**UNIVERSITY OF BLIDA1**

**Faculty of Sciences**
Department of Computer Science

# DOCTORAL THESIS

Specialty : Business Intelligence

## SOCIAL NETWORK ANALYSIS AND ONLINE ANALYTICAL PROCESSING

Defended by

**HANNACHI Lilia**

Doctoral Committee:

| | | | |
|---|---|---|---|
| S. OUKID | Associate Professor, | U. of Blida1 | Chair |
| H. ABED | Professor, | U. of Blida1 | Examinator |
| N. BOUSTIA | Associate Professor, | U. of Blida1 | Examinator |
| L. BELLATRECHE | Professor, | U. of Poitiers (France) | Examinator |
| N. BENBLIDIA | Professor, | U. of Blida1 | Advisor |
| O. BOUSSAID | Professor, | U. Lyon 2 (France) | Co-Advisor |

Blida, November 2016

# Abstract

Recently, social network sites, like Twitter and Facebook, have emerged as a hugely important communication utility. These services not only make users able to exchange information about their personal views, and opinions but also enable them to discover interesting knowledge or news. The result of these services has led to the accumulation of enormous amounts of structured and unstructured data. Although, this content is generally noisy, ambiguous, ungrammatical and unstructured text, but it is a rich data set where most likely users try to pack substantial meaning into this space.

Business Intelligence (BI) represents a set of technologies and systems that play a major role to deliver the right information mined from large amounts of data for decision making and future planning.

The main objective of this thesis is to determine how Business Intelligence techniques like On-Line Analytical Processing tools could help analyzers in providing effective and efficient decision-making about several kinds of complex data arising in real-world situations such as social media services which represents a huge complex and multidimensional content.

To achieve the semantic analysis and to group users according to similar interests, we propose a new model called Community Extraction based on Topic-Driven-Model (CETD).

To determine the list of pertinent information mined from the social network data,we suggest a new data warehousing model, Social Graph Cube to support OLAP technologies on multidimensional social networks.

To extend decision support services on social data with complex features, we propose a new multidimensional model, called Microblogging Cube, for efficient and effective exploration of data contained in the social network sites. To analyze and understand the information behind social network services, we suggest a new dynamic data cubing and mining framework, called Microblogging Community Architecture.It presents the ability to flexibly explore social data and get a fresh and timely perception of the semantic data generated in online social channels.

**Keywords:** Business Intelligence, Data Warehousing, OLAP, Data Mining, Social Network Analysis, Community Extraction.

# Resumé

Récemment, les réseaux sociaux, comme Twitter et Facebook, ont émergé comme des outils de communication extrêmement importants. Ces services non seulement rendent les utilisateurs capables d'échanger des informations sur leurs opinions et leurs perspectives, mais aussi de leur permettre de découvrir les derniers événements et réactions.

Le résultat de ces services a conduit à l'accumulation d'énormes quantités de données structurées et non structurées. Bien que, ce contenu soit généralement bruyant, ambiguës et non structuré, mais il est très riche en terme d'information et des connaissances.

Business Intelligence (BI) représente un ensemble de technologies et de systèmes qui jouent un rôle majeur pour fournir la bonne information extraite à partir de grandes quantités de données. Ces techniques représentent l'informatique à l'usage des dirigeants d'entreprises. Le but majeur de BI et d'offrir une aide à la décision et à la planification future.

L'objectif principal de cette thèse est d'analyser les données générées par les réseaux sociaux à partir de différentes perspectives et avec différentes granularités en utilisant la technique OLAP. Cette étude pourrait aider les décideurs à fournir des décisions efficaces et effectives sur une énorme quantité de données complexes et multidimensionnelles.

Pour réaliser l'analyse sémantique et pour regrouper les utilisateurs en fonction des intérêts similaires, nous proposons un nouveau modèle appelé Community Extraction based on Topic-Driven-Model (CETD).

Pour déterminer la liste des informations pertinentes extraites à partir des données sociales, nous proposons un nouveau modèle d'entreposage appelé Social Graph Cube. Ce modèle permet d'adapter la technique OLAP selon les caractéristiques des données sociales multidimensionnelles.

Nous proposons un nouveau framework appelé Microblogging Community Architecture pour l'extraction et le cubage dynamique des données sociales. Il présente la possibilité d'explorer de manière flexible les données sociales et d'analyser les nouvelles perceptions en temps opportun.

**Mots clés:** Informatique décisionnelle, Entrepôt de données, OLAP, Analyse des réseaux sociaux, Exploration de données.

# Acknowledgments

First and foremost, I wish to express my deepest gratitude to my advisor, Professor Nadjia Benblidia. Without her continuous guidance, support, and encouragement, this thesis would not have been possible. Her keen insight, great passion, and rigorous attitude towards research deeply influenced me.

A huge thank you also goes to Professor Omar Boussaid for his academic guidance, his enthusiasm, and for his invaluable help throughout this period.

Special thanks to Dr. Saliha Oukid for her long-term collaboration and support through my whole Ph.D. Special thanks to Dr. Fadila Bentayeb for all her suggestions, her critical feedback and insightful comments.

I would like to thank all the other thesis committee members, Professor Hafida Abed, Dr. Narhimene Boustia, and Professor Ladjel Bellatreche, for their generous time, commitment, and their constructive suggestions on my thesis.

Lastly but most importantly, I'm grateful to my parents for their endless and unreserved love, who have been encouraging and supporting me all the time. To them I dedicate this thesis.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Context and Motivation

Business Intelligence (BI) represents a set of technologies and systems that play a major role to deliver the right information mined from large amounts of data for decision-making and future planning. One of the most important technologies in BI is On-Line Analytical Processing (OLAP). This technology represents a very powerful and flexible tool to deeply mine and analyze large data warehouses and industry applications. It offers analysts the ability to navigate through data collections at various granularities and from different angles in order to define exceptions and interesting parts. To support handling of user defined views, OLAP tools organize data in a multidimensional model as a data cube. It is based on two main concepts; the concept of fact (measure) which describes the events of interest for an analyst and the concept of dimension which specifies different axes the data can be viewed, and presented. To improve the navigation in the cube, the dimension values are typically organized along hierarchies of one or more Levels. Using operations such as roll-up, drill-down, slice-and-dice and pivot, the result of on-line analysis is viewed as points in a multidimensional space which enables analysts to analyze quickly and navigate through the data from different perspectives and with multiple granularities.

Recently, social network sites, like Twitter and Facebook, have emerged as a hugely important communication utility. They allow people to communicate with the world and share their current activities, spontaneous ideas, photos and videos with the other users. These services not only make users able to exchange information about their personal views, news, opinions, and preferences but also enable them to discover interesting knowledge or news. The result of these services has led to the accumulation of enormous amounts of structured and unstructured data. The social user-generated content has perspectives in many domains such as, friend recommendation, opinion analysis, users' topics, etc. However, this content is generally noisy, ambiguous, ungrammatical and unstructured text, but it is a rich data set where most likely users try to pack substantial meaning into this space. Thus, it is important to understand the information behind messages transmitted in social network services and to detect the latent topics presented by them. To detect these topics, many applications propose to use a topic models like, PLSA (Probabilistic Latent Semantic Analysis) [Hof99b] or LDA (Latent Dirichlet Allocation) [BNJ03] which are effective and powerful approaches to detect the different latent topics. However, it presents each topic by a distribution of words. Thus we cannot extract the semantic concepts associated with each topics.

Typically, the topological structures of the social networks can be modeled as large underlying graphs with vertices representing entities and edges depicting relationship between these entities [AW10]. Multidimensional attributes are usually defined and assigned to the network entities, forming the so-called multidimensional networks. The analysis of the social network structure knows a huge attention from the researchers. It consists on studying the topological structure characteristics of the social networks, in order to derive much potential information such as, social influence, user's communities, etc.

Community extraction methodologies within the social networks are obviously raising an increasing interest from the researchers. They enable the identification of latent cluster by grouping users that share same characteristics and properties (interests). The studies of community extraction generally aim to extract the list of communities mainly according to topological structure. However, in the user graphs, the relations do not present the dynamics between users according to their common views or preferences. For instance, although an existence of a friendship between users, we cannot extract their common interesting information. That's why the study of user-generated content selected from the social network sites attracts much attention in recent years. We think that the relation between users, as defined in the classic methods such as [For10], [DMn04], is not enough when we look for users groups related to the same interest in order to recommend them some information.

In contrast, if we can apply OLAP to analyze the extracted communities from social network sites, we would be able to flexibly explore communities' data and get a fresh and timely perception of the online social channel. Unfortunately, the standard OLAP implementations cannot handle this kind of complex multidimensional data arising in real-world situations. The cause of this limitation, is that the traditional OLAP tools can manage 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 individual data tuples. It also faces great challenges for analyzing unstructured data such as the user-generated content. 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 [RS90a]. Another limitation to apply the OLAP on the social communities is that the existing methods on community extraction generally aim to extract the communities mainly according to topological relationship between users, which we think is not enough when we look for users communities according to semantic relationship.

While researches on modern networks have been in existence for decades [New10], an abundance of applications and algorithms have been proposed for decision makers in relational databases [CD97], [GCB+07], and some studies to support OLAP queries effectively on large multidimensional networks [ZHPL12], [ZLXH11a], [ZCY+08a], none has taken both data, the unstructured user-generated content and the topological structure into account in the multidimensional social network scenario. Moreover, none has proposed to combine OLAP-style multidimensional analysis together with the community extraction methodologies in order to give the analysts an unprecedently comprehensive view of users' and communities' data.

## 1.2 Objectives and Contributions

Considering the above, the ultimate objective of this thesis is to support OLAP-style analysis on information-enhanced multidimensional social networks for efficient information extraction. Particularly, we target four aspects: probabilistic topic models, community extraction methodologies, multidimensional modeling and OLAP operators. Thus, we contribute with the following:

- To achieve the semantic analysis, we propose to use the different words selected in social user-generated content, in order to specify the several treated topics. To detect these topics, various statistical models have recently been proposed such as topic model LDA, which is an effective and useful tool in text analysis. It presents text as a set of latent topics with different proportions. However, these topics are presented by a distribution of words, without associating understandable label with each of them. To overcome this limitation, we propose new model called Topic-Driven Model. In this model not only we are able to specify the label of each topic but also we can define the semantic hierarchy associated with the selected topic. This is achieved by using the Open Directory Project (ODP) taxonomy as an external resource, which represents the largest, and most widely distributed human-compiled taxonomy of web pages currently available. For instance, if we consider a user who writes always tweets in the domain sports like the following real-world tweet: “Contracts for Top College Football Coaches Grow Complicated”. Our proposed Topic-Driven Model will assign automatically the topics “football” to this tweet. Now, if we consider another user who writes the following real-world tweet: “Barcelona win 2-0 at Real Mallorca but Real Madrid return to form and smash Real Sociedad 5-1 ”. The detected topics for this tweet will be also: football although the word “football” is not mentioned in this tweet. This will show the important of our semantic addition to the classic topic model.

- Traditional community extraction methods generally aim to define the list of communities mainly according to the relationships between users. However, in the social networks, the set of relationships do not present the dynamics between users according to their common interesting information. Going beyond these methods, we propose a new community extraction method that permits the identification of latent semantic communities by grouping users that share same characteristics, opinions and interests. In this method, we incorporate both the semantic and the topological relationships into a unified approach to generate the list of communities. To define the semantic relationship we propose to calculate the closeness between the different words selected in the social user-generated content. However, as a word is a very specific unit and connected to different areas, we propose to go further in this study by providing analysts with additional semantic relationships such as the closeness between the different treated topics and the domains characterizing users over a period of time. This is achieved by using the both, the Topic-Driven Model and the ODP taxonomy.

- To determine the list of pertinent information mined from the social network data, we suggest a new data warehousing model, 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. In this model, 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 are used to : (1) illustrate the existence of relationships between the social entities, (2) turn out to define the aggregated network. Going beyond traditional OLAP operations, Social Graph Cube proposes a new method that combines data mining area and OLAP operators to navigate through hierarchies.

- To extend decision support services on social data with complex features, we propose a new multidimensional model, called Microblogging Cube, for efficient and effective exploration of data contained in the social network sites. It presents the possibility to analyze this data according to semantic, geographic and temporal axes. The proposed multidimensional model allows analysts to interactively analyze and navigate structured data together with network structure and unstructured text according to different perspectives and with multiple granularities. Moreover, we propose to combine the multidimensional style analysis together with the community extraction methodologies in a unified method to give the analysts an unprecedently comprehensive view of community's data Furthermore, in opposition to classical multidimensional models proposed in the literature, our Microblogging Cube presents two main advantages: (1) the list of selected measures depends on the hierarchical level and (2) the measures presented in this approach are well suitable for the analyses in the social networks.

- To analyze and understand the information behind social network services, we suggest a new dynamic data cubing and mining framework, called Microblogging Community Architecture. It presents the ability to flexibly explore social data and get a fresh and timely perception of the semantic data generated in online social channels. Unlike a traditional data cube where OLAP aggregations are directly computed by using the simple multidimensional attributes associated with dimensions, Community Cube presents an advanced unstructured text analytics capability for defining aggregations by proposing new clustering method. It consists on grouping individuals according to similar characteristics and interests, which provides to be much more meaningful and comprehensive than classic aggregation in the traditional OLAP techniques. Both topological and semantic relationships between users are combined into one integrated framework for the definition of communities. Moreover, we suggest new solutions to answer different OLAP queries in the multidimensional social network scenario. Besides traditional OLAP queries, our approach introduces a new class of queries, which takes into account the multidimensional attributes associated with networks entities, the social user-generated content and the topological structure of the networks. This type of queries has shown to be effective and useful to navigate through unstructured text and topological structure.

## 1.3 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 outlines the background for the research. Chapter 3 is a brief review of related work in literature. Chapter 4 presents the proposed approach, CETD: Community Extraction based on Topic-Driven-Model. This approach combines the proposed model used to detect pertinent semantic information of the user's tweets based on a semantic taxonomy together with a community extraction method based on the hierarchical clustering technique. Chapter 5 introduces and gives details of the suggested Social Graph Cube which furnishes pertinent responses to OLAP-style multidimensional analysis on information-enhanced multidimensional social networks. Chapter 6 describes the multidimensional model "Microblogging Cube" used to achieve OLAP techniques on the unstructured and related microblogging data. Chapter 7 introduces the Community Cube architecture, which represents advanced data cube architecture for efficient information extraction from the social network services. Chapter 8 presents the different experiments that we conducted on real-world tweets. Finally, we will conclude in chapter 9 and we will suggest some future research work in this area.

# Chapter 2

# Background

## 2.1 ABSTRACT

The main objective of this thesis is to determine how Business Intelligence techniques like On-Line Analytical Processing tools could help analyzers in providing effective and efficient decision-making about several kinds of complex data arising in real-world situations. One example is the social network data which represents a huge complex and multidimensional content.

In this chapter, we discuss the background of our research study from different areas. This includes the introduction and the definition of the different concepts associated with Business Intelligence, Data Warehousing and Social Network Analysis domains.

## 2.2 Business Intelligence

### 2.2.1 Definition

The concept of Business Intelligence (BI) was presented by a set of IT-consultants in the mid-90's, which is identified by the Gartner group [TALS06]. However, the term of decision support system has especially become a recent area of focus since the early 1970's in [GSM71]. Turban et al. (2007) maintains that the notion of Business Intelligence have developed from the decision support systems, while the researchers in [Neg04] asserts that the term of Business Intelligence has substituted various concepts such as management information system and decision support system. Davenport and Harris in [DH07] deduce that the entire area of decision support systems occasionally indicate the concept of Business Intelligence. In the studies presented in [Dav10], [SK10] Business Intelligence is considered as a decision support system that is frequently employed as a fundamental to support decision making. The primary purpose of Business Intelligence process is to treat huge quantities of organizational data produced through the past business transactions or further types of activities in order to get meaningful business information, which could support enterprises to ameliorate the timeliness, quality and accuracy of their performance.

Depending on Davenport [Dav06], the requirement for Business Intelligence process has emerged from heightened competition where all industries propose comparable services and employ the same technical equipment, the major source of discrimination is the business processes. Shared factor between the various determinations presented in the research of Shollo and Kautz's [SK10] is that Business Intelligence is frequently makes reference to a continuous process. This process is characterized as collecting and storing organizational data, such as structures and business transactions, then, based on the analysis this data is converted into information. Finally, the generated information is converted into knowledge that is used to take the most effective decision in given conditions and situations. This process make companies keep up with competitors. Thus, it is critically important for companies to define how to exploit the large amount of organizational data, which in itself is no use for analysis or decision making and how transform it into meaningful business knowledge.

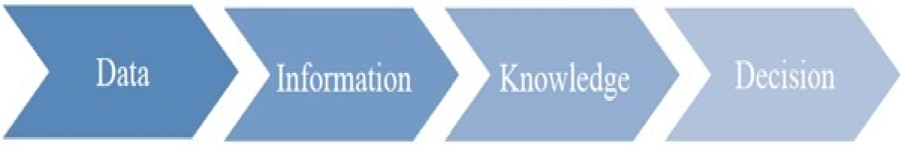


Fig. 2.1: Business Intelligence as a continuous process (Source: Shollo & Kautz, 2010)

## 2.3 Business Intelligence components

As we can notice, the BI represents prominent initiatives for organizations. However, there is few issued research that demonstrates the applications and determines answers that are backed by experimental data. Actually, only few studies submitted by practitioners have considered the aspects of BI from the viewpoint of IT professionals.

Eckerson [Eck03] defines the Business Intelligence (BI) as *"BI solutions create learning organizations by enabling companies to follow a virtuous cycle of collecting and analyzing information, devising and acting on plans, and reviewing and refining the results. To support this cycle and gain the insights BI delivers organizations need to implement a BI system comprised of data warehousing and analytical environments"*.

As we can notice, Eckerson [Eck03] highlights various significant sides of BI. He also proposed a new model to describe the BI Component Framework. In this model, the BI environment is divided into two aspects. In the first aspect, data shall be extracted, cleaned, transformed and loaded into a data warehouse. This aspect represents the data warehousing environment. There are two main sources of data: internal databases, which represent generated data through past business transactions or other type of business activities and external databases, which mostly refer to external data sources, such as text files, web pages, unstructured organizational data, etc. The process of extracting, transforming and loading transaction data into the data warehouse is named the ETL-process (short for Extraction, Transformation and Loading). It represents a key and an initial phase in the Business Intelligence system. In the second aspect, users will be able to retrieve, study and explore the huge amount of data stored in the data warehouse by using analytical tools, which demonstrate the output of BI system. This aspect represents the analytical environment. The model proposed in Eckerson (2003) is illustrated in Figure 2.2.

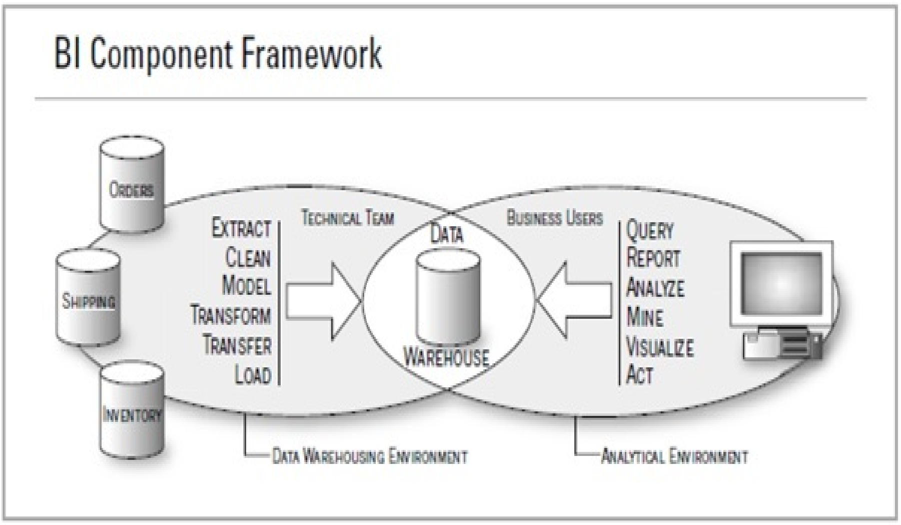


Fig. 2.2: Business Intelligence component framework, Eckerson (2003)

From the concept that all organizational data are stored and available in the same data warehouse, this make the analysis and the study of this huge amount of data sets generated in companies a very simple and easy task. To this effect, there are various analysis and reporting systems proposed in the literature. The two main popular analytical components utilized in Business Intelligence are Online Analytical Processing (OLAP) and Data Mining. The concept of OLAP refers to an analytical technology that permits users, managers and decision makers to better understand the data via quick, consistent, interactive access to a great variety of potential visions of information. It translates relational data models into multidimensional models, which offers analysts the ability to navigate through enterprise data collections at various granularities and from different angles in order to define exceptions and interesting parts. The data mining technology refers the process of extracting and discovering unsuspected useful information hidden in large amount of data sets stored in the data warehouse, in which some pertinent people are concerned. This useful information represents patterns that could demonstrate to be of considerable business value.

