On-Line Analytical Processing on Graphs Generated from Social Network Data

Lilia Hannachi*, Omar Boussaid**,
Nadjia Benblidia*, Fadila Bentayeb**

*LRDSI Laboratory, University of Blida, Algeria
hannachi.lilia@yahoo.fr, Benblidia@yahoo.com
**ERIC Laboratory, University of Lyon 2, France
Omar.Boussaid@univ-lyon2.fr, Fadila.Bentayeb@univ-lyon2.fr

Abstract. Social Network services have quickly become a powerful means by which people share real-time messages. Typically, social networks are modeled as large underlying graphs. Responding to this emerging trend, it becomes critically important to interactively view and analyze this massive amount of data from different perspectives and with multiple granularities. While On-line analytical processing (OLAP) is a powerful primitive for structured data analysis, it faces major challenges in manipulating this complex interconnecting data. In this paper, we suggest a new data warehousing model, namely *Social Graph Cube* to support OLAP technologies on multidimensional social networks. Based on the proposed model we represent data as heterogeneous information graphs for more comprehensive illustration than the traditional OLAP technology. Going beyond traditional OLAP operations, *Social Graph Cube* proposes a new method that combines data mining area and OLAP operators to navigate through dimension hierarchies. Experimental results show the effectiveness of *Social Graph Cube* for decision-making.

1 Introduction

Business Intelligence (BI) represents a set of technologies and systems that play a major role to delivers the right information extracted from large amounts of data for decision-making. One of the most important technologies in BI is on-line analytical processing which represents a very powerful and flexible tool to mine and analyze data deeply. Using operations such as roll-up and drill-down, the result of OLAP is visualized as a data cube which allows decision makers to analyze quickly and navigate through the data from different perspectives and with multiple granularities. The multidimensional model used in OLAP supports handling of user defined views of data.

Over the last few years, social network sites such as Twitter and Facebook have quickly become a rich source of real time information by which people share short microblogs including daily conversations, cultural trends and information news without any concern about writing style, which make them able to exchange information about their personal point of views and interests. This imposes new challenges in the social networks and the microblogging data streams analysis in order to identify interesting/significant information among the hundreds of thousands of data produced in the social network services that are continuously being generated over a period of time. The social networks are typically illustrated as a heterogeneous information networks in which there are several types of network features. They depict the real-world interactions between multiple kinds of objects by very powerful and meaningful representations. For instance, the Facebook network includes users as well as other entities like, posts, videos and photos besides various other types of relationships such as, user-post publishing relationships or post-post joining relationships.

Community extraction methodologies within the social networks are receiving an increasing interest from the researchers in various fields. These methods permit the determination of latent groups by clustering individuals who share same characteristics and properties (interests). Usually, the study of community extraction is designed to determine the list of communities contained in a huge social network depending mainly on topological characteristics associated with the network entities such as the list of relationships. However, in the social network case, the relationships between individuals mostly represent a straightforward communications which indicate that a direct connection has been created during the social interactions. While actually, there is abundant meaningful information between social entities. For instance, although an existence of a friendship between users, we cannot extract their shared valuable information. Therefore, the study of user-generated content selected from the social network services attracts much attention in recent years. We think that the relation between users, as defined in the classic methods such as Fortunato (2010), Donetti and Muñoz (2004) is not enough when we look for users groups related to the same interest in order to recommend them some relevant information. Thus, it is more and more important to analyze the generated data in the social network services by using OLAP technique in order to get a data visualization from different perspectives and according to multiple granularities.

Unfortunately, the standard OLAP techniques does not support this type of complex data arising in real-world situations since the traditional OLAP tools can handle a limited number of hierarchies that ensures correct aggregation by enforcing summarizability in all dimensional hierarchies, which is obviously too rigid for a number of applications. In the case of social network data, OLAP technology does not consider the different kinds of relationships among facts. It also faces great challenges for analyzing unstructured data such as the social user-generated content. Note that the concept of summarizability in the data warehouse area refers to the possibility of correctly computing aggregate values defined at a coarser level of detail taking into account existing values defined at finer level of detail (Rafanelli and Shoshani, 1990). Seamantics, it supports, is weakly expressed.

As a continuation of our previous work Hannachi et al. (2013), in this paper, we explore and extract the pertinent knowledge hidden in the social network services with several interesting tasks by proposing a new data warehouse model, namely *Social Graph Cube* to support OLAP technologies on this complex multidimensional data. *Social Graph Cube* permits the decision makers to interactively analyze and manage structured data which represents the list of multidimensional attributes associated with the social entities together with the topological structure and the unstructured user-generated content produced in the social network services according to several perspectives and with different granularities. As the heterogeneous information networks are omnipresent and demonstrate a crucial component of recent information infrastructure, we represent data in the proposed model as heterogeneous information graphs to capture much richer, significative and comprehensive illustration than traditional OLAP cube. The several perspectives such as the geographic, semantic (i.e. relevant words) and temporal axes determine the dimensions and the different types of the vertices in the *Social Graph Cube*, while the list of measures is used to: (1) illustrate the existence of relationships between the social entities, and (2) turn out to define the aggregated network. Moreover, *Social Graph Cube* breaks the boundaries created in the classic OLAP-style which are based on the simple multidimensional attributes combined with the relational data by suggesting new approach founded on the community extraction methodologies, in order to navigate through the hierarchies and to determine the aggregate networks. It involves the illustration of vertices combined with social entities in coarser levels by defining the list of their associated condensed vertices (i.e. the set of their clusters). Both topological and semantic relationships between vertices are used in the definition of clusters. The definition of the semantic relationships can be achieved by using the Open Directory Project (ODP) taxonomy as an external resource, which represents the largest, and the most widely distributed human-compiled taxonomy of web pages currently available.

