically-related information belonging to the same user in the same cluster using an exact matching, until reaching a top concept from which it is possible to extract a structured representation of an event. Outliers are also group of very low number of tweets belonging to non-active users, so eliminated. Final clusters are promoted to the rest of information extraction pipeline where each cluster is a potential event. Hence, tweets' clustering has two main objectives; gather semantically-related information which will lead to meaningful analysis. Second, reject tweets which don't bring any meaning, so their analysis will

congest the system with unnecessary and useless data which will slow down processing time and limits storage capacity. Clustering technique as a preprocessing work has also other advantages on the global system performance mainly reduce the volume of data to process thanks to outlier removal, reduce the processing time and memory space, improve system accuracy by extracting relevant information.

Hadoop Cluster for Information Extraction

In order to meet data generation and collection frequency, and manage the huge volume of data as well as abstract the complexity of the IE task, the authors propose to implement the task using Hadoop framework. Hadoop is designed to process large volumes of data easily thanks to its distributed storage system in files and parallel processing in map reduce functions across commodity of servers. they develop two loosely coupled Hadoop clusters; named entity recognition cluster (NER-Cluster hereafter) and events extraction cluster (EE-Cluster hereafter) where the output of the first is the input of the next as shown in the figure below:

Figure 3. Hadoop-based Information Extraction Framework

We propose two clusters for two reasons: the first is to reduce the complexity of the task and develop an optimized solution. The second is to offer independent, portable and reusable modules. Thus, each module can be used separately in any other system and gives same result as if both clusters work together. The input of the system is a set of clustered tweets stored in HDFS structure where each cluster represents a separate data file and one or more potential events.

The NER Hadoop cluster implements the cleaning and named entity recognition tasks. The output is a set of named entities promoted to the EE Hadoop cluster. The latter implements the events extraction hybrid approach. The output of the loosely coupled clusters is a set of events.

SOCIAL MEDIA DATA WAREHOUSE

In this section, authors handle social media data modelling and representation in the data warehouse, in particular twitter text. The Social Media Data Warehouse (SMDW) is a structure composed of the Event Data Warehouse (EvDW) representing events extracted from social media content namely the tweet text and the Twitter Data Warehouse (TwDW) for Twitter user information representation like user profile information, followers information, tweet details, etc. All the structures are connected to each others with a bridge table.

Structured information extracted is injected into the event data warehouse. The subject of analysis in this structure is the event. The EvDW conceptual model represents an event as a data structure that could be supported and managed by information technology tools and manipulated by users. The TwDW structure represents information about the social media data source, in our case Twitter, all information about the user and its network.

For the conceptual representation, authors used UML notation and toolkit for the design of the final model. they propose a model which generic classes represent any social media event, and could be specialized to represent domain-specific objects. In this work, researchers focus on drug abuse, so the specialized classes are entities belonging to drug abuse field. Below is the corresponding conceptual class diagram designed with StarUML toolkit:

Figure 4. Class Diagram of Drug Abuse Events

The main entities obtained for the conceptual model are the following:

- Event {id, category, type, date}: the event is the centric entity.
- Participant {id, name}: each event has some participants. The ID corresponds to the User_Id and the name corresponds to the Screen_Name
- Drug_Abuse {id, dose}: this entity models the drug abuse event. It has an identifier (id) and the dose of the drug taken.
- Adverse_Drug_Reaction {id}
- Instrument {id, name}: the instrument is a generic entity, in the case of the chosen application case study, the instrument is the drug.
- Drug {id, name, class}: each drug will have an identifier (id), have a name and belongs to a class.
- Medical Condition {id, name, type, category}: each medical condition has an identifier and a name,
- Physical_Effect {id, name, type, category}: each physical effect has an identifier and a name,
- Psychoactive_Effect {id, name, type, category}: each psychoactive effect has an identifier and a name,
- Route_Of_Intake {id, name, family}: a drug could have many routes of intake, each has a name and a family,

- Place {id, name, longitude, latitude}: an event occurs in a given place which have an identifier, a name, a longitude and a latitude
- Patient {id, gender, profession, marital status, date_of_birth}: an event could have many participants. The patient which is the abuser specified with its demographic information.