The graphical interfaces represent the final tools used in Business Intelligence to display the obtained analysis results for the end users, which are mainly managers, analysts or other business users.

## 2.4 Data Warehouse

### 2.4.1 Definition

The notion of data warehousing solution begin to appear in the industry area since the early 80's, however in the early 90's it's great value for the decision-making purposes was recognized. Among the most frequently referenced definitions of data warehousing technology is provided by Inmon [Inm96]. He describes a data warehouse as: *"a subject oriented, integrated, non-volatile and time-variant collection of data in support of management's decisions."*

The subject orientation refers that the organizational data are organized in the form of area-centered by using experiment metadata like, subject, orientation, and business rather that the details of the current activities of the organization which are application centered. The target of this representation is to develop a powerful data structure that may support the recuperation of transactional data through the use of metadata criteria. As an instance, claims are a crucial business theme for an insurance company. Thus, all claims data stored in the data warehouse are structured around the subject claims rather than being organized around operational applications such as auto insurance or workers' compensation insurance.

The integration characteristic in a warehouse system depends on two phases; first, the pulling of the pertinent data produced by several organizational systems, second, the incorporation of all this data into the same physical location. From the notion that the whole pertinent data are obtained from various operational system which could be databases, files, web pages, etc, the cleaning phase is become indispensable to store the generated data in a data warehouse. In our case, we have used one source of social network data which are real-world tweets. However, as tweets are generally noisy, unstructured data and associated with a set of additional data such as user identifier, Time, Longitude, Latitude, etc. Thus, we use different treatments such as consistent textual and semantic data, consistent additional data, etc.

The third characteristic in a warehouse system is the non-volatility. It is based on the concept that the obtained data are stored in the physical emplacement, without changing or removing previous stored data. In the case of wrong data, it is integrated or the ability of the data warehouse is exceeded which make the task of archive becomes substantial. This feature makes data warehouse in contrast to operational databases where operations like update and delete are allowed to change stored data.

The last feature of a data warehouse is time-variant. As opposed to operational systems where the main objective of these systems is to obtain current information of stored data by supporting day-to-day current operations, the data sets contained in the data warehouse are stored over a long period of time. In the decision making area, if a user is looking to find out the specific reason for the drop in sales in a particular division, the user needs to select pertinent data not only about current state of sales but also about all the sales data which take place in the selected division over a period extending back in time. Therefore, the effectiveness of a warehouse system in the level of request response time becomes an issue. As a solution the data sets stored in a data warehouse are organized as snapshots over past and current periods of time.

### 2.4.2 Warehousing process

As we have seen in the previous subsection, the most popular and widely used definition of a warehousing system is that a data warehouse intended to be used essentially as a store of organizational data sets that has been obtained from operational databases. Despite the fact that this determination is valid, it is incomplete since it leads to a basic misunderstanding between what it is a data warehouse and what it does. As defined in [Kim96], a dynamic visualization of a data warehousing systems can be presented as a process with back-end and front-end tools and where the data warehouse is installed in the middle. In Figure 2.3, we present the warehousing process proposed by Kimball in [Kim96]. The back-end tools, describe the extraction of organizational data from internal and foreign sources, the transformation of resulted data in order to correct data anomalies and finally, loading them into the data presentation field. These tools are displayed in the left hand side of the figure. The front-end tools correspond to querying and analyzing the stored data by utilizing several tools, like ad hoc querying, reporting, OLAP and data mining.

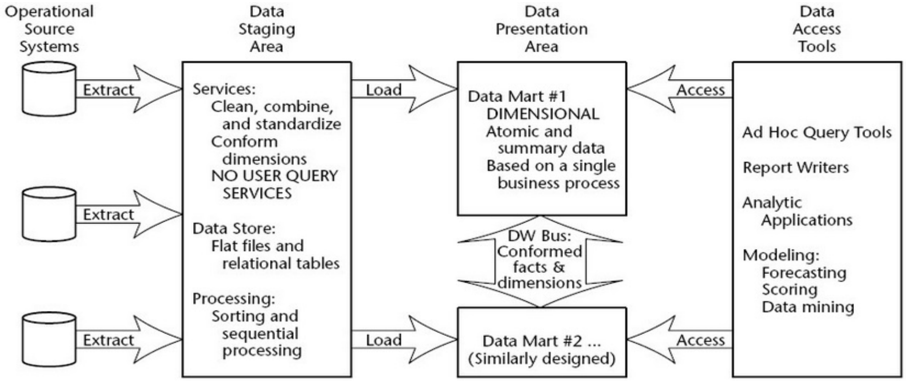


Fig. 2.3: The data warehousing process Kimball (1996)

### 2.4.3 Architectures

In the definition of warehousing system proposed by Inmon [Inm96] is that all pertinent organizational data sets including past and current data are stored and integrated within a data warehouse. Although, the data warehouse considerable value to the decision-making area, more and more, companies argue that a data warehouse system is too expensive. Depending on Phil Blackwood in 2000 *"The average cost of data warehouse systems valued at $1.8 million"* [Bla00]. Therefore, only large companies can afford to buy it. Furthermore, a data warehouse is an extremely complicated system. It can lead to considerable complexity in the business procedure. As an instance, slight modification in the transaction processing system can have main implications on all transaction processing system. Moreover, it can cause time consuming. Instead of traditional data warehouse, many recent studies suggest a virtual data warehouse to supply numerous of the benefit of a real data warehouse, with probably less cost.

The decision to maintain the organizational data in their operational systems or to physically stored them within a data presentation zone in a data warehouse leads to mainly three types of data warehouse architectures: (1) the real data warehouse architecture, where all relevant data are integrated in the same physical zone (2) the virtual data warehouse architecture, where data sets are logically integrated when necessary for use and (3) the remote data warehouse architecture which is the combination between the real and the virtual architectures [Fra97]. These three types of architectures are studied in details as follows.

#### Real data warehouse architecture

It is the typical and the most used architecture in a warehousing system. It is displayed in Figure 2.4. In this architecture, the decision maker directly access the pertinent data produced from various operational systems via the data warehouse. In this case, data warehouse contains not only the metadata and pertinent transactional data of a traditional OLTP system, but also summary data which represents a pre-aggregated data.

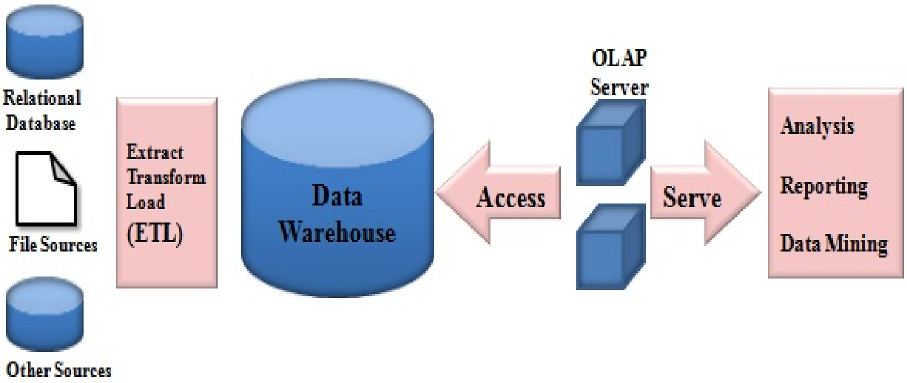


Fig. 2.4: Real architecture of a data warehouse

Summaries are extremely precious in warehousing system since they pre-calculate long activities beforehand which accelerate the request response times. However, this type of data warehouse architecture is not practical for real-time warehousing systems because data update is performed during timeouts.

### Virtual data warehouse architecture

Generally, Data quality issues and complicated integration requests render impossible to provide coherent, incorporated data real-time to several decision making applications. As a result, to overcome these limitations, a new middleware layer is incorporated between transactional systems and decision making systems. This virtual data warehouse architecture is demonstrated in Figure 2.5.

This representation permits a real-time data access where the decision makers express their requests as if the integrated organizational data was physically materialized, while actually, all organizational data sets are kept in their sources. Therefore, the middleware

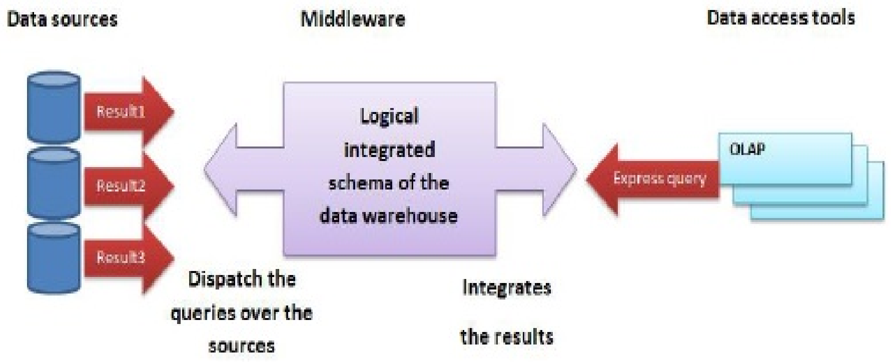


Fig. 2.5: Virtual architecture of a data warehouse [Bou13]

layer is in charge of: first, managing and converting the decision makers request into sub-requests, second, forwarding the sub- requests to their corresponding sources, third, gathering and combining the obtained results into a single result, to be transmitted to the decision makers. In this case, there is no need to update data because this architecture supports the requirements of on demand data processing where organizational data sets are already refreshed in their sources. As a result, some performance mechanism such as pre-aggregated and materialized view [CD97] cannot be utilized.

### Remote data warehouse architecture

This type of data warehouse architecture is displayed in Figure 2.6. In this case, from one side, whenever the organizational data is changing at rapid rate, the warehousing system does not correspond to the proper status of this data. Thus the pertinent data are retained in their organizational system. From other side, due to the great amount of stored data in the data warehouse, query response times are considered very long. Therefore, the Summaries are materialized within the data warehouse. As we can notice, this type of architecture is incorporation between simple and virtual architectures.

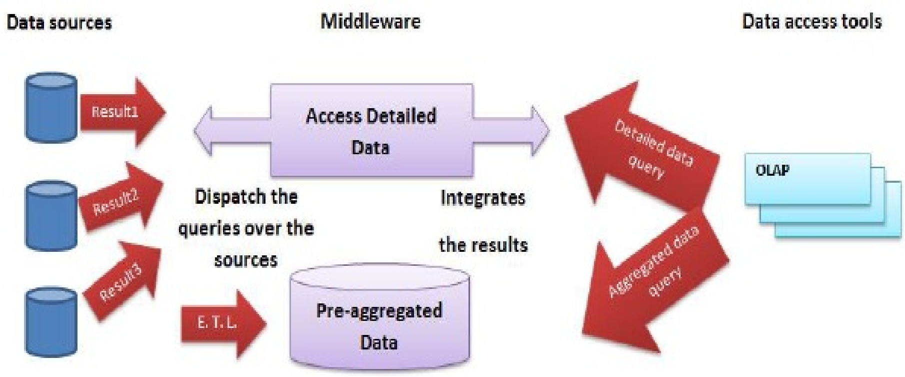


Fig. 2.6: Remote architecture of a data warehouse [Bou13]

## 2.5 Multidimensional Modeling

### 2.5.1 General principle

Characteristically, the development of organizations implies several standards of abstraction. A data model represents the implicit illustration of system or database structures. It consists on a list of conceptual techniques for characterizing the real-world concepts to be modeled in the structure of a database and the relationships between these concepts. Data models vary in the quantity of meaningful detail that can be extracted. The different proposed data models divided into three kinds of classes depending on the level of abstraction as follows:

- The conceptual level where data limitations are explicitly determined.
- The logical level where the underlying implementation of the database is illustrated in order to attain optimal runtime execution and storage space exploitation.
- Physical level where data at the lowest level is demonstrated.

The Entity-Relationship (ER) data model represents one of the most extensively used representatives at the conceptual level. The ER model utilizes three types of concepts to abstract a system or a database: entity sets, relationship sets and attributes. Entities determine the list of objects present in the real world. Attributes define the set of properties associated with the list of entities. A relationship is a combination between entities. The ER model is frequently mapped into a relational model which constitutes the most widely adopted model to perform the data warehouse models. The strength of the relational data model is that this model is based on mathematical foundations. In this model, a relation is used to represent both the list of entities and the list of relationships. While, both the ER model and the relational data model are considered very valuable in different domains, they are not appropriate for modeling data produced in the decision making area [Kim96]. In fact, the data models presented previously are based on the concept of normalization which is applied to avoid data incoherence, inconsistency and redundancy. However, depending on [Fra97], a normalized schema raises the risk of losing domination through data since each decision maker could create custom views of corporate responsibility performance. Thus, in the trend analysis these characteristics such as redundancy of data increases the integration and update cost, however, they can ameliorate the timeliness of the query response. As a result, we need to maintain these characteristics at some cases of granularity levels in order to define the best business decision.

### 2.5.2 Multidimensional model

In the data warehousing area, the next step after determining the business queries and the field of study is to design the information integrated in the data warehouse. The design associated with the data warehouse structure is dissimilar from the design associated with the operational systems. The main source of this dissimilarity is that the multidimensional model associated with data warehouse applications furnishes the decision makers with an analysis-oriented visualization rather than a transactional visualization. Further, in data warehousing systems it is more important to have quick access to the huge amount of data sets than to eliminate irregularities. Thus, as a result, a multidimensional schema has fewer constraints about selected data with regard to a transaction-oriented system. Therefore, the multidimensional models are considered as the most appropriate data model for the data warehousing systems. Depending on Chaudhri and Dayal [CD97], the multidimensional visualization of data stored in the data warehouse is significant when developing tools, database structure and query processing for online analytical processing (OLAP) field. It provides fast answers to analyst queries. The multidimensional models categorize data as being either facts which describe the measures of interest for an analyst, or as being dimensions which specify different axes the data can be presented. In order to support multiple granularities, dimensions are typically organized along hierarchies of one or more Levels.

According to [PJ99] the basic multidimensional fact schema is a two-tuple *S= (F, D)*, where $F$ is a fact type and $D = T_i, i = 1, ..., n$ is its corresponding dimension types. The dimension type $T$ is divided into a set of categories $C = C_j, j = 1, ..., k$. Each category represents the values associated with a level of granularity. $\boldsymbol{T}$ is also presented with a partial order ($\leq_t$) on the $C_j$'s, with $\top_{\boldsymbol{T}} \in C$ and $\perp_{\boldsymbol{T}} \in C$ being the higher and finer level of the ordering, respectively. We note $e \in D$ where $D$ is a dimension of a type $T$, to indicate that $e$ is a dimensional value of $D$, if there is a category $C_j \subseteq D$ such that $e \in \cup_j C_j$. In [PJD99], the authors present a multidimensional object *(MO)* as a four-tuple $M$ = *(S, F, D, R)*, where $S$ = *(F, D =$T_i$)* is the fact schema, $F$ is a set of facts *f*, $D = D_i, i = 1, ..., n$ is a set of dimensions, and $R = R_i, i = 1, ..., n$ is a set of fact-dimension relations. The authors also define the important hierarchy properties of this multidimensional object as follows: given three categories, $C_1$, $C_2$, $C_3$, such that $C_3$ is one of immediate predecessors of category $C_2$ and $C_2$ is one of immediate predecessors of category $C_1$:

- The mapping from $C_1$ to $C_2$ is *Onto* iff $\forall e_2 \in C_2(\exists e_1 \in C_1(e_1 \leq e_2))$. Otherwise, it is *into.*

- The mapping from $C_1$ to $C_2$ is *Strict* iff $\forall e_1 \in C_1(\forall e_2, e_3 \in C_2(e_1 \leq e_2 \wedge e_1 \leq e_3 \Longrightarrow e_2 = e_3))$. Otherwise, it is *non-strict.*

- the mapping from $C_2$ to $C_3$ is *covering* with respect to $C_1$ iff $\forall e_1 \in C_1(\forall e_3 \in C_3(e_1 \leq e_3 \Longrightarrow \exists e_2 \in C_2(e_1 \leq e_2 \wedge e_2 \leq e_3)))$. Otherwise, it is *non-covering* with respect to $C_1$.

### 2.5.3 Data Warehouse Schema Models

A schema is a list of metadata about database objects such as tables, views and indexes used to demonstrate the structure of the database in a formal language. There several methods of ordering schema objects in the schema models developed for data warehousing applications. They can be classified into three forms: a star schema, a snowflake schema, a hybrid schema. In the following subsection, we will explain each schema model selected from these three forms.

#### Star Schema Model

The star schema is probably the most straightforward way to represent the structure of a data warehouse. Depending on Chaudhuri and Dayal [CD97], star schema is utilized in the majority data warehouses to demonstrate the multidimensional model. It takes its name from its visual representation, which resembles a star [Nag06]. A star schema consists of one fact table located at its center which describes the list of measurements of facts related with a specific business domain and of many dimension tables including the attributes that demonstrate the measures which is generally, a descriptive data. All the attributes for a dimension are integrated into one denormalized table. Further, the primary key of a fact table is composed of the association of one of more foreign keys linked to different dimension tables. However, in this schema the dimension tables are not related to each other.

### Snowflake Schema Model

The snowflake schema is a complex variation of the star schema where the dimension tables are normalized. It is schematized with a central fact table surrounded by a list of its constituent dimension tables (such as a star schema), and those dimension tables may additionally normalized by decomposing them into subdimension tables which form a snowflake pattern. These dimensions are linked by a primary key of one dimension and the foreign key of the core dimension at a higher level. By reason of the normalization, the redundancy of data is reduced which minimizes the disk space and therefore improve query performance. However, in the snowflake schema and because of increase number of look ups table, supplementary upkeeps are necessary. The joins are also required to define the list of characteristics of a dimension.

### Starflake Schema Model

A starflake schema is a compromise between the data warehouse structure generated by the star schema and the data warehouse structure produced by the snowflake schema. Therefore, a set of dimension tables are broken up into subdimension tables whereas some others keep the same representation of dimensions without decomposition.

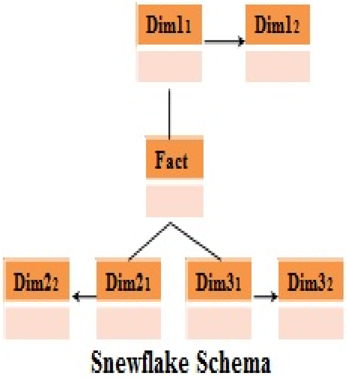


Fig. 2.7: The three representations of a relational multidimentional schema

## 2.6 On-line Analytical Processing (OLAP)

The interest of the multidimensional models is to determine the end-user needs in a formal representation. However, these models primarily concentrated on static visualization of designs generated by data warehouse structure without presenting the possibility of possibility of gaining insight into data, in order to mine significant hidden knowledge that reflect the real dimensionality of the organizational decision-making. In the related literature, the online analytical processing (OLAP) tools permit the decision makers to retrieve and access data stored within the data warehouse by manipulating a multidimensional model. In the following subsection, we will study in details the different concepts associated with OLAP technology.

### 2.6.1 Definition

As an important technology of BI, OLAP technology, which represents a very powerful and flexible tool to mine and analyze data deeply. The concept of OLAP was first invented by Edgar F. Codd [CCS93]. It permits the business analysts to gain insight into the large amount of information that has been processed from raw data through specific interfaces. In OLAP, a multidimensional metaphor named hypercube or data cube for short represents a way to organize data in a multidimensional model in order to support handling of user defined views of data. The data cube is generally displayed as a three dimensional view which can be easily displayed in a graphical representation. An example of this representation is schematized in Figure 2.8. This data cube demonstrates sales that the business analysts desire to analyze along dimensions location, product and time.