The rest of this paper is organized as follows. Section 2 reviews related work. In Section 3, we present and give details of our proposed *Social Graph Cube*. In Section 4, we introduce the suggested approach to define the aggregated networks. Section 5 presents the experiments we performed to validate our approach. Finally we give a conclusion and discuss research perspectives in Section 6.

2 Related Work

Recently, many research works have been done to extend data warehouses and OLAP techniques toward new emerging data arising in real-world situations. In Chen et al. (2009), the authors suggested a new Graph-OLAP framework for graph generated by the social networks. The authors in Qu et al. (2011), proposed additional graph summarization operators for the Graph-OLAP framework. The authors in Zhao et al. (2011) recognize the requirement for OLAP on graphs databases and present a new data warehouse model, Graph Cube, by integrating properties of multi-dimensional networks with existing OLAP techniques. The Graph Cube model is especially a list of all potential aggregations of the implicit multi-dimensional network, by combining attribute aggregation with structure summarization of the networks. However, in the whole of the researches presented above, the analysis is based on the classic OLAP by using the linked set of tuples described via a graph model. They did not create a multidimensional model suitable for the text data contained in social networks and did not also exploit the information (i.e. unstructured data) transmitted in networks. The studies in Tian et al. (2008) and Zhang et al. (2010) present and formally determine two operations comparable to OLAP-style navigations for graph summarization. The first operation SNAP (Summarization by grouping Nodes on Attributes and Pairwise relationships) generates a summary graph. The second less restrictive operation, k-SNAP, control the resolutions of summaries by specifying the number k of node groupings. While these operations are encouraging graph summarization mechanisms, their usability for OLAP technology is questionable. Even that OLAP cube is based on the notion of data dimensions, SNAP and k-SNAP do not operate with this notion at all. In order to define the semantic dimensions, the authors in Zhang et al. (2009) presented a new model called Topic Cube to analyze multidimensional text database. In addition, the authors in Bringay et al. (2011), defined a multidimensional data model for Twitter data. Unfortunately, in these works, OLAP operations, such as roll-up, drill-down, and slice-and-dice, are achieved by using only the traditional OLAP techniques, without considering specific relationships between data tuples. The complex dimension hierarchies defined in our approach might include some irregularities involving summarizability challenges. Many proposed studies aim to solve summarizability problems, such as Pedersen et al. (1999) and Pedersen et al. (2001) where the authors presented a set of algorithms used to automatically transform dimension hierarchies. Moreover, as we aim to construct the user communities based on data mining clustering, we review some studies in this area. The traditional clustering methods, such as Fortunato (2010), propose to group data by using the arcs density in the graph. For instance, the hierarchical clustering techniques like Hastie et al. (2001), aim to identify vertices groups with high similarity. It can be divided into two classes: agglomerative algorithms Donetti and Muñoz (2004), Du et al. (2007) and divisive algorithms Johnson et al. (1993), Newman (2003). In divisive algorithms technique, we do not need to specify the clusters number in advance, like Agglomerative algorithms, but the disadvantage is that many partitions are recovered. In this case we cannot define the best division. In Newman and Girvan (2004), the authors propose a new divisive algorithm. This algorithm is based on the concept of edge betweenness centrality. It works on moderate size networks significantly. However, the need to recompute betweenness values in every step becomes computationally very expensive.

In fact, our proposed *Social Graph Cube* architecture extends data warehousing and OLAP technologies toward such new multidimensional social networks. It provides pertinent responses to OLAP-style multidimensional analysis on information-enhanced multidimensional social network. Aggregation and OLAP operations are performed along the geographic, semantic and temporal dimensions defined upon the social networking services.

3 Social Graph Cube

The suggested *Social Graph Cube* permits the decision makers to quickly examine and understand the features of the topological, structured and unstructured data characteristics produced in the social network services with a view to determine exceptions and meaningful information and to fully take advantage of all the interesting parts contained within the underlying networks. To supply *Social Graph Cube* with accurate, actionable and fast answers for analyst queries, two types of external resources are used in this component: the topological structure of social networks and the different semantic enrichment tools such as: the WordNet dictionary, and the Open Directory Project (ODP) taxonomy. Figure 1, displays an instance of social network data produced in the Twitter website. It is composed of a set of users interrelated with a follower relationship. There are nine vertices (identified with an ID_User) and thirteen edges in the underlying graph, as shown in Figure 1(a).