Multidimensional Model

Multidimensional modelling is the foundation of the data warehouse which offers best analysis capabilities like analysing data from different perspectives. As explained above, its main components are the fact and surrounding dimensions. To represent events extracted from social media, authors consider the event a fact of analysis and its participants as well as circumstances the dimensions according to which events will be analysed. In order to realize detailed and deep and multi-level analysis, the snowflake schema is the right choice. Hence, the social media data warehouse is defined with:

SMDW:(F, D{ }, HDi{ }) where:

F: the Fact table

D: {D1,..., Dn}: set of dimensions defined below,

HDi{ }: set of hierarchies for each dimension Di defined by HDi={h1,..., hk}

Event Fact

The event fact represents an event extracted from social media and considered the subject of analysis. It is defined by F: (NameF, PK{ }, M{ }, I) where:

NameF: is the name of the fact,

PK{ }: the primary key of the fact. It is the union of the primary keys of dimensions but represented as foreign keys in the fact table.

M{ } = (m1,...,mn) a set of measures. Numeric attributes scalar or aggregated such as counters (number of abusers, number of DA cases, number of ADR cases, etc.)

I= inclusion relationship between a fact and another. In fact, two events of the same type or of different types could be nested. This inclusion defined by $I(F_i) = F_j \subset F_i$

Dimensions

Events are facts to be analyzed according to existing dimensions such as drug, psychological effects, physical effects, patient, Time, Place, etc.

A dimension is defined by D: (NameD, A{ }, H{ }, TypeD) where:

NameD: name of dimension,

A{ }=(a1,..., al) is a set of attributes

H{ }=(h1,..., hz) is a set of hierarchies

TypeD $\subset$ [A, I, L, T]: a dimension could be agentive, instrumental, temporal or spatial dimension.

Each entity in the conceptual model developed above is transformed into dimension of the multidimensional model.

Measures

A measure M is defined by M:(NameM, TypeM, FuncM{ }) where:

NameM: name of the measure

TypeM: the type of the measure

FuncM{ }: set of aggregation functions compatible with summarization property of the measure where $\text{FuncM} \subset \{\text{SUM}, \text{AVG}, \text{MAX}, \text{MIN}\}$

Attributes

Each attribute A belongs to a give Dimension D and domain DOM. It is represented by $A_{ij} = (\text{NameA}, \text{DOM}, \text{VAL})$ where A_{ij} is the attribute i of the Dimension j and:

NameA_{ij}: name of the attribute,

DOM_{ij}:Domain of the attribute (String, Number, Boolean...)

VAL_{ij}: Boolean to indicate that the attribute is not always valued

Hierarchies

Figure 5. Drug Abuse EvDW Snowflake Schema

Each hierarchy H_i of the Dimension D_j is defined with $(NameH, P)$ where:

NameH_{ij}: name of the hierarchy

P_{ij}: {p_{i1}, ..., p_{iy}} set of hierarchy parameters

A good and efficient analysis is determined by the choice of measures. For the substance abuse and addiction case study, authors chose the following measures:

- NB_Drugs: the number of different drugs addicted or taken by the abuser,
- Sum_dose: the sum of doses taken by the abuser for each drug,
- NB_med_cond: number of effects
- NB_PSY_EFF: number of psychological effects by drug
- NB_PHY_EFF: number of physical effects by drug
- NB_ROI: number of Routes of Intake by drug
- NB_DA: number of drug abuses by user
- NB_ADR: number of drug effects by user
- AVG_age, MIN_age and MAX_age of abusers,

Twitter Data Warehouse Schema (TwDW)

While event data warehouse is conceived for the tweet content, to manage its semantic, the Twitter Data Warehouse structure is designed for user data and tweet metadata like user_id, user_screen_name, tweet_timestamp, tweet_longitude, tweet_latitude, pwd, url_profile_photo, etc. The following multidimensional star schema is the proposed model for twitter DW where the subject of analysis is the tweet itself.

User information should be also managed during business intelligence process. This information is necessary to discover the social background and social causes of a given phenomenon by managing the spatio-temporal dimension of the tweet, social interactions with other abusers, potential abusers, suppliers, etc.

Figure 6. Twitter Data Warehouse Star Schema

Social Media Data Warehouse Schema

The Twitter data warehouse contains all information about this social media source user. The events data warehouse transforms the content (tweet) posted, shared and liked into manageable entities like patient, drug, effect, etc. However, the enterprise data warehouse is the structure containing operational and transactional data. All structures are connected to each others with a bridge table having the following functions:

Data load from TwDW to EvDW and from the EvDW to EDW to create the hybrid content. As the final objective is to have the social media content integrated smoothly into the EDW without any changes in the existing data warehouse infrastructure and architecture, social content is periodically injected into the EDW.