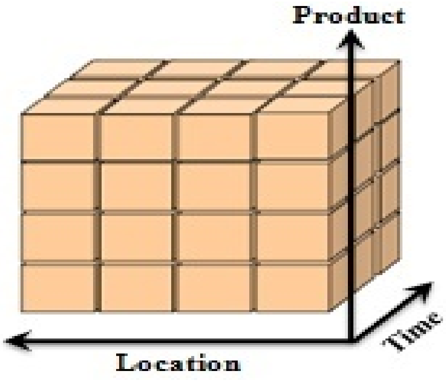


Fig. 2.8: The graphical representation of a data cube

For reasons of clarity, the list of dimensions product, location and time are identified by using a product name, a location name and a day, respectively. However, in excess of three dimensions, this representation of a data cube is unsuitable. To overcome this limitation, a cube is then demonstrated in other ways, for example by mathematical notations. It is based on two main concepts; the concept of fact (measure) which describes the events of interest for an analyst and the concept of dimension which specifies different axes the data can be viewed, and presented. To facilitate navigating the cube, the dimension values are typically organized along hierarchies of one or more Levels. Using operations such as roll-up, drill-down, slice-and-dice and pivot, the result of on-line analysis is viewed as points in a multidimensional space which enables analysts to analyze quickly and navigate through the data from different perspectives and with multiple granularities.

### 2.6.2 OLAP versus OLTP

Online Transaction Processing (OLTP) applications are primarily designed to treat daily transactional operations such as insertions and deletions. They are carried out by means of relational databases which contain a normalized tables and relationships. However, Owing to their aim of utilization, the implicit characteristics of OLTP applications are not appropriate for supporting complicated requests such as the list of queries generated in the decision-making area which could include zooming in to more detailed data or zooming out to less detailed data depending on various aspects of a business. Online Analytical Processing (OLAP) is a broad concept applied to reference to such applications that are fundamentally intended for analytical purposes, but not for transactional operations and data updates. It offers to the business analysts the possibility of extracting pertinent knowledge hidden in the huge amount of data stored in the data warehouse in order to take the best decisions. Table 2.1, due to [Kel97], recapitulates the principal differences between operational databases which employ OLTP and data warehouses which employ OLAP.

### 2.6.3 OLAP Operations

As already mentioned, the basic features of the multidimensional model is that it arranges data depending on several dimensions, and each dimension consist of several levels of abstraction organized by the notion of hierarchy. This characteristic of illustrating data permits the analysts to visualize data from different perspectives and at several levels of details.

To exploit this representation, there are numerous OLAP operations that offer the possibility to materialize these different perspectives and navigate through the data sets in order to mine pertinent information deeply. A basic list of OLAP operations on

Table 2.1: Differences between operational databases and data warehouses

| **Feature** | **Operational Databeses** | **Data Warehouses** |
|---|---|---|
| Users | Thousands | Hundreds |
| Workload | Preset transactions | Specific analysis queries |
| Access | To handreds of records, write and read mode | To millions of records, mainly read-only mode |
| Goal | Depends on applications | Decision-making support |
| Data | Detailed, both numeric and alphanumeric | Summed up, mainly numeric |
| Data integration | Application-based | Subject-based |
| Quality | In terms of integrity | In term of consistency |
| Time coverage | Current data only | Current and historical data |
| Updates | Continuous | Periodical |
| Model | Normalized | Denormalized, multidimensional |
| Optimization | For OLTP access to a a database part | For OLAP access to most of the database |

multidimensional data is presented in the following.

- ***Roll-up***. This operation consists in transforming full measures into abbreviated ones on a data cube. This is achieved by moving up in a hierarchy or by dimension reduction where one or more dimensions are removed from the cube. Consequently, the result of this operation increases the abstraction level associated with multidimensional visualization.

- ***Drill-down***. This operation is the opposite of the roll-up operation. It consists in transforming the visualization from a more general view to a detailed view by zooming into more lower-level granularity in the hierarchy for additional details. A drill-down could also be the equivalent of bringing back a previously deleted dimension into the data cube. Therefore, the outcome of operation decreases the level of abstraction.

- ***Slice-and-Dice***. The slice operation is achieves the selection on one dimension of a particular data cube, resulting in sub-cube. Generally, it is utilized develop cut profiles of complicated view. The representation graphic of a sliced cube is illustrated as a rectangular-shaped because is matches to restricted value of the selected dimensions. The dice operation is a slice operation on two or more dimensions of data cube which means that it determines a sub-cube by realizing a selection of more than one dimension. Dicing a cube limits its data through a complicated predicate on two or more dimensions at the same time.

- ***Pivot*** or ***Rotate***. This operation modifies the dimensional orientation of a data cube. It rotates the axes of visualization associated with data stored in the data warehouse, in order to supply the end-user with an alternative representation graphic of selected underlying data.

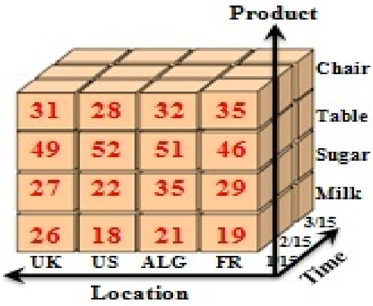


Fig. 2.9: cube representing agregated sales by month and by country

This list of OLAP operations described previously represents the most common used and cited OLAP operators in the related literature. Furthermore to these fundamental operations, OLAP techniques offer a lot of other operations such as merge, push, join and pull which are used in a wide diversity of financial, mathematical and statistical areas. A complete list of OLAP operations and their characteristics are studied in details in [GR09].

## 2.7 Social Network Analyses

### 2.7.1 Social Network Definition

The concept social network has been officially presented by Barnes (1954), although some notions have already been studied earlier. Depending on the description proposed by Wasserman and Faust [WF94], a social network can be determined as *"a finite set or sets of actors and the relation or relations defined on them. The presence of relational information is a critical and defining feature of a social network"*. Thus, the notion social network is based on two concepts; the concept of actor which describes a set of social entities such as users or social organizations and the concept of relationship which characterizes the connections or the exchanges between the social entities. In order to maximize the potential value around the effective representation and visualization of social networks, two main formal representations are then suggested, the first is a formal mathematical illustration (sociometric) and the second is a graph illustration (graph theoretic). However, according to related literature, the graph illustration is the most used representation to characterize social networks. The researches on graph theoretic have been in existence for decades. They are mainly related to mathematics and computer sciences fields. Therefore, since variety of networks could be demonstrated as graphs, it is possible to benefit greatly from the progress made in graph theoretic area. Depending on the definitions proposed by Wilson and Watkins in [WW90], a graph is described as: *"a nonempty set of elements, called vertices and a list of unordered pairs of these elements, called edges. The set of vertices of the graph is called the vertex set of G, denoted by V(G), and the list of edges is called the edge list of G, denoted by E(G)"*. By utilizing the terminology of graph theoretic in the network sociology field, each social actor is depicted by a vertex in the graph, while the interactions between the social actors are represented by a list of edges. Edges can be directed or undirected.

Undirected edges are usually named reciprocal ties. These interactions could refer a casual conversation, joint collaboration, exchange information about personal views, news, opinions, preferences, etc. Interaction intensity between vertices in the social network indicates the strength of the relationship which is illustrated by the edge weight. Figure 2.10 demonstrates a simple instance of social network with directed and weighted edges. The arrows describe the direction of the interaction.

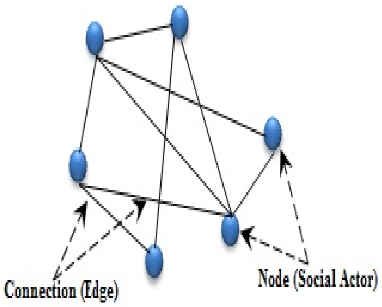


Fig. 2.10: Example of ego network

### 2.7.2 Social Network Analysis (SNA)

Social network analysis plays an important role to provide a good basis for the improvement of an analytical pattern of a social network in the organizational standpoint. From the list of researches marked the beginning of social network analysis, we can cite the study of Mayo in [May33] which represents the primarily researches on social behaviors of employees in enterprises and the study of Warned and Lunt in [War37] and [WL41] that considers the structure of cliques and kinships in the social communities. Social network analysis assess the strength of the relationships among people, social communities or even the whole organizations, in order to determine which are the key actors or what positions and actions they are likely to take" [Kra96]. This type of analysis could reveal abundant meaningful information hidden within the flow of communication that are generated inside the social organizations. The antecedent fields of social network analysis could be attributed to different number of domains. As an instance we can mention psychology [Mor37], [Rap50], anthropology [Str43] [Bar54], [RBEP59] [Mit74] [W.78] or others [Kra96] [MC99].

### 2.7.3 Social Network Analysis Measures

After representing the social network as a graph, it is indispensable to determine the list of methods that permit the exploitation of this pertinent source of data. In this section, we give a brief summary of important measures to perform social network analysis. They permit an in-depth analysis of social structures based on two different levels: actor level and network level. Therefore, these measures can be classified into two separate classes; the first class evaluates the entire network while the second class evaluates a specific vertex in the generated network [HSS10], [vdARS05]. In the network level, the dimensions are concluded from the graph theoretical analysis. In the following subsections, we will list and discuss some existing measures.

#### Measures for an individual level

As part of the administration of social acquaintance, not just from a scientific point of view, but also for commercial or strategic reasons, the requirement to specify the role or the influence of actor (i.e., a vertex in the graph) within the network is very important. It permits the end-user to identify if an actor is a principal leader or is insulated from the remainder of the network. Therefore, the determination of significant measures for weighting the importance of the vertices and / or edges of a particular network is required. There are lots of concepts about actors that can be considered. Thus, we clarify some of the settings that are typically utilized to make these concepts.

- ***Degree Centrality***. The simplest way to measure the importance of a vertex in a particular network is to use the degree centrality, which represents the number of edges directly linked to the selected vertex. This measure can be regarded as the popularity or the social activity of each individual in the social network. Burt (1980) [Bur80] illustrates the degree centrality measure as “ego density” where a high degree reflects a high level of social influence across the network. The degree of a vertex can be determined by using either directed or undirected edges. On the one hand, if an undirected graph is being used, every edge is regarded for computing the degree. On the other hand, if a directed graph is being used, it is possible to make a distinction between two types of edges, incoming and outgoing edges. Therefore, the degree centrality would be divided into two metrics. The first metric represents *in-degree centrality* which evaluates the number of vertices that are directly connected to the selected vertex. The second metric represents *out-degree centrality* which evaluates the number of vertices that the selected vertex points toward.

- ***Betweenness Centrality***. This measure evaluates the impact and the ability that a particular vertex has to control the dissemination of information and the flow of communication across the Network. It is originally suggested by Girvan and Newman [NG04]. Consequently, this measure does not only consider direct and indirect relationships between two vertices, but rather examines the relationships with three vertices implicated. In the social network case, a social actor (i.e., vertex in the graph) with high value of betweenness centrality signifies that it effectuate a critical role in the network since this actor allows the communication between two distinct partitions and if for some reason, this actor is no longer available in the network, then as a result, no exchange of information and acquaintance between these partitions is possible. The betweenness centrality of a vertex $v$ in a particular graph  is based on the concept of shortest path and can be expressed as

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.1}$$

where $s$ and $t$ are vertices in $V$, $sigma_{st}$ is the number of shortest paths relating $s$ to $t$, and $\sigma_{st}(v))$ is the number of shortest paths relation $s$ to $t$ passing through the vertex $v$.

- ***Closeness Centrality***. This measure determines the proximity of each vertex to the other vertices in the network. It determines the effectiveness of a social actor according to not only the access of any other actor in the network more rapidly than some other vertices, but also the independence of selected actor within the network. Contrary to other centrality measures where the high value represent The most influential individual, in this measure a lower value of closeness centrality denotes social actors with crucial role in the network, while the actor with higher value of closeness centrality. Therefore, this measure does not only depend on direct relationships, but also on indirect relationships in the network. The closeness centrality of a vertex $v$ can be expressed as

  $$C_C(v) = \frac{1}{\sum\limits_{u \in V} d(v,u)} \tag{2.2}$$

  where $d(v,u)$ represents the geodesic distance between the selected vertex $v$ and another vertex $u$.

- ***Eigenvector Centrality***. This type of measure is close to the definition associated with the degree measure since it computes the number of relationships that a vertex has in the network. However, this measure goes further in the centrality by considering that the actor which has a lot of influential neighbors, it should be influential as well. It was suggested for the first time by Bonacich in [Bon72]. The eigenvector centrality does not take into account the degree of vertex solely, but the degree of other vertices that are directly related to this selected vertex as well. Degree and closeness centrality propose a more local determination of the importance of each vertex within the whole network, whereas, the betweenness and the eigenvector centrality define a more global evaluation. The eigenvector centrality of a particular vertex $v$ is

$$C_E(v) \propto \sum_{u \in N_v} A_{vu} C_E(u) \tag{2.3}$$

wher $N_v$ is the neighborhood of the vertex $v$ being $x \propto AX$ that implies $AX = \lambda x$

- ***Clustering Coefficient***. This measure is based on the perspective that if a first actor in the social network is related to a second actor, which in its turn is related to a third actor, then there is a heightened probability that the first actor will be also related to the third actor. It was proposed for the first time by Watts and Strogatz [WS98] and has since received a lot of attention from researchers [NWS02] and [SKO$^+$07]. This measure describes the propensity of a network to the cliquishness or clustering and demonstrates the inter-connective power between vertices in a neighborhood within a particular social network. The notion of cliquishness in the social context denotes the network where all its vertices are linked with each other by direct relationships. Therefore, the clustering coefficient is computed by dividing the number of directed connections relating vertex's neighbors by the number of all possible directed connections which may exist between the vertex's neighbors. Consequently, if a social actor has a high value of clustering coefficient, this signifies that this actor is much integrated in the network. While if a social actor has a low value of clustering coefficient, this means that it is a peripheral vertex (i.e., has lack of novel acquaintance and information) and more outlying from all other vertices presented in the network. The clustering coefficient of a vertex $v$ can be expressed as

$$C_v = \frac{n}{N_v(N_v - 1)} \tag{2.4}$$

where $N_v$ denotes all neighbors of vertex $v$.

### Measures for the network level

The list of measures presented previously is limited to a single social actor. However, when making a global network analysis it is indispensable to extract some meaningful characteristics about the entire network. As an instance, the list of pertinent data generated within the network can be exploited to: (1) specify the ability of the network to be divided into smaller sub-networks denoted by clusters, (2) define the particularity of a network such as the density or the sparsity characteristics. With a view to determine kind of information, we demonstrate some of the measures that are usually used to achieve these concepts.

- ***Density***. This measure is correlated to the degree of all vertices contained in the graph, i.e. the correspondence between the current number of edges and the maximum potential number of edges within the whole network, relative to the number of all existing vertices. In the social context, a particular network with a high value of density signifies that everyone communicates with everyone else in the entire social network. The density is defines as

$$Density = \frac{\sum_{i=1}^{v} \sum_{j=1}^{v} e_{i,j}}{v(v - 1)} \tag{2.5}$$

where $v$ indicates the number of vertices represented within the graph and $e_{i,j}$ denote existing edges between vertex $i$ and vertex $j$. The density value associated with each graph relies on its size, since the degree of probable relationships increases exponentially with any vertex inserted to the network.

- ***Clustering coefficient***. This measure defines the likelihood of a network to be subdivided into finitely many sub-networks (clusters). In the social context, a specific cluster is considered a particular category / class in the organization. The clustering coefficient of a graph is computed as the average of all its vertex cluster coefficients. This measure define the clustering coefficient associated with the graph , by the following formula

$$C(G) = \frac{\sum\limits_{v \in V} C_v}{|V|} \tag{2.6}$$

The obtained values range between 0 and 1. From one side, higher values of clustering coefficient characterize the graphs with greater degree of cliquishness between their vertices. While, from the other side, lower values denote the graphs with lower level of cliquishness. Specially, a clustering coefficient value equal to 0 characterizes graphs that do not contain triangles of related vertices, while a clustering coefficient value equal to 1 denotes graphs with the ideal clique.

- ***Centralization***. This measure is immediately related to the individual measures of centrality, demonstrated previously. It computes the degree to which a network is centralized or controlled by some individuals. Thus, the lower the number of leading vertices represented in a specific network, the higher value is the centrality of this selected network. Therefore, the network centralization is dependent on centrality values associated with each vertex contained within the entire network. If for some cause, a vertex with a critical role in the network is eliminated, then consequently, the network rapidly splits into distinct partitions. As a result, placing too much confidence and strength in a single social actor, may generate crucial issues of failure

### 2.7.4 Finding community structures in networks

Social network analysis tools depend significantly on the graphical visualization since it can be regarded as a powerful method to achieve the aim of this analysis and to help end-users comprehend social networks from different perspectives. This is the case of the list of measures demonstrated previously where the analysis of the social network is complimented with a graphical representation. When it comes to large networks, it is difficult and complicated to achieve the different type of social network analysis. To overcome this limitation, there are several algorithms have been proposed to determine the list of communities (sub-networks) contained within the network where the more this structural feature is obvious, the higher a network inclines to be partitioned into clusters of vertices whose connections are denser among vertices belonging to the same cluster compared to the connections between vertices of different clusters which are relatively sparse. From the structural viewpoint, the members that belong to the same communities could share same characteristics and properties (interests). Based on this aspect, the majority of commercial and strategic motivations are founded on the identification of communities inside a network in order to propose or recommend some suggestion. The issue of obtaining the best partitions of networks has a long history. It consists on the determination of the most pertinent clusters in a network depending on the similarity of views and interests between the lists of social actors.

# Chapter 3
# Related Work

## 3.1 ABSTRACT

The research presented in this thesis builds on two major domains. First, the Business Intelligence (BI) domain which represents a list of technologies that plays a major role in all decision support systems. One of these technologies is the On-Line Analytical Processing. It gives users the ability to dynamically analyze data from different perspectives and with multiple granularities. Second, Social Network Analysis domain which provides a good basis for the improvement of several kinds of analytical patterns used to study and analyze social networks.

In this Chapter, we illustrate the main related research works that have been done to extend these techniques toward new emerging studies generated from real-world situations.

## 3.2 Introduction

In this chapter, we will provide a brief overview of related work in the literature. The three main domains closely related to the thesis study are: (1) Online Analytical Processing (OLAP), (2) OLAP for the social network data, and (3) community extraction methodologies. Below, we will represent a brief review of the main studies in each of these associated domains.

## 3.3 Online Analytical Processing (OLAP)

Data warehouses play a major role to analyze and deliver the right information extracted from large amounts of data for decision-making. One of the most used technologies to exploit the data warehouses is On-Line Analytical Processing which represents a very powerful and flexible tool to mine and analyze data deeply [AAD$^+$96], [CD97], [GCB$^+$07]. It has been vastly used in various fields [fJPF07], [LKH$^+$08], [THP08]. OLAP on data warehouses is primarily underpinned by data cubes [CCLR05], [CDH$^+$02]. However, the majority of these studies are not intended to support the unstructured user-generated content such as the text data generated within the social services. The last few years, meaningful progress has been achieved to extend OLAP and data warehousing techniques from the relational databases to novel emerging data sources in various application areas such as streams [HCD$^+$05], sequences [LKH$^+$08], taxonomies [QCT$^+$08] and imprecise data [BDRV07].

## 3.4 Analysis of text data

In literature, there are different previous researches which have tried to examine and analyze the text data. From this list of researches, we can cite the two principal studies in Probabilistic Topic Modeling [BNJ03], [Hof99b]. Therefore, several topic models have been widely investigated in recent years [BL06], [MSZ07]. In the study presented in [SSRZG04], the researchers suggest a novel author-topic model to illustrate the words in a multi-author document like the outcome of a mixture of each authors' topic mixture. The study of [MWN$^+$09] defines a Polylingual topic models to reveal the closeness between documents aligned across several languages by defining their list of treated topics. The authors of [NAXC08] discuss the issue of integrating the text data contained in a list of documents and their citations in the topic modeling framework by taking benefits of liaison structures between documents and provide an enhanced estimate of latent topics. The works presented in [RHMGM09], [ZBZ$^+$08] examine the list of semantic topics treated in the web pages by associating the content of web pages with the user-generated tags.