Figure 1(b) shows real-world tweets interchanged between this set of users. They are extracted by using available tools and techniques. The main tool in the most popular social networks is the API (Application Programming Interface). It permits users to retrieve data in different formats which is usually in an Extensible Markup Language (XML) or JavaScript Object Notation (JSON). As we can notice, the unstructured user-generated content included within these tweets is a very rich data set, where most likely users aim to put significant information about their current activities or opinions. However, as this text is generally noisy and unstructured data, a treatment step has become indispensable. Thus, we clean our dataset

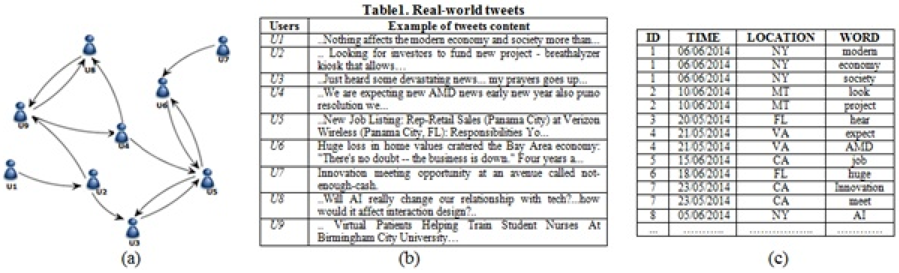

Table1. Real-world tweets

Users	Example of tweets content
U1	..Nothing affects the modern economy and society more than...
U2	.. Looking for investors to fund new project - breathalyzer kiosk that allows…
U3	..Just heard some devastating news... my prayers goes up...
U4	..We are expecting new AMD news early new year also puno resolution we...
U5	..New Job Listing: Rep-Retail Sales (Panama City) at Verizon Wireless (Panama City, FL): Responsibilities Yo...
U6	Huge loss in home values cratered the Bay Area economy: "There's no doubt -- the business is down." Four years a...
U7	Innovation meeting opportunity at an avenue called not-enough-cash.
U8	..Will AI really change our relationship with tech?...how would it affect interaction design?..
U9	.. Virtual Patients Helping Train Student Nurses At Birmingham City University…

ID	TIME	LOCATION	WORD
1	06/06/2014	NY	modern
1	06/06/2014	NY	economy
1	06/06/2014	NY	society
2	10/06/2014	MT	look
2	10/06/2014	MT	project
3	20/05/2014	FL	hear
4	21/05/2014	VA	expect
4	21/05/2014	VA	AMD
5	15/06/2014	CA	job
6	18/06/2014	FL	huge
7	23/05/2014	CA	Innovation
7	23/05/2014	CA	meet
8	05/06/2014	NY	AI
...	....	....	....

FIG. 1 – *Example of real-world tweets and a multidimensional social network*

by removing stop words, URL, noisy words, etc. Then, we convert the extracted data from its previous state into the desired state by using different semantic enrichment tools. To achieve the syntactic transformation of textual data, we utilize a linguistic knowledge. It is based on different techniques such as: stemming, spelling correction with WordNet dictionary, etc. Each tweet picked from the Twitter website is combined with supplementary data like user identifier, Time, Longitude, Latitude, etc. While the temporal specifications are captured explicitly in this metadata, the geographical specifications can be defined implicitly. They can be determined in different ways: First, by using the geographical coordinates (longitude and latitude). Second, manually filled by the user in his profile. Third, through the use of user's time zone.

We use the geographic database *Geonames*, to enhance the information about locations. This database is available for free under a Creative Commons Attribution license. The result of this treatment is represented in Figure 1(c). The structural and the semantic characteristics associated with this sample social network depicted in this figure, such as user ID (as primary key), time, location (in state) and word, are represented as a tuple in a vertex attribute table. The topological structure of the graph, together with the multidimensional attributes associated with vertex, forms a multidimensional network.

Definition 1. [***Social Graph Cube***] *Given a multidimensional network $N = (V, E, S, U)$, where V and E are the list of vertices and edges contained in the network, while S is the structural data such as the geographic and the temporal axes, U illustrates the user-generated content which represents the semantic axe, it is determined by selecting the list of pertinent words. The Social Graph Cube is obtained by reorganization of this multidimensional network in all possible cuboids produced by using the structured data S and the user-generated content U. For each cuboid C' obtained by S and U, the measure could be a homogeneous or heterogeneous weighted graph $G' = (V', E', W_{V'}, W_{E'})$ w.r.t. C'. The V' in the G' is either a simple set of vertices as defined in the initial multidimensional network or a set of condensed vertices. The E' represents the set of edges illustrated in the graph G', while $W_{V'}$, $W_{E'}$ are the list of weights associated with each edge and each vertex in the weighted graph, respectively. They are determined by using the semantic or the topological characteristics or the both.*

3.1 The Social Graph Cube Lattice

The lattice structure has performed a significant role in numerous views of data cubes since it can assist to enhance performance of data cube computations. Figure 2 describes a *Social*

Graph Cube lattice.