Content integration between the three structures. In fact, a user in the TwDW corresponds to a participant in the EvDW such as a patient. To achieve the thesis goal, patient information are injected into the EDW once becomes mature enough for management and for use into the business activity. For example, a drug abuser detected from social media is considered a real patient and injected into the healthcare organization data warehouse for effective management by the latter.

Empty fields in the social media data warehouse structure which are domain related but cannot be extracted from the social text, like the types of drugs and effects, category of effects and family of drugs, are completed.

The bridge table is defined by:

BT:(NameBT, A{ }, O{ }, DateO) where:

NameBT: name of the bridge table,

A: {a1,..., az}: the set of attributes, in general its attributes are only the primary key which is composed of a set of Foreign keys which are the primary key of the event fact table and the primary keys of the other fact tables of the existing system architecture,

O: {o1, ..., on}: is the set of operations that could be operated between EDW and the SMDW,

DateO: each operation should have a date for future analysis purposes. This attribute could be represented also with a separate table with its possible hierarchies if necessary for future analysis purposes.

Events extracted from social content and loaded into the EvDW could be mined using data mining techniques and analysed as well as visualized using OLAP and reporting tools directly. However, the main objective of this work is to combine both transactional and social media data seeking for new

Figure 7. Global Data Warehouse Schema

insights and discoveries for better decision making.

Data Normalization

After data extraction, structured data should be transformed to meet the types of variables defined in the conceptual model.

- o Normalization of dates to the format defined during creation of the logical model of the data warehouse.
- o Normalization of names of drugs, adverse reaction, routes of intake and units expressed using slang terms or abbreviations to the real medical concept name. For example « Coca » or « Coke » should be « COCAINE », « ingest » should be « ORAL », « mg » should be « milligram », etc.
- o Doses which are not numeric values but expressed with letters « One », « Half », etc. should be transformed into digits as this item is defined as a numerical attribute.
- o Information about the abuser corresponds to the tweet user personal information obtained from the TwDW.

Concerning Twitter Data Warehouse feeding algorithm, authors refer to clusters metadata. In fact, each cluster is represented with a file managed by the HDFS and has a metadata. User and Twitter information are stored as metadata. The following is an excerpt of the TwDW feeding algorithm:

Algorithm: TwDW Feeding

Input: TwDW: Schema, C: Meta Data Clusters
DIM_USER.USER_ID = C.UID
DIM_USER.User_screen_name=C.SN
DIM_USER.URL_photo=C.url_profile_photo
DIM_USER.email=C.email
DIM_USER.address=C.address
DIM_USER.date=C.date

CONCLUSION

The integration of social media data in the business intelligence process is a real challenge regarding their complexity and particularity. This challenge appeals for new technologies to adapt data warehouse to support this new wave of data. In fact, the data warehouse is a mature business intelligence technology supported by a wide range of studies, fundamentals and methodologies that cannot be easily abandoned. Authors chose to continue to use it since it is still the single version of truth and the trusted basis of an accurate decisional system. Therefore, they propose to handle all the complexity in a staging phase to data warehousing process. They implement a set of novel methods in a large scale framework for structured information extraction from streaming social text namely events. They propose also a tailored multidimensional model for their conceptual representation. In the future work, authors will focus on OLAP for social content exploratory analysis and hybrid data analysis to show the advantage of using external events in the decisional process, together with operational and transactional data and the impact of social perspectives on final decisions.