The topic models have been used successfully to a wide variety of domains such as text mining issues. We can mention, the hierarchical topic modeling to illustrate the text's structure by revealing the relationship between topics [BGJT04], [Hof99b], opinion mining to identify the attitude of individuals about some topics [TM08], sentiment analysis to determine feelings expressed in text [MSZ07], multi-stream bursty pattern finding to define correlated bursty topic patterns and their bursty periods within different sources of text streams [WZHS07a], spatiotemporal text mining to study and analyse the text data in a temporal and spatial context [MLSZ06], information retrieval [WC06] and social network analysis [MCZZ08]. All these studies indicated that probabilistic topic models are very beneficial for examining and extracting latent topics hidden in the text data, and they are some of the most efficient text mining applications. However, all these presented studied are concentrating on the pure text data.

As we use, in our proposed approaches, the general probabilistic topic model LDA [BNJ03], in which the text collection is represented as a distribution of topics, and each topic is represented by a words distribution, we will present some related works which use the topic model LDA on twitter. The researchers, like [WLJH10a], use the standard topic model LDA in micro-blogging environments in order to identify influential users, but the proposed influence measure is based on the number of tweets, however the twitter users usually publish a large number of noisy posts. In addition, the work of [ZJW$^+$11] compares the tweets content empirically with traditional news media by using a new Twitter-LDA model in order to detect their topics. They consider that each tweet usually treats a single topic, but this is not always the case. The works of [MM10] present an approach to discover a Twitter user's profile by extracting the entities contained in tweets based on Wikipedia's user-defined categories.

## 3.5 OLAP for text analysis

As the proposed approach presented in this thesis aims to combine the text data analysis with OLAP technology in order to permit an efficient and effective exploration of pertinent information contained in both structured and unstructured data, we will present in this subsection, some previous researches that have attempted to study the unstructured text data by supporting OLAP techniques. The authors in [BK07] suggest a new tool to analyze and visualize the blogScope. They treat the unstructured text by representing it as a fact of interest. As an instance, the result of end-users queries can be determined as the term frequency. However, such study cannot handle some basic OLAP operations on the text dimension such as drill-down and roll-up operations. In the study presented in [WSR07], the unstructured text data is considered as a character field. In this case, the database records that contain the most pertinent data in the text filed in accordance with keyword query will be selected as an answer of end-users needs. However, this approach cannot handle the OLAP techniques on the text dimension. The authors in [CKKS02] suggest a new approach to represent text as a classified data. It utilizes classification techniques in order to organize document into different categories by associating each document with a class label. This process permits the end-users to achieve different type of OLAP operations on the text data along the category dimension such as the drill-down and the roll-up. However, in this approach only high-level function demonstrations are provided with no algorithms given. In [SBSR08], the authors suggested to represent the text data like a component of OLAP technology by integrating the keyword search and the OLAP tools in a unified framework, in order to get a powerful determination of pertinent data contained in the multidimensional text databases. In [LDH+08], the authors proposed a new data cube model called text cube, in which they incorporate the strength of classic OLAP and IR tools for text by utilizing the list of IR measures of terms to outline the unstructured text data in a cell. In [ZZH09a] and [ZZH+09b] a novel data model called topic cube model is suggested in orderto overcome the list of issues generated in the multidimensional text database by combining the traditional OLAP techniques with probabilistic topic modeling approaches and the obtained results are stored as probabilistic content measures. In addition, the works of [ZZH11], [ZZH13] suggested a new infrastructure called MicroTextCluster Cube, in which a compact representation of the unstructured text data contained in large collection of documents is displayed. This is achieved by determining the micro-clusters of text listed in the documents and store them in the multidimensional text databases.

## 3.6 OLAP for Graph databases

Typically, the topological structures of the social networks are modeled as large underlying graphs with vertices representing the social entities and edges depicting relationship between these entities. Thus, we will illustrate in this subsection, a few previous studies that have aimed to analyze graphs by supporting OLAP techniques. The principal purpose of the study presented in [ZCY+08b] is to analyze and examine the graphs data by using OLAP functionalities. This study is based on the aggregation of several graphs by considering as an input a list of related graphs and a list of attributes associated with the graph vertices, and as an output a summary static graph. As the OLAP operations are based on the representation of data in different aggregations, we cite some realized works related to the field of graph summarization such as the research of Navlakha et al. in [NSK09] which introduces a new graph summarization technique to cluster some large graphs generated in several scientific applications like the synthesis of proteins and DNA testing, which make the biologists able to solve their problems. In addition, the studies presented in [GKT05] and [NRS08] define the summarization of the original graph by using different compression methods. The authors in [ZCY09] propose to divide the input graph into a list of parts. Further, these parts can be clustered according to different properties. In [LT10] a new approach suggested to summarize the structure of original graphs by utilizing a random world model. Its main aim is to enhance the precision and the accuracy of similar queries about the graph structure such as adjacency, degree and eigenvector centrality. The study presented in [Wat06] combines the techniques of graph visualization and data reduction in a one unified framework in order to get refined and clarified visions of complex graphs. Moreover, in the works of Tian et al. [THP08] and Zhang et al. [ZTP10] 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 [CYZ+09], the researchers studied the requirement for OLAP techniques on the huge amount of graphs generated in several applications since existing researches did not consider the list of connections between data tuples. As a result, they suggested a new Graph-OLAP framework for graphs. This framework allows the user to obtain a multidimensional and multilevel visualization over graphs. In a subsequent paper, Qu et al. [QZY+11] proposed additional graph summarization operators for the Graph-OLAP framework. They suggest two characteristics to improve considerably the data analysis. The first one consists on the determination of the higher level summaries (Roll-Up operation) by utilizing pre-aggregated networks. The second one facilitates pruning the area of research while running a query. However, they consider the input data as a static graph and consequently do not include the impact of changes in the obtained graphs. Furthermore, dimensions in the proposed framework are illustrated as graph snapshot labels, instead of vertices and connections. The researchers in [ZLXH11b] recognize the requirement for OLAP on graphs databases and present a new data warehousing model, Graph Cube, by integrating properties of multidimensional 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 all previous researches, 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 exploit the information (i.e. unstructured data) transmitted in networks.

## 3.7 Community Detection methodologies

Typically, social networks are composed of communities matching to some social clusters. This characteristic permits the user to make a difference between the different social networks and other complex networks [NP03]. Community property is a similar process to illustrate the social network as a mesoscopic demonstration of the network topology [New06], [For10], [New11]. As in our approach, we aim to determine the list of communities, various studies have been carried out to examine the community properties of online social networks. From this list of realized researches in this area, we can cite [For10], [KLN08], [POM09], [SZ10], [TKMP11], [ZLZ11]. The issue of finding related vertices in a particular social network and placing them in as set of groups has been widely reviewed in several areas such as Physics, Bioinformatics and Computer Science [New03], [NBW06]. In the context of revealing the community structure of a network, two primary kinds of algorithms exist in the literature: (1) disjoint communities' algorithms, where each vertex listed in the network belongs to only one community, [GN02a], [New06], [BGLM08], [RB08] (2) overlapping communities' algorithms, in which the network vertex has the possibility to be considered as a membership of one of multiple communities, [PDFV05], [ABL10]. The community extraction concept is generated from the graph partitioning field in the graph theory, like the Kernighan-Lin algorithm [KL70], spectral partitioning [Fie73], [HK92], and hierarchical clustering [Sco00], [New11]. The traditional clustering methods, such as [For10], are based on the arcs density in the graph. For instance, the hierarchical clustering techniques like [HTF01], aim to identify vertices groups with high similarity. It can be divided into two classes: Agglomerative algorithms [DMn04], [DFLJ07] and Divisive algorithms [JMN93] and [New03]. 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 [NG04], 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. Modularity is extensively considered as a powerful measure to determine the set of communities listed in a particular network [New06], [BGLM08]. It is presented as the total number of edges in one group minus the expected edges placed at random, to define how well a list of graph vertices are clustered together [New06]. In fact, different from the existing works, we propose in this thesis several approaches to extend data warehousing and OLAP technologies toward such new complex data produced in the social network services. It provides relevant answers to OLAP-style multidimensional analysis on information-enhanced multidimensional social network. We also suggest a new model to cluster users tweets, in which, we combine both semantic hierarchy presented in ODP taxonomy and topic model (LDA) in order to improve the obtained results by adding semantics relations. Aggregation and OLAP operations are performed along the user dimension defined upon the social networking services. We aim to construct communities by considering the semantic of tweets contents as the only one inputs data instead of being as an additional information.

# Chapter 4

# Community Extraction Based on Topic-Driven-Model

## 4.1 ABSTRACT

Twitter has become a significant means by which people communicate with the world and describe their current activities, opinions and status in short text snippets. Tweets can be analyzed automatically in order to derive much potential information such as, interesting topics, social influence, user's communities, etc. Community extraction within social networks has been a focus of recent work in several areas.

Different from the most community discovery methods focused on the relations between users, we aim to derive user's communities based on common topics from user's tweets. For instance, if two users always talk about politic in their tweets, thus they can be grouped in the same community which is related to politic topic. To achieve this goal, we propose a new approach called CETD: Community Extraction based on Topic-Driven-Model. This approach combines our proposed model used to detect topics of the user's tweets based on a semantic taxonomy together with a community extraction method based on the hierarchical clustering technique.

## 4.2 Introduction

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 view and interests. For instance, Twitter is an online social networking service that has become a significant means in which users are able to send and read text-based posts of up to 140 characters, known as "tweets". It enables its users to communicate with the world and share current activities, opinions, spontaneous ideas and organize large communities of people. The service rapidly gained worldwide popularity. This imposes new challenges in the social networks and the microblogging data streams analysis in order to identify of 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. As a result, several researches have attempted to study and analyze the characteristics of tweets content in order to derive much potential information such as interesting topics, social influence, user's communities, etc. The tweets studies have perspectives in many domains such as, friends' recommendation, opinions analysis, users' topics, etc. However, the text of tweets is generally noisy, ambiguous, unstructured text data, ungrammatical, but it is a rich data set to analyze and most likely users try to pack substantial meaning into the short space, one subject by one tweet. Thus, it is important to understand the information behind the tweets and to detect the topics presented by them. Different proposed approaches in the literature are attempted to resolve the problems related with text mining area by determining the set of topics listed in the text data. One of the most powerful and effective approaches to define and mine the latent topic in a large text collection are the probabilistic topic models like, Probabilistic Latent Semantic Analysis (PLSA) [Hof99b] or Latent Dirichlet Allocation

(LDA) [BNJ03]. In fact, these probabilistic topic models have recently been considered as very efficient and successful techniques in a wide range of text mining issues. As an instance, we can cite, the learning of topic hierarchies [Hof99a], [BGJT04], the unsupervised learning techniques for author-topic analysis [SSRZG04], the spatiotemporal theme patterns for the text mining field [MLSZ06], the probabilistic models for capturing the mixture of topics and sentiments analysis simultaneously [MLW$^+$07], and multi-stream bursty pattern finding [WZHS07b]. Specifically, we will use in our approach the Latent Dirichlet Allocation model [BNJ03] so that the immense amount of complex data produced in social network services such as the twitter site would carry the parameters and the settings of a probabilistic model that can illustrate the Interesting/significant information among the text content. However, this model represents each document by a distribution of topics and each topic by a distribution of words. Table 4.1 presents an example of obtained results by using the LDA model. It characterizes the topic distribution of two different topics treated in a particular document. As we can notice, each topic is identified with its number and the list of word distributions associated with it, without determining the semantic label related with this selected topic. Therefore, it is difficult to extract semantic concepts of latent topics generated by this model since the word is a very specific unit and connected to different domain categories. To overcome this challenge, we suggest a new model called Topic-Driven-model to determine the concepts reconstructing the semantics of the distributed words by topic model in order to detect automatically the high level topics presented by the tweets. This is achieved by using the human-edited Taxonomy Open Directory Project (ODP) as an external resource which is considered as the most important and effective taxonomic directory on the web. The data in the ODP taxonomy is organized as a hierarchical structure with parent-child relationships between categories nodes. Thus, we define the domains distribution characterizing each topic, by identifying for each word

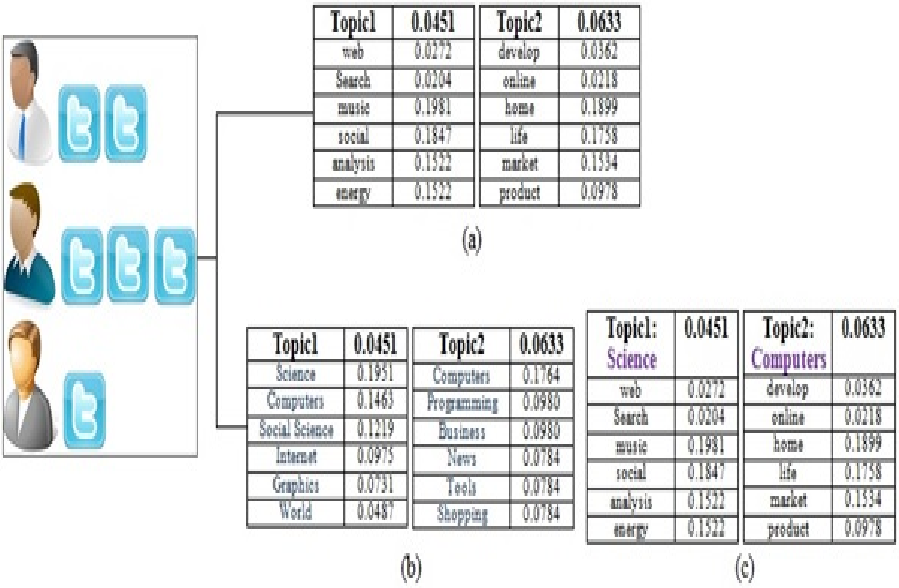
Table 4.1: Topics and their probabilities extracted from a particular document. Each topic is identified by the top five words and their probabilities given the corresponding topic

| | **Topics** | | | |
|---|---|---|---|---|
| | **Topic1** | **0.1863** | **Topic2** | **0.0791** |
| **Words** | germany | 0.0592 | develop | 0.0284 |
| | kitchen | 0.0278 | online | 0.0160 |
| | business | 0.0204 | product | 0.0143 |
| | flavor | 0.0186 | market | 0.0143 |
| | concept | 0.0094 | life | 0.0125 |

selected from the top-K most representative words characterizing the selected topic, its first top-J categories with their top three levels from the ODP taxonomy. The number of levels is selected based on the experiment results presented in [TG04a]. Communities are a user's groups that share same characteristics and interests. It is used in many applications, such as social networks, data mining, web searching, etc. 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 properties. 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 or common treated topics. 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 [DMn04] and [For10] is not enough when we look for users groups related to the same interest in order to recommend them some information. Different from the most existing community discovery methods which focus on the relations between users, we aim to derive user's communities based on common topics from user's tweets. For instance, if two users always talk about politic in their tweets, thus they can be grouped in the same community which is related to politic topic. To achieve this goal, we propose to combine both the high level-topics model and the user communities' extraction in a unified approach called CETD: Community Extraction based on Topic-Driven-Model. This approach is divided into two phases; in the first one, we aim to discover the different treated topics by users and their different related categories by constructing a topics tree based on ODP taxonomy. In the second phase, we aim to derive user's communities by grouping users according to three separated cases; topics similarity, domains similarity and topics-domains similarity. These three cases will be extracted from the first part. For instance, if we consider a user who writes always tweets in the domain sports like the following real-world tweet: "Contracts for Top College Football Coaches Grow Complicated". Our proposed topic-driven model will assign automatically the topics "football" to this tweet. Now, if we consider another user who writes the following real-world tweet: "Barcelona win 2-0 at Real Mallorca but Real Madrid return to form and smash Real Sociedad 5-1 ". The detected topics for this tweet will be also: football although the word "football" is not to the classic topic model. We note that the two users treat the same topic football. Thus we can regroup these two users in the same community based on their topics of interesting football.

## 4.3 Topic-Driven Model for Clustering Users Tweets

In this section, we present a new ontological topic model called Topic-Driven Model which is used to modelize the users' interests as high-level topics extracted from user-generated content by examining the terms they mention in their tweets and determine the domain distributions associated with each generated topic. To achieve this goal, we propose an ontological topic clustering model based on ODP taxonomy Open Directory Project as an external knowledge source in order to automatically derive from users' tweets the high level topics and the list of their related categories which explicitly determine the semantic and the meaningful of each generated topic. Thus, different from the works that use topic model on Tweets data by representing the topics as multinomial distributions over words, we will define semantics to each topic by constructing for it a multi levels semantic hierarchy. A semantic hierarchy is defined for each topic which allows detecting common topics between users that would not have been detected with LDA only. This model is considered as the first phase of our proposed approach.



| Topic1 | 0.0451 | Topic2 | 0.0633 |
|---|---|---|---|
| web | 0.0272 | develop | 0.0362 |
| Search | 0.0204 | online | 0.0218 |
| music | 0.1981 | home | 0.1899 |
| social | 0.1847 | life | 0.1758 |
| analysis | 0.1522 | market | 0.1534 |
| energy | 0.1522 | product | 0.0978 |

(a)

| Topic1 | 0.0451 | Topic2 | 0.0633 |
|---|---|---|---|
| Science | 0.1951 | Computers | 0.1764 |
| Computers | 0.1463 | Programming | 0.0980 |
| Social Science | 0.1219 | Business | 0.0980 |
| Internet | 0.0975 | News | 0.0784 |
| Graphics | 0.0731 | Tools | 0.0784 |
| World | 0.0487 | Shopping | 0.0784 |

(b)

| Topic1: Science | 0.0451 | Topic2: Computers | 0.0633 |
|---|---|---|---|
| web | 0.0272 | develop | 0.0362 |
| Search | 0.0204 | online | 0.0218 |
| music | 0.1981 | home | 0.1899 |
| social | 0.1847 | life | 0.1758 |
| analysis | 0.1522 | market | 0.1534 |
| energy | 0.1522 | product | 0.0978 |

(c)

Fig. 4.1: Comparison between LDA results and Topic-Driven Model results

Figure 4.1 presents a comparison between the list of results generated by using the classic LDA model and the list of results obtained by using the Topic-Driven model. Figure 4.1 (a) illustrates an example of topics generated by using the topic model LDA. As we can see, although the detected multinomial word distributions associated with the latent topics are frequently intuitively significant features, a major challenge facing this representation is to accurately expound the significance of each generated topic. In fact it is usually very hard for a simple individual to comprehend a topic depending solely on its multinomial distributions over words, essentially when the individual is not habitual with database area. Further, it is very complicated to evaluate the difference between one distribution and another word distribution.

In actual studies of multinomial topic models the researchers mostly, define the top words selected from the multinomial distribution as straightforward topic labels [BNJ03], [GS04], [Hof99b], [BL06]. However, this representation is not satisfactory since employing the top words is not very beneficial to illustrate the consequential significance of the latent topics.

Figure 4.1 (b) and (c) present the list of results obtained by using the Topic-Driven Model. The Figure 4.1 (b) illustrates the automatic produced domain weights associated with each topic. It is used to make the interpretations of topics easier and to facilitate the distinction between the different produced topics from several views. Figure 4.1 (c) displays the semantic labels associated with the two topic word distributions illustrated in Figure 4.1 (a). These labels are used to get a comprehensive vision about the meaning of the topic and the overall domain treated by it. To obtain these results, the proposed model combines the meaning of the topic terms by selecting for each term belongs to the top words list characterizing each topic its semantic hierarchy generated from the ODP taxonomy. Further, the utilization of our model is not restricted to the labeling topic models; our model may also be employed in every text management techniques in which the multinomial word distributions can be measured, like the labeling of the social user-generated content, the determination of the most treated domains and the summarization of unstructured text.