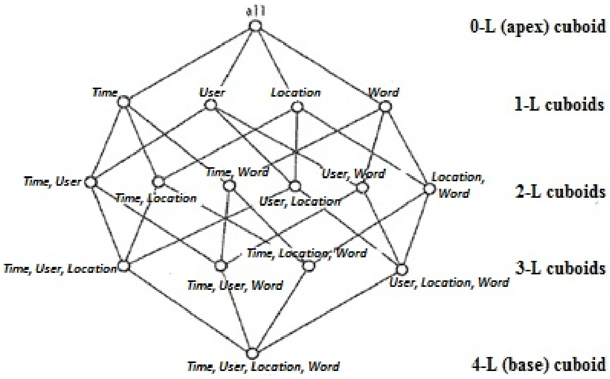

FIG. 2 – *The Social Graph Cube lattice*

Each node in this lattice is a cuboid of the *Social Graph Cube* produced from the initial multidimensional network by generating all the possible subsets of the particular dimensions. The edges in the lattice determine the parent-child relationship between two cuboids. Given a multidimensional network G with the dimensions time, user, location and word there are 2^4 cuboids in the *Social Graph Cube*. Each cuboid shows a different level of summarization. Unlike the conventional definition of the cube lattice structure, the cuboid in the *Social Graph Cube* that describes the highest degree of summarization is referred to the base cuboid. While the cuboid which describes the lowest degree of summarization is called the apex cuboid and it is characteristically symbolized by all. In *Social Graph Cube* a weighted graph matching to an ancestor cuboid is more homogeneous than the weighted graph matching to one of its descendant cuboids, which contains more topological properties and semantic details. By using the *Social Graph Cube* structure, end-users can examine the initial network in several multidimensional spaces by passing through the produced lattice. In this manner, a list of weighted graphs with different summarized resolution can be explored and analyzed for decision-making and business intelligence aims. In the following subsections, we will describe in details the list of cuboids (i.e., weighted graphs) produced in each level within the *Social Graph Cube* lattice.

3.1.1 The First Level in the Social Graph Cube Lattice

The cuboids generated in the first level *C'* with number of dimension $|dim(C')| = 1$, are represented with homogeneous weighted graphs, where in each weighted graph, we have only one type of topological entities. They can be computed to request some complex queries that could be asked on a multidimensional network such as:

- *What is the semantic network structure between the several users?*
- *What is the semantic relationship between the most mentioned words in the social user-generated content?*

As a result, we have four types of homogeneous weighted graphs; user-user graph, location-location graph, time-time graph and finally word-word graph. In the case of user-user graph, location-location graph and time-time graph, the first step is to aggregate all the social user-generated content sent by the same user, transmitted from the same location or produced at the same time interval, respectively. Then we clean this generated data by converting it from its previous state into the desired state using different semantic enrichment tools. The second step is to calculate the semantic distance between users, location, time intervals, respectively. The researchers in Rosen-Zvi et al. (2004) present the distance between individuals as the symmetric Kullback-Leibler divergence between the topics distribution conditioned on each of the individuals, as follows:

$$dist(i,j) = \sum_{t=1}^{T} [\theta_{it} \log \frac{\theta_{it}}{\theta_{jt}} + \theta_{jt} \log \frac{\theta_{jt}}{\theta_{it}}] \tag{1}$$

where i, j: represent $useri$ and $userj$. T: is the number of topics. θ_{it}, θ_{jt}: The probability of $topict$ according to $useri$ and $userj$, respectively.

Inspired by this study, we propose to calculate the semantic distance between users as the Kullback-Leibler divergence between the words distribution rather than the topics distribution. This distribution is computed by using the normalized TF-IDF measure, where values are taken in the range $[0, 1]$. To avoid the division by zero, we utilize +0.0001 standard deviations instead of zero for the TF-IDF weights. The distance between $user_s$ and $user_l$ is then represented by the following formula:

$$\begin{aligned} dist_{S_1}(user_s, user_l) &= \textstyle\sum_{k=0}^{K} TFIDF_s(w_k) log \frac{TFIDF_s(w_k)}{TFIDF_l(w_k)} + \\ &= TFIDF_l(w_k) log \frac{TFIDF_l(w_k)}{TFIDF_s(w_k)} \end{aligned} \tag{2}$$

where T represents the top-K most representative words characterizing each user, while $TF-IDF_s(w_k)$ and $TF-IDF_l(w_k)$ represent the TF-IDF weights associated with word w_k according to $user_s$ and $user_l$ respectively.

We can use this formula to compute the distance between locations and time intervals by replacing $user_s$ and $user_l$ with $location_s$ and $location_l$ or $time_s$ and $time_l$.

However, this expression compares only the weights associated with the same words without considering the semantic relatedness among the different words. As an instance, if we consider the following users with their top 3 most representative words according to the normalized TF-IDF measure.

- *User s : media 0.61, event 0.24, people 0.07*
- *User l : journal 0.56, company 0.37, media 0.02*

By using the previous formula, the distance between $user_l$ and $user_s$ is calculated depending only on the TF-IDF weights associated with word media. It does not take into account the semantic relationship between the different words such as media and journal or company and people which presents an important source of information. It could reveal that the list of words are linked semantically and often appear in the same fields. Based on this information, we can consider that these users are close semantically. From this idea, we propose to calculate the distance between users according to several words by combining the Kullback-Leibler divergence between the TF-IDF weights associated with different words distribution with another measure like the well-founded measure “Normalized Google Distance (NGD)" introduced by Cilibrasi and Vitanyi’s in Cilibrasi and Vitanyi (2007). This measure does not depend on a particular dictionary or corpus; contrariwise, it takes advantage of the vast knowledge available on the web where all possible interpretations for a word are considered. The NGD measure consists of calculating the distance between two words w_i, w_j as follows:

$$NGD(w_i, w_k) = \frac{max\{\log f(w_i), \log f(w_k)\} - \log f(w_i, w_k)}{\log P - min\{\log f(w_i), \log f(w_k)\}} \tag{3}$$