REFERENCES

- Boussaid, O., Bentayeb, F., & Darmont, J. (2008). An MAS-based ETL Approach for complex data. ERIC/BDD, Universit'e Lumi'ere Lyon 2.
- Chary, M., Genes, N., McKenzie, A., & Manini, A. F. (2013). Leveraging social networks for toxicovigilance. *Journal of Medical Toxicology; Official Journal of the American College of Medical Toxicology*, 9(2), 184–191. doi:10.1007/s13181-013-0299-6 PMID:23619711
- Chen, S. (2010). Cheetah: A high performance, custom data warehouse on top of map reduce. *Proceedings of VLDB Endowment*, 3(2).
- Das, T. K., & Mohapatro, A. (2014). A study on big data integration data warehouse. *International Journal of Computer Trends and Technology*, 9(4), 188–192. doi:10.14445/22312803/IJCTT-V9P137
- Gupta, V., & Rathore, N. (2013). Deriving Business Intelligence from Unstructured Data. *International Journal of Information and Computing Technology*, 3(9), 971-976.
- Krishnan, K. (2013). Data warehousing in the age of big data. Newnes.
- Kumar, S., Athigopal, M., & Vetrivel, S. (2014). Extract, Transform and Load Strategy for unstructured data into data warehouse using map reduce paradigm and big data analytics. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(12). doi:10.1007/978-3-642-23544-3_8
- Liu, X., Thomsen, C., & Bach Pedersen, T. (2011). ETLMR: a highly scalable dimensional ETL framework based on map reduce. *Proceedings of 13th International Conference on Data Warehousing and Knowledge*, 96-111. doi:10.1007/978-3-642-23544-3_8
- Liu, X., Thomsen, C., & Bach Pedersen, T. (2013). *CloudETL: Scalable dimensional ETL for Hive*. In International Conference on Data Warehousing and Knowledge, Toulouse, France. doi:10.1007/978-3-642-23544-3_8
- Moalla, I., Nabli, A., Bouzguenda, L., & Hammami, M. (2016, November). Data warehouse design from social media for opinion analysis: The case of Facebook and Twitter. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)* (pp. 1-8). IEEE.
- Moulai & Drias. (2018). From Data Warehouse to Information Warehouse: Application to Social Media. In *LOPAL'18: International Conference on Learning and Optimization Algorithms: Theory and Applications*. ACM. doi:10.1145/3230905.3230914
- Piskorski, Yangarber, Manning, Mihai, Bauer, Finkel, Bethard, & McClosky. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. Academic Press.
- Poibeau, T. (Eds.). (2013). Multi-source, Multilingual Information Extraction and Summarization 11. Theory and Applications of Natural Language Processing. DOI 10.1007/978-3-642-28569-1__2
- Qu, Z., & Zhang, S. (2012). The WAMS Power Data Processing Based on Hadoop. *IPCSIT*, 25. doi:10.1007/978-3-642-23544-3_8
- Rehman, N., Mansmann, S., Weiler, A. H., & Sholl, M. (2012). Building a Data Warehouse for Twitter Stream Exploration. *IEEE ACM International Conference on advances in Social networks analysis and mining (ASONAM)*. doi:10.1109/ASONAM.2012.230
- Santoso & Yulia. (2017). Data Warehouse with Big Data Technology for higher education. *Procedia Computer Science*, 124, 93–99.
- Thusoo, A., Borthakur, D., Murthy, R., Shao, Z., Anthony, S., Jain, N., Sarma, J., & Liu, H. (2010). *Data Warehousing and Analytics infrastructure at Facebook*. In SIGMOD'10, Indianapolis, IN.
- Valenzuela-Escarcega, M. A., Hahn-Powell, G., & Surdeanu, M. (2015). *Description of the Odin event extraction framework and rule language*. arXiv preprint arXiv:1509.07513.
- Valenzuela-Escárcega, M. A., Hahn-Powell, G., Surdeanu, M., & Hicks, T. (2015, July). A domain-independent rule-based framework for event extraction. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations* (pp. 127-132). Academic Press.

Yangui, R., Nabli, A., & Gargouri, F. (2015). Towards Data Warehouse Schema Design from Social Networks-Dynamic Discovery of Multidimensional Concepts. *ICEIS*, (1), 338-345. doi:10.5220/0005383903380345

Yangui, R., Nabli, A., & Gargouri, F. (2017). DW4SN: A Tool for Dynamic Data Warehouse Building from Social Network. *Research in Computing Science*, 134(1), 191–205. doi:10.13053/rcs-134-1-15

ENDNOTE

- ¹ <https://cran.r-project.org/web/packages/Storm/index.html>

Ferdaous Jenhani obtained her PhD in 2021 in the field of Business Computing. She is an active member in SMART Lab. Her fields of interests are big data, data science, data mining, NLP and data warehouse. Her research resulted on interesting papers published in international conferences. She obtained her M.Sc. degree, she worked as an ERP consultant in a Tunisian SSII (SIMAC Tunisie) for 2 years. Then, she joined Smart Lab in 2015 for scientific research. She worked as a teacher at the Higher Institute of Information and Communication Technologies belonging to the University of Carthage since 2016 until today.