The architecture of this model is illustrated in Figure 4.2. We can divide it in four steps:

- ***Cleaning Database***, to clean the tweets corpus by using both a linguistic knowledge and a semantic knowledge.
- ***ODP-Based Adapted LDA***, to apply Latent Dirichlet Allocation (LDA) on the cleaned tweets data by using ODP taxonomy.
- ***ODP-Based Topics Semantic***, construct concepts sub-trees in order to detect the semantic relations between the words of each topic resulted after the applying of LDA.
- ***N-depth High Level Topics***, to present the topic in a higher level as a distribution of domains where domain is a set of topics.

Each part of the Topic-Driven Model is studied in detail in the following subsection.

### 4.3.1 Cleaning Database

In this step, we index the data collection; the Tweets corpus is stored in a database. Then, we organize Tweets data with the following attributes: *"Id, User, Time, Content, hashtag, URL and at_user"*. In order to clean our data base, we use both a linguistic knowledge and a semantic knowledge to process the Tweets corpus. Because linguistic knowledge does not capture the semantic relationships between terms and semantic knowledge does not represent linguistic relationships of the terms. In the linguistic processing phase, we do the following steps:

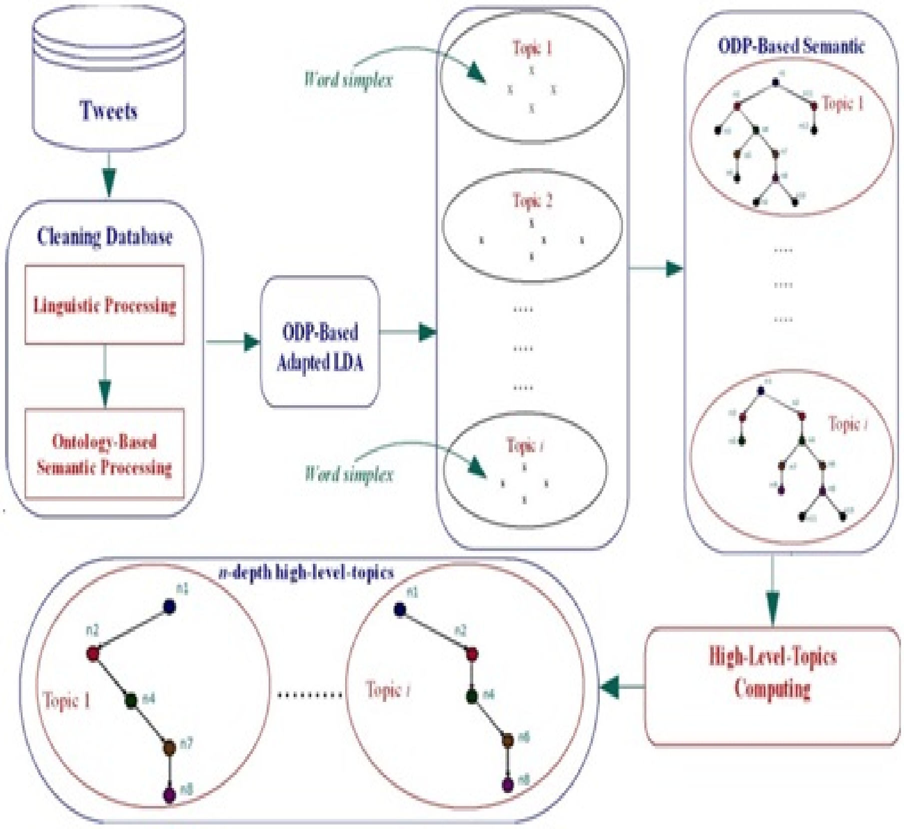


Fig. 4.2: Topic-Driven Model Architecture

- Remove stop-words such as: *the, is, at, which, on,* etc.
- Remove user names, hashtag and URL.
- Stemming: getting the word root (ex: *plays, playing*, etc, will be: *play*).
- Spelling correction.
- Use the WordNet dictionary to remove the noisy words.

After the first cleaning based only on the linguistic processing, we notice that many noisy or unrecognized words still existed after this step. To solve this problem, we clean the corpus by using a semantic knowledge, such as, the ODP Taxonomy as an instance of a general ontology. Here, we use ODP categories as a stopword filtering mechanism before applying the LDA model. Thus, we achieve the following steps to keep only the relevant words:

- We verify the existence of word in the results of the ODP indexing. If it does not exist, we remove it from the Tweets. For example, a noise words such as: *suprkkbwp, mirsku*, etc. can be removed in this step.
- If the word exists, we compute the number of documents (Web pages) which support it $ND(w_i)$. As cited in DMOZ site, a word may be helpful if the number of its WebPages more than 20. Thus, we suppose a threshold: $ND(w_i) >= 20$.Therefore, we remove all words that have a number of supported documents less than 20; i.e. $ND(w_i) < 20$. For example, if we consider the word *"awry"*, the number of its supported pages is 15, thus this word will be removed in this cleaning step. Figure 4.3 illustrates a small example of this task. We notice that many irrelevant words are removed after the cleaning step based on a semantic knowledge.

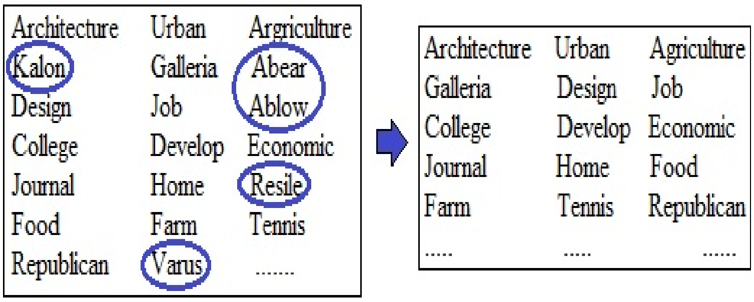


Fig. 4.3: An example of cleaning step

### 4.3.2 ODP-Based Adapted LDA

In this step, we apply Latent Dirichlet Allocation (LDA) on the cleaned tweets data. To apply LDA, we have to define the number of iterations, the number of words allocated to each topic and the number of topics. Thus, each collection of tweets generated by users is represented as a distribution of topics and each topic is represented as a words distribution. However, in order to verify the utility of using ODP Taxonomy in cleaning tweets data, LDA is applied two times: to the cleaned data by using ODP, and without using it. After comparison between the results in the two cases, we notice that the topics, sorted from the not cleaned corpus, have some noisy words with a high probability. For instance, if we specify the number of topics as three topics and five words for each topic, from one side, the words distributions for the list of topic generated by using only the linguistic phase are illustrated in the Table 4.2.

Table 4.2: Example which shows the five topics extracted from cleaned data without using the ODP taxonomy

| **Topic 1** | | **Topic 2** | | **Topic 3** | | **Topic 4** | | **Topic 5** | |
|---|---|---|---|---|---|---|---|---|---|
| agriculture | 0.0731 | sport | 0.0995 | art | 0.1218 | architecture | 0.0726 | pilot | 0.0864 |
| farmland | 0.0224 | college | 0.0262 | hi | 0.0309 | job | 0.0510 | go | 0.0290 |
| wheat | 0.0224 | football | 0.0262 | design | 0.0187 | right | 0.0294 | open | 0.0250 |
| farm | 0.0224 | top | 0.0219 | photograph | 0.0187 | create | 0.0151 | yahooansw | 0.0250 |
| urban | 0.0152 | baseball | 0.0219 | graphic | 0.0127 | auto | 0.0151 | question | 0.0250 |
| spring | 0.0152 | golf | 0.0133 | tweetmyjob | 0.0127 | remove | 0.0079 | year | 0.0168 |
| take | 0.0152 | via | 0.0133 | new | 0.0127 | sign | 0.0079 | new | 0.0168 |
| matter | 0.0152 | contract | 0.0133 | paint | 0.0127 | article | 0.0079 | world | 0.0127 |
| gap | 0.0152 | coach | 0.0133 | one | 0.0127 | risk | 0.0079 | medical | 0.0127 |
| gender | 0.0152 | grow | 0.0133 | photo | 0.0127 | chi | 0.0079 | wshgkir | 0.0127 |

From the other side, the words distributions for the list of topic generated by using ODP taxonomy are are illustrated in the Table 4.3.

As we can notice, the topics resulted, in the second case, are more significant and more homogeneity between words. As an instance, the list of results generated by using only the linguistic treatment such as the set of words: *"right, take, via, hi, new, tweetmyjob, one, sign, risk, chi, go"* are either a set of irrelevant words or very general entities that are used in various areas which make the interpretation of the topic meaning a major challenge. This difficulty is mainly related with the lack of available information to understand the direction of topics since the list of labels produced by using this list of

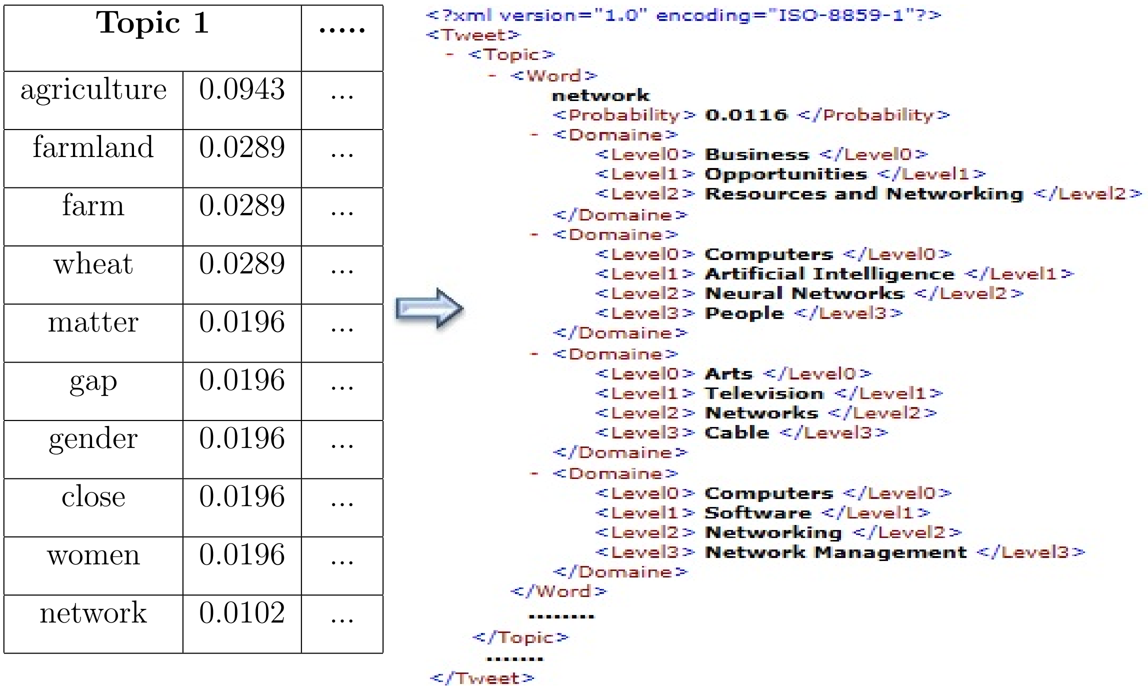
Table 4.3: Example which shows the five topics extracted from cleaned data by using the ODP taxonomy

| **Topic 1** | | **Topic 2** | | **Topic 3** | | **Topic 4** | | **Topic 5** | |
|---|---|---|---|---|---|---|---|---|---|
| agriculture | 0.0943 | sport | 0.1172 | art | 0.1335 | architecture | 0.1122 | pilot | 0.1134 |
| farmland | 0.0289 | college | 0.0309 | job | 0.0322 | job | 0.0344 | open | 0.0327 |
| farm | 0.0289 | football | 0.0309 | design | 0.0196 | auto | 0.0233 | question | 0.0327 |
| wheat | 0.0289 | top | 0.0258 | photograph | 0.0196 | cad | 0.0233 | year | 0.0220 |
| matter | 0.0196 | baseball | 0.02588 | urban | 0.0132 | create | 0.0233 | medical | 0.0166 |
| gap | 0.0196 | golf | 0.0157 | paint | 0.0132 | wow | 0.0122 | boomer | 0.0166 |
| gender | 0.0196 | contract | 0.0157 | photo | 0.0132 | indianapolis | 0.0122 | arab | 0.0112 |
| close | 0.0196 | coach | 0.0157 | galleria | 0.0132 | associ | 0.0122 | street | 0.0112 |
| women | 0.0196 | grow | 0.0157 | follow | 0.0132 | engine | 0.0122 | iowa | 0.0112 |
| network | 0.0102 | inca | 0.0157 | graphic | 0.0132 | manage | 0.0122 | obama | 0.0112 |

global words cause the meaning associated with each topic is not distinguish from the other topics. Thus, in our model we depend on these results which are cleaned by ODP taxonomy.

### 4.3.3 ODP-Based Topics Semantic

As we mentioned previously, the LDA model defines the meaning of each topic by presenting it as a words distribution. However, in this case, the end-users cannot observe the several trends and orientations of the selected topic since from the list of properties characterizing a word is that it is a very specific unit and mostly related to various fields and topics categories. Thus, end-users interpret and describe the meaning of the LDA results depending on their personal background and experiences. As a result of this gap, the performance of this probabilistic topic model may be decreased. To overcome this challenge, we suggest providing the end-users with a topical hierarchy to each latent topic. Thus, to achieve this purpose, we construct the concepts trees characterizing each unsupervised topic in order to detect the semantic relationships between the words of the selected latent topic resulted after applying the LDA model. The process used in this phase is that for each word in the unsupervised topic (latent topic), we generate the semantic sub-tree (fragments) from the ODP taxonomy. This human-edited Taxonomy



| Topic 1 | | ..... |
|---|---|---|
| agriculture | 0.0943 | ... |
| farmland | 0.0289 | ... |
| farm | 0.0289 | ... |
| wheat | 0.0289 | ... |
| matter | 0.0196 | ... |
| gap | 0.0196 | ... |
| gender | 0.0196 | ... |
| close | 0.0196 | ... |
| women | 0.0196 | ... |
| network | 0.0102 | ... |