P: is the total number of web pages indexed by search engine; $f(w_i)$ and $f(w_k)$ are the number of hits for each words w_i and w_k, respectively; and $f(w_i, w_k)$ is the number of web pages on which both w_i and w_k occur. In our study, we generalize the NGD measure by using the open directory project as frequency source. The result of the combination between the proposed formula and the NGD measure is represented in the following expression:

$$\begin{aligned} dist_{S_2}(user_s, user_l) &= dist_{S_1}(user_s, user_l) + \frac{1}{2} \times \\ &\textstyle\sum_{k=0}^{K} \sum_{i \in \{K-k\}}^{K} TFIDF_s(w_k) log \frac{TFIDF_s(w_k)}{TFIDF_l(w_i)} \\ &+ TFIDF_l(w_i) log \frac{TFIDF_l(w_i)}{TFIDF_s(w_k)} + NGD(w_k, w_i) \end{aligned}$$

The third step is to compute the weight associated with each social entity (i.e. vertex) in the weighted graph. From the idea that depending on mentioned words in the social user-generated content, the users, locations or times candidates may be more or less important, different criteria are proposed to define the influence degree of each entity in the network. In the social network analysis area, users have a tendency to follow people that expected to be interesting. Based on this idea, the number of relationships between users such as following relationship is considered as the most important criteria to define user importance. The most popular reference to this measure is the closeness centrality Bavelas (1950), Beauchamp (1965), which is used to indicate the importance of a particular user. Considering this measure, the importance of a user s depends on the sum of its distances to all other users. It can be represented by the following measure:

$$Centrality(s) = \frac{1}{\sum_{s \neq l} distance(s, l)} \tag{4}$$

In our approach, we calculate the importance of users, locations or times in each weighted graph by adapting this measure. Unlike traditional closeness centrality measures where the distance between users is determined by using the length of shortest path between them, in our process, the distance between users is defined by using the semantic distance defined previously in Equation 4. The fourth step is to construct the semantic weighted graphs associated with each cuboid in the first level of *social graph cube* lattice. The main goal of these weighted graphs is to maximize the potential value around this effective visualization by representing a portion of the abundant meaningful information between social entities. By considering the list of words associated with the nine users presented in Figure 1, we present in Figure 3 the semantic weighted user-user graph. This graph enriches the modeling of the multidimensional networks with another type of relationships, which vividly describes the closeness of interests and views between the social users, locations or furthers the time intervals. This relationship is determined by using the semantic distance equation.

To construct this semantic weighted graph, we propose the following process: First, we define the list of vertices which represents the selected users. Second, we create an edge from the vertex s to the vertex l, if the semantic distance between these two $user_s$ and $user_l$ is less

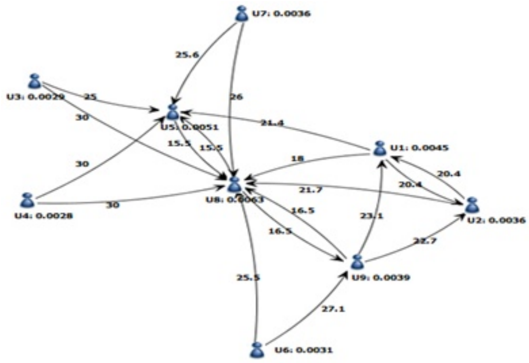

FIG. 3 – *An instance of the weighted user-user graph*

than or equal to a threshold parameter ε which can be defined by either the end-users or by the mean associated with user s, which is calculated by the following formula:

$$mean_s = \frac{\sum_{l=0}^{L} dist_{s2}(user_s, user_l)}{L} \tag{5}$$

where, L: represents the total number of selected users. The weight of this edge is calculated as the distance between these two users for the selected list of words. While, the weight of each vertex is computed as the closeness centrality described previously. In the case that we get a large number of semantic edges compared to the number of vertices, we repeat this phase. In the case of word-word graph, the first step is to define the most representative words selected from the social user-generated content by using the normalized TF-IDF measure. The strategy used to construct this type of graph is divided into three steps: First, compute the semantic distances between the selected words by using the NGD measure. Second, calculate the importance of words by modifying the closeness centrality measure. In the adapted measure, we integrate in the closeness centrality, the word weights computed by using the mean of all the normalized TF-IDF values associated with the selected word.

$$Centrality(word_i) = \frac{1}{\sum_{l \neq j} distance(i, j)} + Mean(TFIDF(word_i)) \tag{6}$$

The word weights reflect the importance of each word in the selected users or locations data. Third, evaluate the existence of edges between words. The guiding principle for creating an edge from the word vertex i to the word vertex j in this weighted graph is that, the semantic distance between these two vertices is less than or equal to a threshold parameter ε.