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Tweet>
  - <Topic>
    - <Word>
        network
        <Probability> 0.0116 </Probability>
      - <Domaine>
            <Level0> Business </Level0>
            <Level1> Opportunities </Level1>
            <Level2> Resources and Networking </Level2>
        </Domaine>
      - <Domaine>
            <Level0> Computers </Level0>
            <Level1> Artificial Intelligence </Level1>
            <Level2> Neural Networks </Level2>
            <Level3> People </Level3>
        </Domaine>
      - <Domaine>
            <Level0> Arts </Level0>
            <Level1> Television </Level1>
            <Level2> Networks </Level2>
            <Level3> Cable </Level3>
        </Domaine>
      - <Domaine>
            <Level0> Computers </Level0>
            <Level1> Software </Level1>
            <Level2> Networking </Level2>
            <Level3> Network Management </Level3>
        </Domaine>
    </Word>
       ........
   </Topic>
    .......
 </Tweet>
```

Fig. 4.4: An example of XML file

is available for free under a Creative Commons Attribution license. It is considered as the most important and effective taxonomic directory on the web. It covers more than 5,223,457 sites filed into over than one million categories (topics). ODP's data is organized as a hierarchical structure with parent-child relationships between categories nodes. In our case, we consider only the first five categories with their top three levels. This choice of the total number of categories is because the first five categories are the more specific categories and in the other hand to simplify the model implementation, while the number of levels is selected based on the experiment results presented in [TG04a]. We repeat the same process for all latent topics generated by the probabilistic topic model LDA. Next, we construct XML file for each topic, called *"Topics-XML"*, and represent each one by fragments. Fig. 3 presents an example of this process.

### 4.3.4 N-Depth High Level Topics

The last phase in our model is inferring the semantic higher levels associated with each unsupervised topic produced by using LDA model. This is achieved by determining the domain distributions characterizing the global orientation and meaning of the selected latent topic. As we mentioned previously, for each word in the unsupervised topic, we generate its semantic fragments from the ODP taxonomy. If we consider the top ten ODP hierarchies associated with the words agriculture, farmland, farm seen previously in topic 1, then the result of this process will be the domain tree presented in Figure 4.5. In this tree, nodes represent the set of domains selected from the ODP taxonomy, while the links represent superdomain-subdomain relationships. The different colors represent the several levels of this hierarchy.

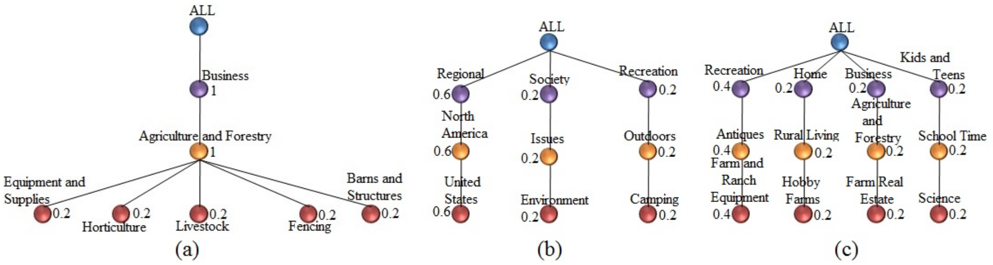


Fig. 4.5: The domain tree associated with the words agriculture, farmland and farm, respectively

The weight of each node in these fragments is computed illustrates the importance of a domain in the chosen word. It is defined by the occurrence number of a particular domain in the selected word divided by the total number of domains associated with this word.

$$P(D_{j,i}) = n_j N \tag{4.1}$$

where, $n_j$ : the occurrence number of domain $j$ in the word $i$ selected from topic $k$. $N$: the total domains occurrence for word $i$. $D_j$: The domain $j$.

Based on these sets of fragments generated by using the list of words associated with the latent topic, we construct the global tree which characterizes this topic. Then we calculate the weight of each domain node in this tree. This weight is computed based on two types of probabilities:

- The first one represents the importance of a domain in the chosen word, which is computed previously.
- The second represents the importance of the picked word $i$ in the topic $k$. This probability is obtained from the result produced previously by the LDA topic model: $P(w_i|T_k)$

Thus, to determine the probability of each domain in the selected topic, we propose the following formula:

$$P(D_j, T_k) = \sum_{i=1}^{I} n_j N \times P(w_i|T_k) \tag{4.2}$$

where, $T_k$: Topic k. I: the number of the most representative words characterizing the picked topic $T_k$.

For instance, if we consider the Topic 1 resulted after applying the ODP-Based Adapted LDA phase presented previously, the domain *"Business"* is generated to the words *"Agriculture"* and *"Farm"*. In this case, the weight of "Business" in this topic is computed by the sum of the two weights in the two words. Finally, each domain's node in the topical tree is labeled with both the name and the weight of this category. Figure 4.6 presents the resulting semantic tree for the Topic 1. In this tree, the nodes represent the categories generated by using ODP taxonomy, while the links between the nodes represent the relationships which are of the type superdomain-subdomain. For instance, the node *"Antiques"* has a superdomain *"Recreation"* and a subcategory *"Farm and ranch Equipment"*. We notice, from the tree, the different levels of this hierarchy and also each node is attached with both domain name and domain weight.

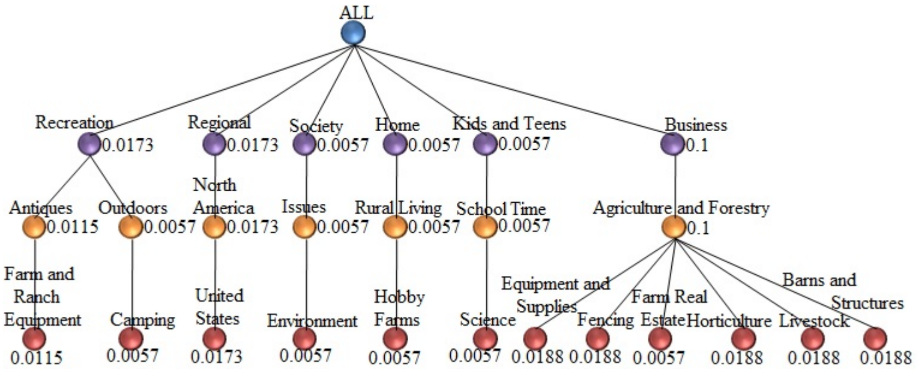


Fig. 4.6: The topical tree of Topic 1

The final step is to infer the high-level topics in the several levels. In each topic, $T_k$, and for each level, we select the node which has the maximum weight. Thus, to do this, we propose the following formula:

$$SD_{k,l} = ArgMax(\sum_{i=1}^{I} n_{j,l} N \times P(w_i|T_k)) \tag{4.3}$$

$SD_{k,l}$: selected domain for Topic $k$ in level $l$. $n_{j,l}$: the occurrence of the domain $j$ in word $i$ for level $l$.

For instance, in the semantic tree of topic 1, shown in Figure 4.6, the high level topics inferred for each level will be:

- Level 1: Business
- Level 2: Agriculture and Forestry
- Level 3: Horticulture

In our model, the topic inferred from the third level, is considered as the representative title of this topic. Therefore, the title which characterizes the topic 1 will be *"Horticulture"*.

## 4.4 Relationship between users and domains

The proposed Topic-Driven Model allows the individuals to detect the users' topics of interest on Twitter. These topics can be linked to different domains or categories. Usually, users treat a set of topics in different domains and areas with different percentages. Thus, we can define the relation between users and domains. Here, we propose to combine the importance of topic $T_k$ for the user Us from the distribution of user-topics produced by the probabilistic topic model LDA and the importance of the domain $j$ for the selected topic $T_k$ from the weight of topic-domain which are produced by our model. By the following formula (2), we can define the probability that particular user $s$ treats a specific domain $j$. Thus, it is computed by the sum of multiplication two measures:

- The probability of treating topic $T_k$ by user $s$: $P(U_s, k)$
- The weight of the category $j$ in the topic $T_k$: $W(C_j, k)$

$$P(U_s, D_j) = \sum_{k=0}^{K} P(U_{s,k}) \times P(D_{j,k} \tag{4.4}$$

Table 4.4 presents the categories weights for five users selected from our data set, as an instance of users.

in the case of user domain we create two files in the first we present domains and its identifying and in the second we present the user domain distribution as a matrix where

Table 4.4: Example of domain distribution over user

| **User - Domain** | | | | | |
|---|---|---|---|---|---|
| ***Domains*** | ***User 1*** | ***User 2*** | ***User 3*** | ***User 4*** | ***User 5*** |
| Arts | 0.003715 | 0.135456 | 0.039334 | 0.154858 | 0.001651 |
| Business | 0.097715 | 0.078215 | 0.097643 | 0.000432 | 0.097715 |
| Recreation | 0.205599 | 0.062232 | 0.039284 | 0.081719 | 0.074937 |
| Science | 0.000191 | 0.001836 | 0.001240 | 0.002050 | 0.001025 |
| Computers | 0.017724 | 0.005260 | 0.092696 | 0.006303 | 0.131343 |

the lines are users and columns are domains, for example the probability that the user1 treated the domain (Arts) is 0.0037

## 4.5 Topic Communities Extraction

Most existing works on community extraction in data mining and social network analysis areas, generally aim to define the list of users communities mainly according to the topological structure which represents the connection or the communication between users. However, in the network structure, the relationships between users do not display the dynamics between them according to their common views. For instance, although an existence of a friendship between users, we cannot detect their common opinions or interesting information. Therefore, we haven't the possibility to recommend users some relevant information such as the list of communities extracted from the social network based on the semantic relationships between users at topic level or at domain level. To overcome the list of limitations presented previously, we present in this section, the second part of our proposed approach which is a method of extraction user's communities based on the list of topics or domains extracted from user tweets content by using our proposed Topic-Driven Model for users' tweets. In our study, we suggest to consider a semantic clustering to answer some queries as *"what is the network structure grouped* *by domain?", "what is the network structure grouped by Topic?"*. In this method, we calculate the distance between users according to the common topics or domains or the both and then the results will be used to construct the communities. For instance, if there is a friendship between two users, and these users always talk about basketball in their social content, thus they will be assigned in the same community which is related to sport orientation although they mention different list of topics.

To achieve this goal, we propose the following process: First, we calculate the semantic distance between users according to three different cases: the first one represents the list of topics that users treated them within their social messages. The second one represents the set of domains discussed among the social user-generated content. The third one represents both treated topics and domains. Second, we construct the semantic graph which presents the different closeness relations between users; third, we detect the list of users communities based on the semantic graph constructed previously; fourth, we evaluate the obtained communities, in order to get the best result. In the following subsections, we will study in detail each phase in this method.

### 4.5.1 Distances between Users

Usually, users mention in their messages different topics to demonstrate the same idea or they use same topics but not necessary in the same orientation and domains. This is related to the acquire meaning from the way these topics are used or to the different personal background of users. Thus, in order to support queries like: *"what is the semantic distance between users according to the set of treated topics within the social user-generated content?", "what is the most closest or distant individuals to a specified user in a period of time depending on the list of domains they discuss?"* etc. Thus, in order to compute the closeness between users according to different axes of semantic information, we present in the suggested method, three types of distances that are illustrated as follows:

### Distances based on topics weights

Inspired by the study presented in [WLJH10b], we calculate the semantic distance between user $i$ and user $j$ as the Jensen-Shannon Divergence between the topics distributions on users presented by the following formula:

$$dist_T(i,j) = \sqrt{2 \times D_{JS}(i,j))} \tag{4.5}$$

$D_{JS}(i,j)$: the Jensen-Shannon Divergence between the two topic distributions $DT_i$ and $DT_j$. It is defined as:

$$D_{JS}(i,j) = \frac{1}{2}(D_{KL}(DT_i||M) + D_{KL}(DT_j||M)) \tag{4.6}$$

$M$: the average of the two probability distributions. $M = \frac{1}{2}(DT_i + DT_j)$ $D_{KL}$: the Kullback-Leibler Divergence which defines the divergence from distribution $Q$ to distribution $P$ as:

$$D_{KL}(P||Q) = \sum_i P(i)log(P(i)Q(i)) \tag{4.7}$$

In the following Table 4.5, we present the similarity between users according to topics, for example the similarity between user 1 and user 2 is 12.60

Table 4.5: Similarity between users according to topics

| **User - User (Topic)** | | | | | |
|---|---|---|---|---|---|
| | ***User 1*** | ***User 2*** | ***User 3*** | ***User 4*** | ***User 5*** |
| ***User 1*** | 0.000000 | 12.603026 | 6.683783 | 10.589195 | 13.140529 |
| ***User 2*** | 12.603026 | 0.000000 | 13.062113 | 13.506515 | 10.849284 |
| ***User 3*** | 6.683783 | 13.062113 | 0.000000 | 13.928627 | 8.679685 |
| ***User 4*** | 10.589195 | 13.506515 | 13.928627 | 0.000000 | 9.710527 |
| ***User 5*** | 13.140529 | 10.849284 | 8.679685 | 9.710527 | 0.000000 |

### Distances based on domains weights

In our study, as each topic is related to different domains with different probabilities, there is a possibility that two users treat the same topic but not necessarily the same orientation. For example, let us consider the two users $i, j$ which treat the same topic *"President Obama"* but they are not in the same orientation, because user $i$ talk about the politic and user $j$ talk about health. On the other side, they may be talk about different topics but with the same orientation. In this case, we propose another type of distance measure which calculates the distance between users as the Jensen-Shannon Divergence between domains distributions over users as follows:

$$dist_D(i,j) = \sqrt{2 \times D_{JS}(i,j))} \tag{4.8}$$

$D_{JS}(i,j)$: the Jensen-Shannon Divergence between the two domain distributions $DD_i$ and $DD_j$. It is defined as:

$$D_{JS}(i,j) = \frac{1}{2}(D_{KL}(DD_i||M) + D_{KL}(DD_j||M)) \tag{4.9}$$

In the following Table 4.6, we present the similarity between users according to domain, for example the similarity between user 1 and user 2 is 8.40.

Table 4.6: Similarity between users according to domains

| **User - User (Domain)** | | | | | |
|---|---|---|---|---|---|
| | ***User 1*** | ***User 2*** | ***User 3*** | ***User 4*** | ***User 5*** |
| ***User 1*** | 0.000000 | 8.407701 | 4.335734 | 9.594203 | 8.418274 |
| ***User 2*** | 8.407701 | 0.000000 | 5.909619 | 11.449679 | 5.729813 |
| ***User 3*** | 4.335734 | 5.909619 | 0.000000 | 11.641353 | 4.997698 |
| ***User 4*** | 9.594203 | 11.449679 | 11.641353 | 0.000000 | 6.813813 |
| ***User 5*** | 8.418274 | 5.729813 | 4.997698 | 6.813813 | 0.000000 |

### Distances based on topics and domains weights

In order to provide the end-users with additional information such as the closeness between users based on both the most treated topics and domains within the social user-generated content, we propose a third distance that can evaluate the proximity between individuals by selecting the two aspects, domains and topics. This semantic information allows end-users to realize many types of analysis. For instance, this distance can define the most closest users that not only they treat the topic Photography but also different domains such as Arts, Business, and Recreation. This new measure allows decreasing the distance between users who do not treat only same topics but also same domains. The distance between users in this measure is computed by using:

$$dist_{TD}(i,j) = \sqrt{2 \times D_{JS}(i,j))} \qquad (4.10)$$

$D_{JS}(i,j)$: the Jensen-Shannon Divergence between the domain distributions $DD_i$, $DD_j$ and the topic distributions $DT_i$, $DT_j$. Here, the divergence is also computed by the following formula:

$$D_{JS}(i,j) = (\frac{1}{2}(D_{KL}(DT_i||M)+D_{KL}(DT_j||M)))+(\frac{1}{2}(D_{KL}(DD_i||M)+D_{KL}(DD_j||M))) \tag{4.11}$$

Table 4.7 shows the distance between users based on topics-domains. We consider, in this table, only five users as an example of the test collection which will presented in the experimentation section. For instance, the distance between user 3 and user 5 according to topics-domains is 1.142 which is the minimum distance. That means, these two users are the closest in comparison to others.

Table 4.7: Similarity between users according to topic-domain

| **User - User (Topic-Domain)** | | | | | |
|---|---|---|---|---|---|
| | ***User 1*** | ***User 2*** | ***User 3*** | ***User 4*** | ***User 5*** |
| ***User 1*** | 0.000000 | 21.010727 | 11.019518 | 20.183398 | 21.558803 |
| ***User 2*** | 21.010727 | 0.000000 | 18.971732 | 24.956194 | 16.579097 |
| ***User 3*** | 11.019518 | 18.971732 | 0.000000 | 25.569980 | 13.677383 |
| ***User 4*** | 20.183398 | 24.956194 | 25.569980 | 0.000000 | 16.524340 |
| ***User 5*** | 21.558803 | 16.579097 | 13.677383 | 16.524340 | 0.000000 |

### 4.5.2 Graph Construction Based on Users Closeness Relations

The graphs have a great expressive and they are simple for modeling. They are based on two concepts nodes and edges. In the social network case, the nodes represent a set of social entities such as users or social organizations while the edges between nodes indicate that a direct relationship has been created during social interactions. In this approach, we create the graph which consists of nodes and edges as in the existing works but the nodes, in our method, represent the users and the edges represent the closeness between them according to selected topics, domains or both selected topic-domain and not the communication relationship as in the existing works. Thus, we can define three

|  | $Topic_1$ | $Topic_2$ | $Topic_3$ | $Topic_4$ | $Topic_5$ |
|---|---|---|---|---|---|
|  | *Arts* | *Society* | *Business* | *Regional* | *Sports* |
| *User1* | 0.01654 | 0.00150 | 0.00150 | 0.97894 | 0.00150 |
| *User2* | 0.00183 | 0.00183 | 0.99866 | 0.00183 | 0.00193 |
| *User3* | 0.95283 | 0.00087 | 0.00087 | 0.04454 | 0.00087 |
| *User4* | 0.00060 | 0.00668 | 0.00060 | 0.01276 | 0.97933 |
| *User5* | 0.07915 | 0.90694 | 0.00664 | 0.00060 | 0.00664 |

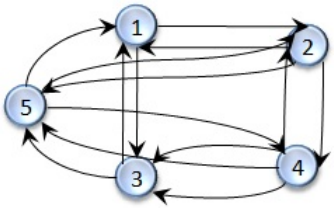


Fig. 4.7: The Topic graph

types of graphs based on the user's closeness according topics, domains and topics-domains. Topics or domains are selected either by the choice of users or we consider all topics and domains of users which are produced by our model if he/she dose not choose any topic or domain. In the topic graph, we create an edge from the user i to the user j, if the user j is the closest to the user i for the topic k, the weight of this link is calculated as the distance between them for this selected topic k. Moreover, if there is another edge from the user i to the user j for another topic, it is enough to choose the minimal distance between these two users i and j. For example, the table illustrated in 4.7 shows the topics distributions for the experimented five users. We selected, as an instance, the five most related topics for these five users. Figure 4.7 shows the corresponded graph(topic graph). We use the same method to create the other two graphs which are based on the domains closeness or topics-domains closeness.

### 4.5.3 Construct Users Communities

Based on the constructed graphs, we can extract the users communities by adapting the approach of Newman [GN02b] which is based on the divisive classification. The divisive is a top down approach which starts with all nodes as an only community and applies the division method. The algorithm [GN02b] process is based on the following steps: firstly calculate the betweenness scores for all edges in the network, secondly, find the edge with the highest score and remove it from the network, thirdly, recalculate the betweenness for all remaining edges and finally, repeat the research and the remove until get the communities.

We improve this approach to create the communities, the classic approach uses the existing graphs to calculate the betweenness by using the communications relation between users, but in our adaptation, we are based on the semantic graphs shown in the previous section. Figure 4.8 shows our adapted method which is based on divisive approach.

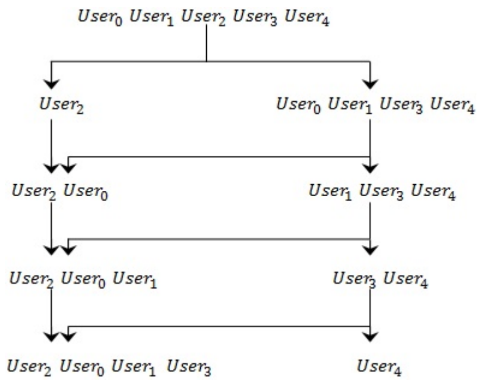


Fig. 4.8: The adapted divisive approach

### 4.5.4 Evaluation of obtained Communities

Here, we will evaluate the extracted communities based on the closeness of interests and views between social entities. The question asked, here, is *how to get the best result?* The researchers in [GN02b] 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 [ADFG07] present an extension of modularity for directed graphs. The adapted formula is:

$$Q = \frac{1}{m}\sum_{i,j\in V}(A_{ij} - \frac{k_i^{out}, k_j^{in}}{m})\delta(C_i, C_j) \tag{4.12}$$

Where $A_{ij}$ : 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. $\delta(C_i, C_j)$ equal 1 if i and j belong to the same community, and 0 otherwise.

However, in our case we have not the same classic graphs that use the existing links. Thus, we will compute the modularity by using the semantic graphs constructed previously. After modularity computation for the resulted communities over the experimented users, we note that the division of the two communities where the first one which contains user1, user3, user4 and the second which contains user0, user2 have the maximum value of modularity. This division is shown in Figure 4.8.

## 4.6 Conclusion

In this Chapter, we proposed a new approach to construct user's communities based on high-level-topics and domains which were extracted from semantic hierarchy, ODP taxonomy, to cluster users' tweets. Generally, the existing works have constructed the users communities based on the links between users. While, in our approach, the user community is based on the common topics. Thus, our method allows focusing on the relation between users according to their topics of interest. Also, we can detect the emerging topics or domains in each community. In order to construct these communities, we proposed a model to cluster users' tweets based on ODP taxonomy as an external resource to derive high level topics and the topics domains. The existing works use a generative probabilistic model, such as LDA (Latent Dirichlet Allocation), on the tweets data to identify topics for these tweets as words distribution without considering the semantic relations. The contribution of our proposed model is using a semantic hierarchy (ODP taxonomy) to assign automatically high-level-topics to each user tweet.

# Chapter 5

# On-Line Analytical Processing on Graphs generated from Social Network data

## 5.1 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 Chapter, we suggest a new data warehousing model, 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 hierarchies.

## 5.2 Introduction

Business Intelligence (BI) technologies improve dramatically an important capability to perform the business process by analyzing large amounts of data and achieving a set of brighter and clever decisions at each level of the business granularity. 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 by operating complex computation. OLAP tools can swiftly generate a visual result for complicated analytical and ad-hoc queries by using a graphical user interface (GUI). Furthermore, this result can be illustrated from different perspectives and with multiple granularities.