3.1.2 The Second Level in the Social Graph Cube Lattice

Consider the following query presented in Figure 1: *“What is the semantic weighted graph structure between the user U2 and the most representative words?"* The answer is a cuboid generated in the second level with number of dimension $|dim(C')| = 2$. As it accumulates two different multidimensional spaces of the graph, it is represented as a heterogeneous weighted graph with two types of topological entities. Further, in this level of lattice, we have six types of heterogeneous weighted graphs that leverage the rich semantic meaning of the social data structure. These weighted graphs are represented as follows: location-word graph, location-time graph, location-user graph, word-user graph, word-time graph, time-user graph. The process used to generate this type of graphs is: First, define the two types of dimensions involved in the request of the decision makers as topological entities. Second, aggregate and clean all the social generated content produced by these dimensions. Third, calculate the centrality of each entity as described in the previous subsection. Fourth, compute the semantic distance between all the topological entities presented in this graph. In the case of heterogeneous graphs produced through the semantic dimension *word* such as, word-user, word-time and word-location graph, the semantic distance is calculated by computing the average of the NGD measure between words and the top-K most representative words characterizing users or locations obtained from TF-IDF vector. As an example, in the following formula, we calculate the semantic distance between $user_s$ and $word_w$:

$$dist\_word(user_s, w) = \frac{\sum_{k=1}^{K} NGD(w_k, w)}{K} \tag{7}$$

where K is the number of top-K most representative words selected to represent users. In the other graphs, the semantic distance is computed as illustrated in Equation 4. Finally, keep the relevant relationships which describe the most closely entities. In Figure 4, we display the answer of the preceding query.

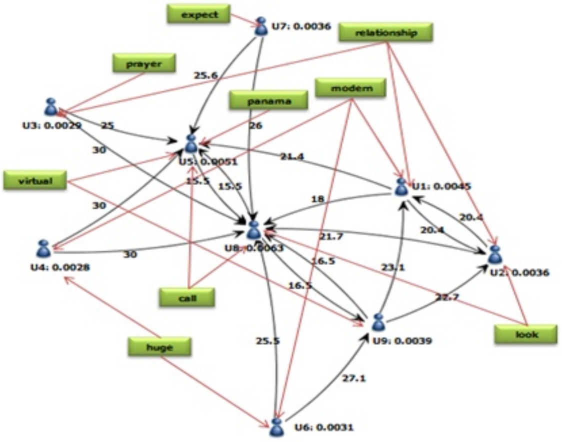

FIG. 4 – *The heterogeneous weighted user-word graph*

3.1.3 The Third Level in the Social Graph Cube Lattice

In this level, end-users can explore the original network by traversing three kinds of multidimensional spaces. In this way, we can examine, analyze and answer some complex queries that could be very useful for decision support and business intelligence purposes. Accordingly to the number of entities involved in the end-users requests, heterogeneity of graphs generated in this level is increased. As a result, these heterogeneous weighted graphs leverage the rich semantic surprisingly rich knowledge hidden in the massive social structure. The list of the produced heterogeneous weighted graphs is: user-location-time, user-location-word, word-location-time, user-word-time graphs. The process utilized in this level is similar to the process illustrated in the second level. The only difference is that, rather than defining the two multidimensional spaces concerned in the end-users needs, we start by determining the three multidimensional spaces involved in the analysis request.

3.1.4 The Fourth Level in the Social Graph Cube Lattice

It represents the union of all the four dimensional cuboids that develops the most heterogeneous graph in the *Social Graph Cube*. The set of relationships presented in this heterogeneous weighted graph could relate different kinds of multidimensional spaces. We can cite the word-user, word-location and word-time relationships which characterize the semantic closeness between the selected word and the other dimensions. In the case of user-location, user-time and location-time relationships, we have two types of connections. The first type describes the social relationships. For instance, we relate a user to a specific location or time interval if this user belongs to this location or sent a message within the selected time interval. The second type demonstrates the semantic closeness between the top-K most representative words characterizes the first part and the second part in this relationship. As a result, the list of visions displayed in this level of lattice not only captures much richer knowledge than in the other levels, but also the various relationships across the different types of topological entities can carry several semantic significations.

4 The Social Graph Cube Aggregations

In the case that the decision makers are concerned with zooming into more high-level granularity in order to get a summarized view of generated multidimensional networks, a roll-up operation may be carried out. The proposed *Social Graph cube* integrates OLAP technologies, community extraction methodologies and data mining clustering in a unified approach in order to represent the social data in a summarized visualization. In our previous work presented in Hannachi et al. (2012), only the semantic axis is used to define the list of users' communities. However, in this proposed approach all possible aggregations given a social multidimensional network can be determined by using both the set of the topological attributes associated with networks entities such as the number of followers, the selected language, etc, and the semantic, geographic and temporal axes. For instance, if the topological and the semantic relationships between vertices are considered in the aggregation phase, then the used process to determine the list of clusters is as follows: first, compute the topological and the semantic distances using the length of the shortest path and the semantic distance presented in Equation 4, respectively. Second, the agglomerative strategy is utilized to extract the users clusters or location clusters by using the content and topological distance computed previously. The agglomerative is a bottom-up approach that uses nodes as clusters and combines these nodes according to distances, until getting the dendrogram which represents the visualization of the nodes coalescing in clusters. Third, the extracted clusters are evaluated to get the best result. The researchers in Girvan and Newman (2002) give the answer by his popular modularity measure that evaluates the extracted communities. It is calculated by comparing the number of edges within community minus expected number in an equivalent network with edges placed at random. Moreover, authors in Arenas et al. (2007) present an extension of modularity for directed graphs. The adapted formula is:

$$Q = \frac{1}{m}\textstyle\sum_{i,j \in V}(A_{ij} - \frac{k_i^{out},k_j^{in}}{m})\delta(C_i, C_j) \qquad (8)$$

where A_{ij} are the elements of the adjacency matrix of $G(E, V), E$: edge, V: vertex. k_j, k_i: the in-degree and out-degree of nodes j, i. m: the number of edges, $\delta(C_i, C_j)$ is 1 if i and j belong to the same community, and 0 otherwise.

However, we think that this measure is more suitable for clusters, that are defined, based on the topological properties, contained in classic graphs. It computes only the link between users contained in classic graph without considering the semantic relationship between them. Thus, in our approach, we compute the modularity by using the weighted graphs constructed in *Social Graph Cube*. In an aggregated graph $G' = (V', E', W_{V'}, W_{E'})$, the V' is a set of condensed vertices, while the E' represents the set of condensed edges illustrated in the graph G'. $W_{V'}$ is the list of weights associated with each condensed vertex. It is computed as the mean of all the centrality values associated with each entity in the condensed vertices. $W_{E'}$ is the weights of the condensed edges. It defines the value of the semantic closeness between two condensed vertices by choosing the minimal distance among all the relationships values that relate these condensed vertices.

5 Experimentation

In this section, we present brief experimental studies evaluating the functionalities that the *Social Graph Cube* can provide for analyzing social network data. We trained the proposed *Social Graph Cube* on data collected by crawling one month of public tweets. The total number of tweets contained in our collection is approximately 4 millions. We select the first 3000 relevant users according to a list of criteria such as: the number of followers, the total number of retweets, etc. All our experimental methods are implemented in Java and tested on a Windows PC with dual Intel Xeon processors (3.06 GHz six-cores) and 12G of RAM. In these experiments, we are concerned with the features of tweets locations from different perspectives such as semantic, geographic and temporal axes. As an instance of the several remarkable obtained results in *Social Graph Cube*, we present in Figure 5 (a) a compressed vision of the list of countries which are considered semantically related.

As an instance, by considering this representation, we can answer several queries such as: *"Semantically, what is the closest (or distant) location to China?"* or *"How is the relationship between two countries is developed from one period of time to another?"*. In Figure 5 (b), we illustrate the top six most pertinent words detected in each cluster presented in the Figure 5 (a). This list of words is selected by using *TF-IDF* weight. From the table presented in Figure 5 (b), we notice homogeneity within each cluster where most of its countries treat related words with height values. From this information we can derive the semantic of each cluster. For instance, it seems that cluster C1 is mostly related to technology area, cluster C2 is concerned by the political field; cluster C3 focuses on economic and cluster C4 on travel.

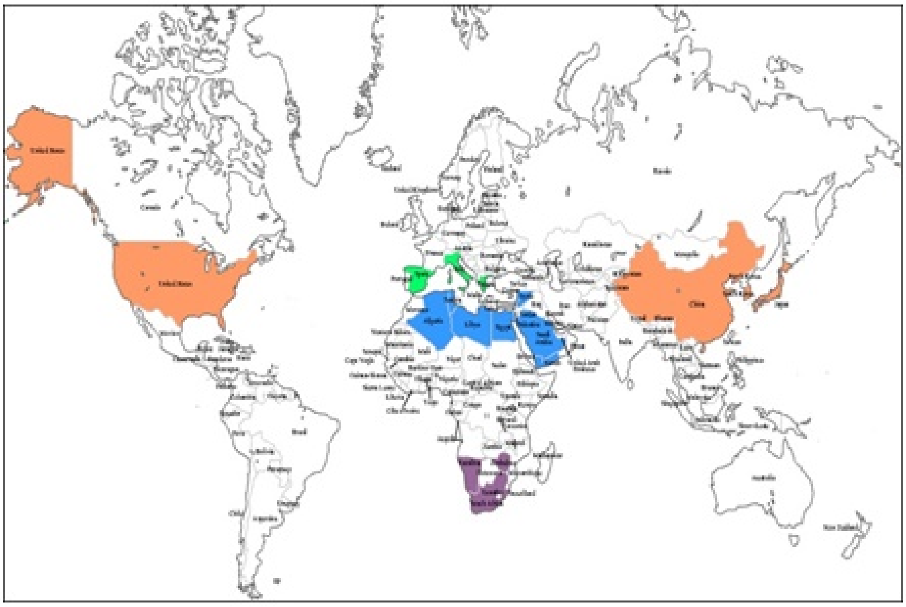

FIG. 5 – *Compressed vision with the top used words of semantically related countries*

6 Conclusion and Future Work

In this paper, we proposed a new data warehouse model, called *Social Graph Cube* for analyzing social networks data. Our *Social Graph Cube* is designed to support OLAP-style multidimensional analysis on information-enhanced multidimensional social network. It provides pertinent answers to analysis queries. Going beyond traditional OLAP operations where OLAP aggregations are directly computed by using the list of attributes associated with the relational data, *Social Graph Cube* proposes a new method that combines data mining techniques and OLAP operators to navigate through dimension hierarchies. It consists in grouping network entities into different clusters according to their similar interests, characteristics and views, which provides to be much more meaningful and comprehensive than classic aggregations in the traditional OLAP techniques. Different from the most community extraction methods that focused on the relations between users, our proposed approach suggests a new clustering method in order to represent the social data in a summarized vision. The set of clusters are determined by using both the topological structure of the social network and the semantic relationship between the network entities. Moreover, it permits the end-users to detect the emerging interests or orientations in each cluster. The experimental results show the efficiency and the effectiveness of our proposed *Social Graph Cube* for decision-making based on social Network data. It may help to develop more effective strategies in this area. In the future work, we will plan to go further in this analysis by using other types of measures proposed in social network area.

References

Arenas, A., J. Duch, A. Fernandez, and S. Gómez (2007). Size reduction of complex networks preserving modularity. *CoRR abs/physics/0702015*.

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America 22*(6), 725–730.

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science 10*(2), 161–163.

Bringay, S., N. Béchet, F. Bouillot, P. Poncelet, M. Roche, and M. Teisseire (2011). Towards an on-line analysis of Tweets processing. In *Proceedings of the 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Part II*, pp. 154–161.

Chen, C., X. Yan, F. Zhu, J. Han, and P. S. Yu (2009). Graph OLAP: A multi-dimensional framework for graph data analysis. *Knowl. Inf. Syst. 21*(1), 41–63.

Cilibrasi, R. L. and P. M. B. Vitanyi (2007). The Google similarity distance. *IEEE Trans. on Knowl. and Data Eng. 19*(3), 370–383.

Donetti, L. and M. Muñoz (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, P10012.

Du, H., M. W. Feldman, S. Li, and X. Jin (2007). An algorithm for detecting community structure of social networks based on prior knowledge and modularity. *Complexity 12*(3), 53–60.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*(3-5), 75 – 174.

Girvan, M. and M. E. J. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*(12), 7821–7826.

Hannachi, L., O. Asfari, N. Benblidia, F. Bentayeb, N. Kabachi, and O. Boussaid (2012). Community extraction based on topic-driven-model for clustering users Tweets. In *Proceedings of the 8th International Conference on Advanced Data Mining and Applications, ADMA 2012*, Volume 7713 of *Lecture Notes in Computer Science*, pp. 39–51. Springer.

Hannachi, L., N. Benblidia, F. Bentayeb, and O. Boussaid (2013). Social microblogging cube. In *Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP, DOLAP'13*, pp. 19–26. ACM.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.

Johnson, E. L., A. Mehrotra, and G. L. Nemhauser (1993). Min-cut clustering. *Math. Program. 62*, 133–151.

Newman, M. (2003). Fast algorithm for detecting community structure in networks. *Physical Review E 69*.

Newman, M. E. J. and M. Girvan (2004). Finding and evaluating community structure in networks. *Physical Review E 69*(026113).

Pedersen, T. B., C. S. Jensen, and C. E. Dyreson (1999). Extending practical pre-aggregation in on-line analytical processing. In *Proceedings of 25th International Conference on Very Large Data Bases, VLDB'99*, pp. 663–674. Morgan Kaufmann.

Pedersen, T. B., C. S. Jensen, and C. E. Dyreson (2001). A foundation for capturing and querying complex multidimensional data. *Inf. Syst. 26*(5), 383–423.

Qu, Q., F. Zhu, X. Yan, J. Han, P. S. Yu, and H. Li (2011). Efficient topological OLAP on information networks. In *Proceedings of the 16th International Conference on Database Systems for Advanced Applications, DASFAA'11, Part I*, pp. 389–403. Springer-Verlag.

Rafanelli, M. and A. Shoshani (1990). STORM: A statistical object representation model. In *Proceedings of the 5th International Conference on Statistical and Scientific Database Management, SSDBM'90*, Volume 420 of *Lecture Notes in Computer Science*, pp. 14–29. Springer.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smyth (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI'04*, pp. 487–494. AUAI Press.

Tian, Y., R. A. Hankins, and J. M. Patel (2008). Efficient aggregation for graph summarization. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD'08*, pp. 567–580. ACM.

Zhang, D., C. Zhai, J. Han, A. Srivastava, and N. Oza (2009). Topic modeling for OLAP on multidimensional text databases: Topic cube and its applications. *Stat. Anal. Data Min. 2*(56), 378–395.

Zhang, N., Y. Tian, and J. M. Patel (2010). Discovery-driven graph summarization. In *Proceedings of the 26th International Conference on Data Engineering, ICDE 2010*, pp. 880–891. IEEE.

Zhao, P., X. Li, D. Xin, and J. Han (2011). Graph Cube: On warehousing and OLAP multidimensional networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD'11*, pp. 853–864. ACM.

Résumé

Les services du réseau social ont rapidement devenir un puissant moyen de communication par lequel les gens partagent des messages en temps réel. Typiquement, les réseaux sociaux sont modélisés par des grands graphes. Répondant à cette nouvelle tendance, il devient extrêmement important de visualiser et d'analyser cette quantité massive de données à partir de différentes perspectives et sous plusieurs granularités. Bien que le traitement analytique en ligne (OLAP) est une technique trés performante pour analyser les données structurées, il fait face à des défis majeurs pour manipuler et traiter ces données complexes. Dans cet article, nous proposons un nouveau modèle multidimensionnel, appelé *Social Graph Cube* pour étendre l'analyse multidimensionnelle OLAP à l'analyse des réseaux sociaux. *Social Graph Cube* représente les données sous forme d'un graphe d'information hétérogène pour une illustration plus exhaustive que la technologie OLAP traditionnelle. Allant au-delà des opérations multidimensionnelles OLAP, *Social Graph Cube* propose une nouvelle méthode qui combine l'OLAP et la fouille de données OLAP afin de naviguer à travers les hiérarchies de dimension. Les résultats expérimentaux montrent l'efficacité de notre modèle pour la prise de décision.