Over the last few years, social network sites have been quickly increasing and most particularly within the last ten years. The participation of people in these sites plays a crucial role by publishing real time information about their personal views and interests. This includes daily conversations, cultural trends and information news without any concern about writing style. This kind of communication between users is not only significant in view of the fact that it permits the diffusion of information but it is also considered as the key factor to keep track of the different reports and knowledge expressed in the social messages. 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. As an 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.

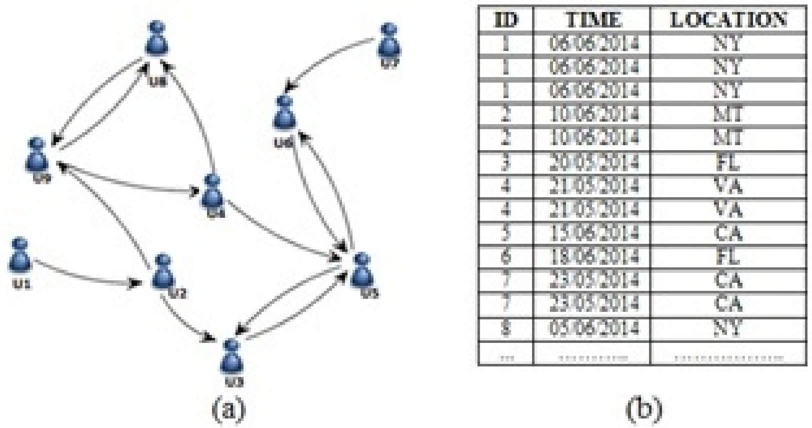
Community extraction methodologies within the social networks are receiving an increasing attention in several areas. It consists in determining a good classification of the social networks by discovering the most similar clusters who share same characteristics and properties. Typically, the community extraction methods aim to define the list of clusters mined from the networks by considering that inside each cluster (i.e., community) the density of the topological relationships between the social entities must be dense and among clusters must be sparse [GN02b]. However, the relationships in the social networks do not illustrate the quantity, quality and frequency of the semantic meaning information shared between individuals. It only reflects a straightforward communications established during a social interactions. Therefore, the study of user-generated content selected from the social network services represents a rich source of information when the end-user aims to determine the list of communities that share the same interest or attention. 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 with 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 individual data tuples. It also faces great challenges for analyzing unstructured data such as the social user-generated content. 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 [RS90b].

In this Chapter, we explore and extract the pertinent knowledge hidden in the social network services with several interesting tasks by proposing a new data warehousing model, *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 are used to: (1) illustrate the existence of relationships between the social entities, (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.

## 5.3 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 5.1, 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 5.1(a).



(a)

| ID | TIME | LOCATION |
|---|---|---|
| 1 | 06/06/2014 | NY |
| 1 | 06/06/2014 | NY |
| 1 | 06/06/2014 | NY |
| 2 | 10/06/2014 | MT |
| 2 | 10/06/2014 | MT |
| 3 | 20/05/2014 | FL |
| 4 | 21/05/2014 | VA |
| 4 | 21/05/2014 | VA |
| 5 | 15/06/2014 | CA |
| 6 | 18/06/2014 | FL |
| 7 | 23/05/2014 | CA |
| 7 | 23/05/2014 | CA |
| 8 | 05/06/2014 | NY |
| ... | .......... | ............... |

(b)

Fig. 5.1: Example of a multidimensional social network and its real-world metadata

Typically, the user-generated content in the social networks are associated with different metadata such as the geo-information which can be considered as a standard to facilitate people physically comprehend and study these networks. This can be reinforced by integrating the temporal information as well to examine and recognize the spatio-temporal representation of social networks. It can be used to analyze the spatial diffusion of information and also to explore the individual dynamic with changing relationships such as the correlation between the social relationships and the geographic location of individuals. In our approach, 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 5.1(b). The structural characteristics associated with this sample social network depicted in this figure, such as user ID (as primary key), time and location (in state) are represented as vertex attributes. The topological structure of the graph, together with the multidimensional attributes associated with vertex, forms a multidimensional network.

Table 5.1 shows real-world tweets interchanged between this set of experimental 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).

Table 5.1: Real-world tweets

| **Users** | **Example of tweets content** |
|---|---|
| U1 | ...this series is not over we need a strong 3rd period get some goals... |
| U2 | ...DeepTrawl helps webmasters find and fix site errors quickly, simply... |
| U3 | ...turn off the Busyness - which is the 1 enemy to Intimacy, get in daddy's face,.... |
| U4 | ...Right now not so much harmony in my life, working on it. Work in progress!... |
| U5 | ...polls suggest deeds surge in Va.: A surveyUSA poll shows deeds the democratic... |
| U6 | ...huge loss in home values cratered the Bay Area economy: the business in down... |
| 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... |
| U9 | ...Virtual patients helping train student nurses at Birmingham city university... |

As we can notice in this figure, the unstructured user-generated content included within these tweets is a very important source of data since it can generates explicit and implicit information about how individuals react with events, personal experiences and related ideas or opinions. However, in the same time, these social media streams are different from authored news and conventional text where the quality and the short length of the user-generated content pose a list of new issues. Thus, the traditional analysis techniques face major challenges in manipulating and treating this type of social media streams. The main cause of this limitation is that these streams are generally noisy, unstructured and most of time contain unusual spelling, emoticons and idiosyncratic abbreviations. To address this challenge, we start by aggregating all the tweets send by the same individuals into different documents, in order to get a more complete and comprehensive visualization.

Then, in order to prepare the data for the analysis tasks, we use a cleaning processing. Different techniques are used in this phase such as the syntactic transformation, the semantic enrichment and the removing stop words. In the removing stop words, we eliminate URL, numbers, noisy words, etc. The different semantic enrichment tools are used to convert the extracted data from its previous state into the desired state. In the syntactic transformation of textual data, we utilize several linguistic knowledge such as: stemming, spelling correction with WordNet dictionary, etc. Furthermore, to enhance the pertinent of the cleaning procedure, we use different semantic enrichment tools such as the WordNet dictionary and the ODP taxonomy to determine the most relevant words. The result of this treatment is represented in Table 5.2. The structural and the semantic properties combined with the sample social network displayed in the Figure 5.1 are illustrated as a tuples containing various type of data such as user ID (as primary key), time, location (in state) and word (semantic unit).

We notice that many irrelevant words are removed after the cleaning step. The structural and the semantic characteristics (i.e., pertinent words) associated with this sample social network depicted in this table, such as user ID, 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.

Table 5.2: The multidimensional attributes

| **ID** | **TIME** | **LOCATION** | **WORD** |
|---|---|---|---|
| 1 | 06/06/2014 | NY | period |
| 1 | 06/06/2014 | NY | goal |
| 2 | 10/06/2014 | NY | webmaster |
| 2 | 10/06/2014 | MT | error |
| 3 | 20/06/2014 | FL | business |
| 3 | 20/05/2014 | FL | enemy |
| 4 | 21/05/2014 | VA | life |
| 4 | 21/05/2014 | VA | work |
| 5 | 15/06/2014 | CA | poll |
| 5 | 15/06/2014 | CA | democratic |
| 6 | 18/06/2014 | FL | economy |
| 6 | 18/06/2014 | FL | business |
| 7 | 23/05/2014 | CA | innovation |
| 7 | 23/05/2014 | CA | opportunity |
| 8 | 05/06/2014 | NY | AI |
| ... | .... | ... | .... |

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. They represent the main components of a typical graph. The set of the structural data* $S$ *such as the geographic and the temporal axes, represents the selected metadata used to display better visualization of the treated date.* $U$ *illustrates the semantic data set selected from the user-generated content. It is defined by choosing the list of relevant words. The Social Graph Cube is generated by reorganization of this multidimensional network in all possible cuboids produced by using the structured data* $S$ *and the user-generated content* $U$.

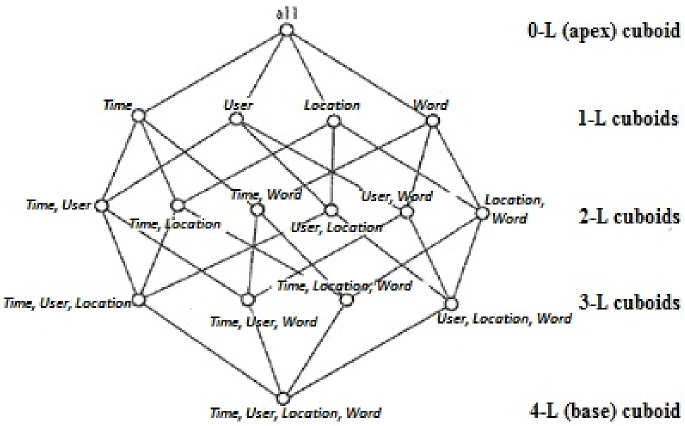


Fig. 5.2: The Social Graph Cube lattice

*The list of measures associated with each cuboid $C'$ obtained by $S$ and $U$, can either be demonstrated as 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.*

### 5.3.1 The Social Graph Cube lattice

OLAP tools display a hypercube lattice structure to navigate through the different perspectives and granularities. This structure has performed a significant role in numerous views of data cubes since it can assist to enhance performance of data cube computations. Figure 5.2 describes a *Social Graph Cube* lattice.

The cuboid of the *Social Graph Cube* are illustrated as nodes in this lattice is a (i.e. special OLAP view). They are generated from the initial multidimensional network by presenting all the possible combination of the particular dimensions. The edges in the lattice determine the parent-child relationship between two cuboids. It leads the end-users from one cuboid to another by defining the OLAP navigation links. If we consider all the available structural and semantic dimensions in a given multidimensional network $G$ such as time, user, location and word, then the number of the generated cuboids in the *Social Graph Cube* will be $2^4$ cuboids. Each cuboid illustrates a different vision of the information according to the level of the aggregation and the chosen dimensions.

In the classical determination of the cube lattice structure, the base cuboid corresponds to *C[Time, User, Location, Word]* defines the finest level of granularity, where a detailed view is displayed. However, the base cuboid in the *Social Graph Cube* designates the highest degree of summarization, in which a global vision about the interaction between all the dimensions in the lattice, while the cuboid which describes the lowest degree of summarization is called the apex cuboid and it is characteristically symbolized by all. The weighted graph associated with a descendant cuboids in *Social Graph Cube*, is more heterogeneous than the weighted graph associated with one of its ancestor cuboid which contains less topological properties and semantic details. The structure of *Social Graph Cube* permits analysts to study the original social network according to various multidimensional spaces by exploiting the generated lattice of cuboids. As a result, a set of weighted graphs with several aggregated resolutions can be mined and examined to offer the decision makers the possibility to navigate through a huge amount of complex data in order to realize some 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.

### The first level in the Social Graph Cube lattice

The number of selected dimensions for cuboids generated in the first level of the cube lattice structure is $|dim(Cuboid)| = 1$, which mean that only one type of nodes is chosen to illustrated the selected dimension in the graph. As a result, four kinds of homogeneous weighted graphs could be generated in this level; user-user graph, location-

location graph, time-time graph and finally word-word graph. They can be computed to request some complex queries that could be asked on a multidimensional network such as:

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

The semantic aspect represents a leading role to define the different relationships between the topological entities. In this level we use two different ways to evaluate the existence of the relationship between the different kinds of the topological entities (i.e., dimensions). The first one is used in the word-word graph, where the closeness between words is determined by using the well-founded semantic measure named "Normalized Google Distance (NGD)" introduced by Cilibrasi and Vitanyi's in [CV07]. 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:

$$NGD(w_i, w_k) = \frac{max\{\log f(w_i), \log f(w_k)\} - \log f(w_i, w_k)}{\log P - min\{\log f(w_i), \log f(w_k)\}} \tag{5.1}$$

$P$: 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 second one is used in the case of user-user graph, location-location graph and time-time graph. It is based on four steps. The first step is based on the aggregation and the cleaning of all the social user-generated content sent by the same user, transmitted from the same location or produced at the same time interval, respectively. In the second step , the semantic distance between users, location, time intervals is calculated respectively. The researchers in [RZGSS04] present the distance between individuals as the symmetric Kullback-Leibler divergence between the topics distribution conditioned on each of the individuals, as follows:

$$dist(i,j) = \sum_{t=1}^{T} [\theta_{it} \log \frac{\theta_{it}}{\theta_{jt}} + \theta_{jt} \log \frac{\theta_{jt}}{\theta_{it}}] \tag{5.2}$$

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

Inspired by this study, we suggest using the Kullback-Leibler divergence to compute the semantic closeness between the different topological entities (i.e., selected dimension values). However, instead of using the distribution of topics generated by the probabilistic topic model, we utilize in the *Social Graph Cube* the distribution of the most representative words computed by using the normalized  measure. The values in this measure are taken in the range $[0, 1]$. To avoid the division by zero error that may result, we utilize +0.0001 standard deviations instead of zero for the  weights. The distance between $user_s$ and $user_l$ is then represented by the following formula:

$$dist_{S_1}(user_s, user_l) = \sum_{k=0}^{K} TF\text{-}IDF_s(w_k) log \frac{TF\text{-}IDF_s(w_k)}{TF\text{-}IDF_l(w_k)} + TF\text{-}IDF_l(w_k) log \frac{TF\text{-}IDF_l(w_k)}{TF\text{-}IDF_s(w_k)} \tag{5.3}$$

Where $T$ represents the top-K most representative words characterizing each user, while $TF\text{-}IDF_s(w_k)$ and $TF\text{-}IDF_l(w_k)$ represent the  weights associated with word $w_k$ according to $user_s$ and $user_l$ respectively. The same formula can be used to calculate the distance between the different locations and time intervals.

However, the main issue of this expression is that it only considers the comparison between the same words without taking into account the degree of correlation between the various words. As an instance, if we examine different words as strings, they are determined as unrelated, however, if we define the relatedness between this set of words from the semantic point of view, we reveal that there are some value of similarity between them.

As an instance, if we consider the following users with their top 3 most representative words according to the normalized  measure.

- *User s : business 0.52, company 0.15, community 0.03*
- *User l : design 0.59, business 0.19, degree 0.02*

By using the previous formula, only the  weights associated with word business is considered to compute the distance between these two users. However, the list of other words such as company, design, community, etc, is not taken into account in this expression. This substantial source of information plays a principal role to evaluate and uncover the closeness between the different social entities even they use dissimilar concepts but in the same time, they are correlated semantically and often occur in the same domains. Based on this point of view, we suggest defining the semantic distance between users by

combining the Kullback-Leibler divergence between the  weights associated with different words distribution with another measure like the NGD measure, which can reveal the level of the closeness between the evaluated entieies. The result of the combination between the proposed formula and the NGD measure is represented in the following expression:

$$dist_{S_2}(user_s, user_l) \quad = \quad dist_{S_1}(user_s, user_l) + \frac{1}{2} \times$$

$$\sum_{k=0}^{K} \sum_{i \in \{K-k\}}^{K} TF--IDF_s(w_k) log \frac{TF\text{-}IDF_s(w_k)}{TF\text{-}IDF_l(w_i)}$$
$$+TF\text{-}IDF_l(w_i) log \frac{TF\text{-}IDF_l(w_i)}{TF\text{-}IDF_s(w_k)} + NGD(w_k, w_i)$$

In the third step, we define the weight of each social entity (i.e. nodes) in the weighted graph in order to determine the most important of influential entities in the social graph. This identification is based on several criteria such as the level of interaction. In our approach, the weight of each entity is defined based on its semantic similarity and closeness with the other entities. One of the most popular measures proposed in the social network analysis field the closeness centrality [Bav50], [Bea65], where the importance of an entity depends on its shortest path length with the other entities. It can be represented by the following measure:

$$Centrality(s) = \frac{1}{\sum_{s \neq l} distance(s, l)} \tag{5.4}$$

In our approach, we use this measure to compute the weight of the topological entities. However, instead of using the shortest path to define the weight of nodes, we determine the importance of the social entities depending on the list of mentioned words. In this case, based on the social user-generated content, the users, locations or times candidates may be more or less important.

The main goal of the fourth step is to present an effective illustraction and visualization of the abundant meaningful information between social entities by constructing the semantic weighted graphs. Only homogeneous graphs are constructed in this level of the *social graph cube* lattice to demonstrate the generated cuboid. By considering the list of selected words associated with the nine users presented in Figure 5.1, we present in Figure 5.3, the semantic weighted user-user graph.

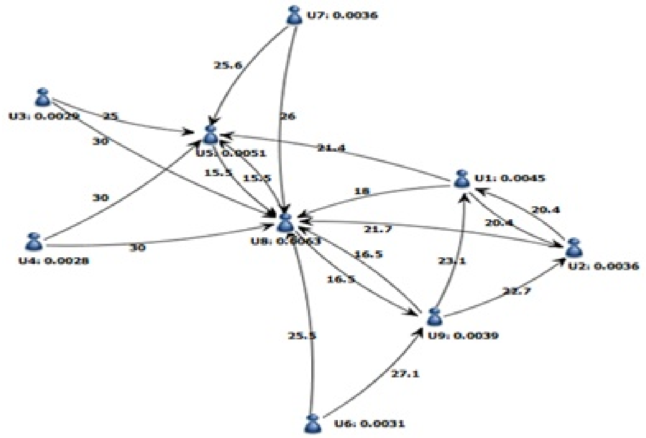


Fig. 5.3: An instance of the weighted user-user graph

Different from classical multidimensional graph representation, where the relationships between the topological candidates describe the interaction or the communication between them, our approach illustrates 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.

The process used to define the structure of the generated semantic weighted graphs is represented as follows: First, the type and the number of nodes are determined based on the selected dimension. Second, the existence of a link between two nodes $node_s$ and $node_l$ is dependent on the semantic distance between them. This means that a link is created between these two nodes only in the case that the semantic distance between them is less than or equal to a threshold parameter $\varepsilon$ which can be defined by either the end-users or by the mean associated with $node_s$. Otherwise, no link between these nodes is created.

The mean associated with $node_s$ is computed by the following formula:

$$mean_s = \frac{\sum_{l=0}^{L} dist_{s2}(node_s, node_l)}{L} \tag{5.5}$$

where, L: represents the total number of selected nodes.

From one side, the list of values obtained from the semantic distance represents the weight of each link in the constructed graph. From the other side, the list of values generated from the closeness centrality illustrates the set of weights associated with each node. In the case that we get a large number of semantic links compared to the number of nodes, we repeat this phase.

The process used to construct the weighted semantic graph associated with the dimension word is divided into four phases. They are represented as follows: First phase consists on the definition of the most pertinent words picked out by using the normalized  measure. The second phase determines the weight associated with nodes and links. The weight of links is defined by the semantic measure NGD. It is utilized to determine the correlation and the closeness between the several words. The weight of nodes is computed by adapting the closeness centrality measure. In the modified measure, we integrate in the closeness centrality, the word weights computed by using the mean of all the normalized  values associated with the selected word.

$$Centrality(word_i) = \frac{1}{\sum_{l \neq j} distance(i,j)} + Mean(TF\text{-}IDF(word_i)) \tag{5.6}$$

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 node i to the word vertex j in this weighted graph is that, the semantic distance between these two nodes is less than or equal to a threshold parameter $\varepsilon$.

### The second level in the Social Graph Cube lattice

The second level in the *Social Graph cube* lattice represents the first step to generate heterogeneous weighted graphs. As the number of dimensions $|dim(C')| = 2$, the set of cuboids produced in this level accumulate two different multidimensional spaces. As a result, two kinds of nodes are illustrated in the heterogeneous weighted graphs to figure the two selected dimensions. The list of generated cuboids can provide meaningful answers for different queries such as *What is the semantic weighted graph structure between the user U2 and the most representative words presented in Table 5.2?* or *What is the semantic closeness between the location New York and the set of experimental users presented in Figure 5.1?*. 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. To construct this type of graphs we use the following process: First, depending on the selected dimensions involved in the request of the decision makers define the number and the type of nodes in the graph. Second, aggregate and clean all the social content generated by these dimensions. Third, calculate the closeness centrality of each entity. 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 vector. As an example, in the following formula, we calculate the semantic distance between $user_s$ and $word_w$:

$$dist\_word(user_s, w) = \frac{\sum_{k=1}^{K} NGD(w_k, w)}{K} \tag{5.7}$$

where $K$: the number of top-K most representative words selected to represent users.

In the other graphs, the semantic distance is computed as illustrated in the first level. Finally, keep the relevant relationships which describe the most closely entities. In Figure 5.4, we display the answer of the preceding query.

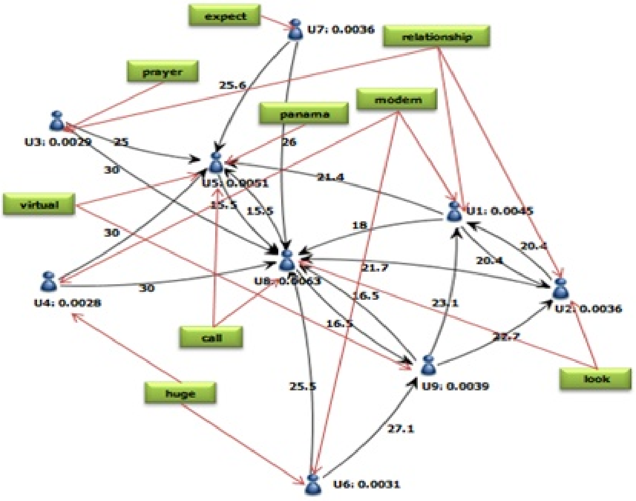


Fig. 5.4: The heterogeneous weighted user-word graph

### 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. As a result, more heterogeneous graphs are generated in this level compared to the previous level. The list of cuboids in produced in this stage is displayed in a very comprehensive visualization, where some complex queries could be analyzed and answered in a very beneficial 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.

### 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.

## 5.4 The Social Graph Cube aggregations

In the case that the decision makers are concerned with zooming into more higher-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 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. As an instance, if the topological and the semantic relationships between vertices are considered in the aggregation phase then the process used to determine the list of clusters is as follows: first, compute the topological and the semantic distances by using the length of 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 [GN02b] 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 [ADFG07] present an extension of modularity for directed graphs. The adapted formula is:

$$Q = \frac{1}{m} \sum_{i,j \in V} (A_{ij} - \frac{k_i^{out}, k_j^{in}}{m}) \delta(C_i, C_j) \tag{5.8}$$

Where $A_{ij}$ : 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. $\delta(C_i, C_j)$ equal 1 if i and j belong to the same community, and 0 otherwise.

However, we think that this measure is more suitable 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.

## 5.5 Conclusion

In this Chapter, we proposed a new data warehousing model, called *Social Graph Cube* for analyzing social networks data. Our suggested approach is designed to support OLAP-style multidimensional analysis on information-enhanced multidimensional social network. *Social Graph Cube* provides pertinent answers to analyst 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 field and OLAP operators to navigate through hierarchies. It consists in grouping network entities into different clusters according to similar interests, characteristics and views, which provides to be much more meaningful and comprehensive than classic aggregation in the traditional OLAP techniques. Different from the most community extraction methods 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. Our proposed approach permits the end-users to detect the emerging interests or orientations in each cluster. It may help to develop more effective strategies in this area.

# Chapter 6
# Social microblogging cube

## 6.1 ABSTRACT

In this Chapter, we suggest a new multidimensional model called *Microblogging Cube* to achieve OLAP techniques on unstructured microblogging data. It provides the possibility to analyze microblogs users and locations according to semantic, geographic and temporal axes. The semantic axe is defined by using the Open Directory Project (ODP) taxonomy. Different from existing classical multidimensional models, the measures in Microblogging Cube may vary depending on the aggregation levels. Further, in order to define the multiple granularities associated with microblogs users we propose a new process to extract the list of their communities.

## 6.2 Introduction

On-Line Analytical Processing presents a powerful tool for the analysis of numerous large data warehouses and industry applications. It offers analysts the ability to navigate through data collections at various granularities and from different angles in order to define exceptions and interesting parts. The multidimensional models used in OLAP provide fast answers to analyst queries. They categorize data as being either facts which describe the measures of interest for an analyst, or as being dimensions which specify different axes the data can be presented. In order to support multiple granularities, dimensions are typically organized along hierarchies of one or more Levels.

In recent years, microblogging services like Twitter have emerged as a hugely important communication utility. These services allow people to share current activities, opinions, etc and to discover interesting knowledge or news. The study of messages transmitted in microblogging sites knows a huge attention from the researchers. The Twitter site calls these messages *"tweets"*, which even though they can't exceed 140 characters, they can show the interesting domains of the user, his/her own views on a specific event, etc.

Thus, it is more and more important to analyze the information present in social network services by using OLAP technique. Unfortunately, the standard OLAP cannot handle this kind of huge complex multidimensional data arising in real-world situations. The cause of this limitation, is that the standard OLAP technology can manage hierarchies that ensure correct aggregation by enforcing summarizability in all dimensional hierarchies, which is obviously too rigid for a number of applications. The concept of summarizability in OLAP technique 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 [RS90a].

In this Chapter, we study extending decision support facilities on microblogging services by proposing a new multidimensional model, *Microblogging Cube*, for efficient and effective exploration of data contained in these services. It presents the possibility to analyze this data according to semantic, geographic and temporal axes. The concept of summarizability is taken into account in our model. Going beyond traditional OLAP operations which are based on the simple multidimensional attributes associated with the relational data, *Microblogging Cube* proposes new method to navigate through user hierarchy. It consists on the representation of users in a highest level by defining the list of their communities. Both topological and semantic relationships between users are used in the definition of communities.

To achieve the semantic analysis, we propose to use the different words selected in microblogs. However, as a word is a very specific unit and connected to different areas, we propose to go further in this study by providing analysts with additional information such as the domain distribution characterizing each user or location over a period of time. This is achieved by using the ODP taxonomy as an instance of a general ontology. In this way, we suggest a new semantic hierarchy, which consists of words in lower level and domains in higher level.

Moreover, in opposition to classical multidimensional models proposed in the literature, our *Microblogging Cube* presents two main advantages: (1) the list of selected measures depends on the hierarchical level and (2) the measures presented in this approach are well suitable for the analyses in the social networks.

## 6.3 Microblogging Cube

The proposed *Microblogging Cube* model allows analysts to study and analyze the characteristics of the unstructured text data contained in microblogs, in order to fully take advantage of all the meaningful information. This analysis is organized according to semantic, geographic and temporal axes. As a specific tweet example, we consider the following tweet: *#News Brazil stocks open higher, tracking gains in Europe - Wall Street Journal - Media http://t.co/BM80YLxM.* It is associated with different information such as *user identifier, Time, Longitude, Latitude, etc.* While the temporal specifications with their hierarchy are captured explicitly in the meta-information, the geographical and semantic specifications are defined implicitly. To specify these two axes in this model, we process as follows:

First, information about location and its associate hierarchy can be determined in different ways:

- By using the two types of attributes (longitude and latitude), which characterize the geographical coordinates when the user sends this tweet. The following information present the set of location attributes associated with geographical coordinates of previous tweet: *(Address: 320224 Foster Ave, City: "", Administrative-division: New York, Country: United States).*

- Manually filled by the user in his profile. This information can be directly usable, or used after some text transformations.

- Through the use of user's time zone.

In our study, we use the geographic database *"Geonames"*[1] to enhance the information about locations. This database is available for free under a Creative Commons Attribution license.

Second, the ODP taxonomy is applied on the content of tweets in order to extract the semantic specifications.

The proposed model is divided into three parts, which are studied in detail in the following subsection.

### 6.3.1 Microblogging Word Cube

The text contained in microblogs is a very rich data set, where most likely users aim to put significant information into this short space. However, this text is generally noisy and unstructured data. Thus, in order to clean the database, we utilize a linguistic knowledge by using different techniques such as: stemming, spelling correction, deleting noisy words with WordNet dictionary,etc.

[1] http://www.geonames.org/

In order to achieve the different possible analysis over words, such as, the most representative words of a location or a user over a period of time, the semantic distance according to words, the classification of users or location according to the sentiment they express, etc. We present in Figure 6.1 the star schema of *Microblogging Word Cube.* In this model, *Microblog_Word* is the fact type, while *Location, User, Time, Community* and *Word* are the dimension types.

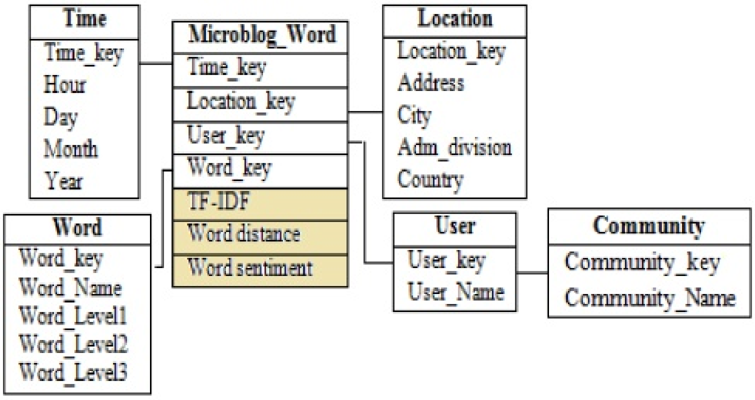


Fig. 6.1: Star Schema of a Microblogging Word Cube

The location dimension hierarchy in the tweet example seen in previous section is non-covering because the rural address: *"320224 Foster Ave"* does not show the city; the latter address is thus directly part of an administrative division. Therefore, if we compute aggregates at the city level, we will have no value for this address. Thus, the facts mapped to this address will not be considered in aggregations by administrative division through the computed aggregations by city. To overcome this type of summarizability problem, we use the algorithm presented in [PJD99] which proposes to insert an intermediate value in the city level and link this value with the address and the administrative division level.

From the several possible measures that can be used in the fact type, we choose the following ones:

**The Term Frequency Inverse Document Frequency ():** in order to determine the top-K most representative words of a location or a user, we represent the aggregation of all microblogs transmitted by each user or transmitted from each location as a normalized  vector, which reflects how important a word is to a collection of microblogs.

**The semantic word distance**: Usually, users in microblogs mention different words to describe the same idea or they use same words but not necessary in the same orientation. This is related to the acquire meaning from the way these words are used or to the different personal background of users. Thus, In order to support queries like: *"what is the semantic distance between a specified location and a selected word?", "what is the most closest or distant word to a specified user in a period of time?"*, etc, we use the well-founded measure *"Normalized Google Distance (NGD)"* introduced by Cilibrasi and Vitanyi's in [CV07]. 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 $w_i, w_k$ as follows:

$$NGD(w_i, w_k) = \frac{max\{logf(w_i), logf(w_k)\} - logf(w_i, w_k)}{logP - min\{logf(w_i), logf(w_k)\}} \tag{6.1}$$

where $P$ 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 Normalized Google Distance by using different web search engines as frequency source. Thus, the semantic word 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  vector. As an example, in the following formula, we calculate the semantic distance between *user s* and *word w*:

$$dist\_word(user_s, w) = \frac{\sum_{k=1}^{K} NGD(w_k, w)}{K} \tag{6.2}$$

where $K$: the number of top-K most representative words selected to represent $user_s$.

**Sentiment measure**: the content of microblogs is a rich source of sentiment analysis. In order to automatically define the sentiment of microblogs content, in this measure, we use the top-K most representative words characterizing each user or location as candidate sentiments words. For each word in this collection we calculate the two binary polarization weights (positive, negative) by using Naive Bayes classifiers. These weights are summed to represent the *word sentiment.* If the sum of the top-K most representative *words sentiment* is greater than 0, then this collection is regarded as positive. In the other side, if the sum is less than 0, then this collection is regarded as negative. As we can see, we use the simplest method to calculate the user's sentiment or the location's sentiment. However, our aim is to present a quick overview of the opportunities of the mapping between the sentiment analysis field and the OLAP techniques in the analysis of microblogging services.

The Location dimension is associated with the following order on its category types: $\perp_{Location}$ = Address < City < Administrative division < Country < $\top_{Location}$. In addition the Time dimension type has the following order on its category types: $\perp_{Time}$ = Hour < Day < Month < Year < $\top_{Time}$. These two dimensions are utilized for two purposes: First, they are used in the selection of user's words. For example, in the case that the analyst's interest is to study the reaction of users in a specific location towards event over a specific time period, he express his request by the following query *"What is the semantic distance between tweets transmitted over the time period t and users of location l?".* The two dimensions (time, location) are utilized to define the list of words mentioned over the period $t$ and define the list of words mentioned by users of the location $l$. Only the top-K words characterizing period $t$ and users of location $l$

are used to request previous query. Second, they are used in the grouping of users and words at a desired level of detail like the list of users' communities based on users' location. However, in the case of word dimension, classic OLAP system cannot handle the grouping of users according to the words' closeness. Thus, in the following subsubsection, we propose a new method to map the community extraction area to OLAP query processing in order to define the aggregation of users according to word dimension.

### Word Community Extraction

Traditional data cube and OLAP techniques use the structured attributes dimensions to support the navigation along any dimension hierarchies. However, they face a great challenge to answer some complex queries of the form *"What is the list of communities extracted from the social network based on the semantic relationships at word level?"*. In our model, we propose a new method to navigate along user hierarchy by introducing the granularity community which represents the highest level in this hierarchy. This method consists on the aggregation of users into a set of clusters according to the topological structure of the social network and the semantic relationship between users at word level. To achieve this goal, we propose the following process:

In our study, the distance between a pair of users ($user_s$ and $user_l$) is given by:

$$dist(user_s, user_l) = \gamma \times dist_S(user_s, user_l) + \theta \times dist_T(user_s, user_l) \tag{6.3}$$

where $dist_S$ and $dist_T$ represent respectively the semantic and the topological distance measures, while $\gamma$ and $\theta$ are the weighting factors of each distance specified by the user.

Inspired by the distance defined in [RZGSS04], we propose to calculate the semantic distance between users as the Kullback-Leibler divergence between the  of the top-K most representative words characterizing each of them. This distance is then represented by the following formula:

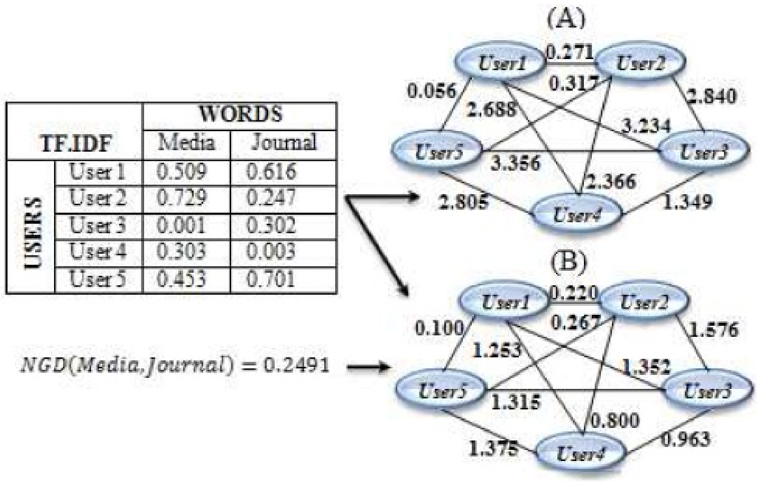
$$dist_{S_1}(user_s, user_l) = \sum_{k=0}^{K} \left( TF\text{-}IDF_s(w_k) log \frac{TF\text{-}IDF_s(w_k)}{TF\text{-}IDF_l(w_k)} \right) + \left( TF\text{-}IDF_l(w_k) log \frac{TF\text{-}IDF_l(w_k)}{TF\text{-}IDF_s(w_k)} \right)$$

However, this expression compares only the weights associated with the same words without considering the semantic relatedness among the different words. As the example illustrated in Figure 6.2 (A), this formula calculates the distance between users according to words: *Media, journal.* It does not take into account the semantic relationship between these two words which could reveal that they are linked semantically and often appear in the same fields. From this idea, we propose to combine this equation with another measure like NGD distance. The result of this combination is represented in formula 7.3. It is divided into two parts: The first part represents the distance between the selected users based on the weight of same words. This part is multiplied by the weighting factor $\alpha$. The Second part represents the distance between users based on the weight and the semantic relatedness of different words. It is multiplied by the weighting factors $\beta$. These two weighting factors ($\alpha$, $\beta$) indicate the relative importance or impact of each part in formula 7.3. They are taken in the range [0, 1].

$$dist_{S_2}(user_s, user_l) = \alpha \times dist_{S_1}(user_s, user_l) + \beta \times \frac{1}{2} \times \sum_{k=0}^{K} \sum_{i \in \{K-k\}}^{K} \left( TF\text{-}IDF_s(w_k) log \frac{TF\text{-}IDF_s(w_k)}{TF\text{-}IDF_l(w_i)} \right) + \left( TF\text{-}IDF_l(w_i) log \frac{TF\text{-}IDF_l(w_i)}{TF\text{-}IDF_s(w_k)} + NGD(w_k, w_i) \right)$$

Figure 6.2 (A) and (B) present the distances between five users depending on equation 7.3 and equation 7.3, respectively. Only two words *(media, journal)* are considered in the calculation of these equations. In Figure 6.2 (B), we give more importance to alpha weight by assigning it the value *0.65* compared with beta weight which is equal to *0.35*.

As we can see, even *user3* and *user4* use the same words with different values, the fact that the NGD between *media* and *journal* show a high degree of relatedness, the distance between these two users is decreased from *1.348* in Figure 6.2 (A) to *0.963* in Figure 6.2 (B).



| TF.IDF | | WORDS | |
|---|---|---|---|
| | | Media | Journal |
| USERS | User 1 | 0.509 | 0.616 |
| | User 2 | 0.729 | 0.247 |
| | User 3 | 0.001 | 0.302 |
| | User 4 | 0.303 | 0.003 |
| | User 5 | 0.453 | 0.701 |

Fig. 6.2: Example of the semantic word closeness

The topological distance $dist_T(user_s, user_l)$ is computed by the length of shortest path between $user_s$ and $user_l$.

Second, the agglomerative strategy is utilized to extract the users' communities by using the content and topological distance computed previously. The agglomerative is a bottom-up approach that uses nodes as communities and combines these nodes according to distances, until get the dendrogram which represents the visualization of the nodes coalescing in communities.

Third, the extracted communities are evaluated to get the best result. Newman in [GN02b] proposes the popular modularity measure $Q$ which compares the number of edges within community with the expected number in an equivalent network with edges placed at random. Moreover, the authors in [ADFG07] present an extension of modularity for directed graphs. The adapted formula is:

$$Q = \frac{1}{m} \sum_{s,l \in V} (A_{sl} - \frac{k_s^{out} k_l^{in}}{m}) \delta(C_s, C_l) \tag{6.4}$$

where $A_{sl}$: the elements of the adjacency matrix of graph *G(E, V)*, $E$: edge, $V$: vertex. $k_l$, $k_s$: the in-degree and out-degree of nodes $l$,$s$. $m$: the number of edges. $\delta(C_s, C_l)$ equal 1 if $s$ and $l$ belong to the same community, and 0 otherwise.

However, we think that this measure is more suitable for community extraction methods based on the topological properties. It computes only the link between users contained in classic graph without considering the semantic relationship between them. Thus, we propose to construct another type of graph where the nodes represent the users and the links represent the semantic relationship described by the closeness between users according to selected words. In this graph, we create a link from the *user s* to the *user l*, if the distance $dist_S(user_s, user_l)$ is less than or equal to a threshold parameter $\varepsilon$ which is determined experimentally, the weight of this link is calculated as the distance between them.

Figure 6.3 (A) and (B) present an example of semantic graph generated from Figure 6.2 (B) with $\varepsilon = 1$ and classic graphs respectively. The nodes in these graphs represent the five experimented users while the links in Figure 6.3 (A) represent the semantic closeness between them according to a list of words and in Figure 6.3 (B) represent the communication relationships.

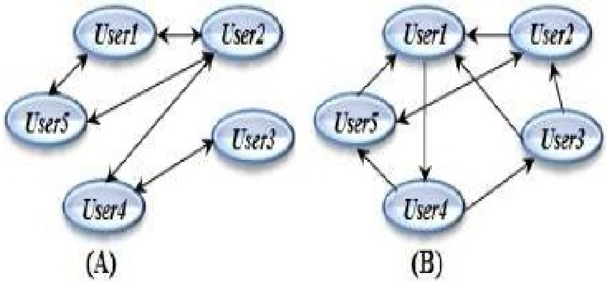


Fig. 6.3: The two types of graphs semantic and classic

Finally, to evaluate the list of communities detected in our method, we propose to adapt the modularity measure as the average of the modularity obtained from the two types of graphs (semantic and classic), as follows:

$$Q = 1/2 \times \left( \gamma \times \frac{1}{m_S} \sum_{s,s \in V_S} (AS_{s,l} - \frac{kS_s^{out} kS_l^{in}}{m_S}) \delta(C_s, C_s) \right) + \\ \theta \times \left( \frac{1}{m_T} \sum_{s,l \in V_T} (AT_{s,l} - \frac{kT_s^{out} kT_l^{in}}{m_T}) \delta(C_s, C_l) \right)$$

Based on this equation, we find that the best result that can be obtained is to define *user1, user2 and user5* as a *community1* and *user3, user4* as *community2.*

### 6.3.2 Microblogging Domain-Word Cube

In order to analyze the unstructured data presented in microblogs, word represents a very specific unit. Thus, analysts cannot automatically generate the overall areas treated by each user or location; they understand the results according to their personal background. To overcome this challenge and to derive much potential information such as, the different treated domains, we use in this work the ODP as an external resource. This human-edited Taxonomy is available for free under a Creative Commons Attribution license. It is considered as the most important and effective taxonomic directory on the web. It covers more than *5,223,457* sites filed into over than one million categories (topics). ODP's data is organized as a hierarchical structure with parent-child relationships between categories nodes.

In Chapter 4, we have used the ODP taxonomy to derive the set of categories associated with each topic detected by using the topic model Latent Dirichlet Allocation (LDA). In this work, we define the domains distribution characterizing each user or location, by identifying for each word selected from the top-K most representative words characterizing a user or location, its first top-J categories with their top three levels from the ODP taxonomy. The number of levels is selected based on the experiment results presented in [TG04b]. We repeat the same process for all selected users or locations, in order to construct their sematic hierarchies. If we consider the top ten ODP hierarchies associated with the word *journal* seen in previous tweet, then the result of this process will be the domain tree presented in Figure 6.4. In this tree, nodes represent the set of domains selected from the ODP taxonomy, while the links represent superdomain-subdomain relationships. The different colors represent the several levels of this hierarchy.

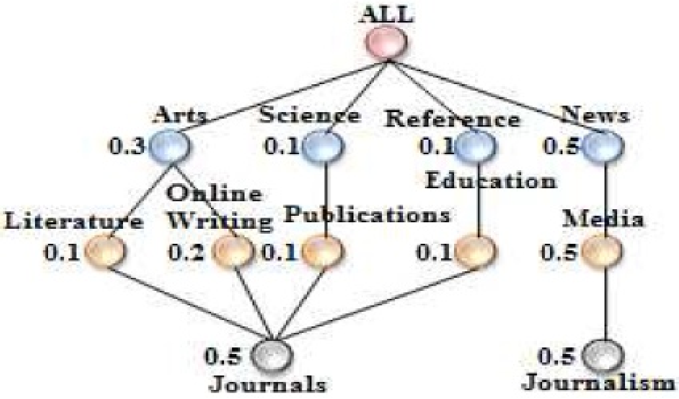


Fig. 6.4: The domain tree associated with the word *journal*

Traditional OLAP systems only enable strict hierarchies where every lower-level value belongs to at most one single higher-level value. However, as we can see in Figure 6.5, the words in the lower level have several parents in the domain level. Therefore, if we compute the total count of words at the higher level, we will have the double counting problem, which causes an incorrect result.

To overcome this limitation, we use the MakeStrict algorithm presented in [PJD99]. The basic idea of this algorithm is that it aggregates for each word his set of domains parent in each level into one fused value.

This semantic hierarchy allows users to realize many types of analysis by selecting domains or words. For example, the top-K most representative words characterize a

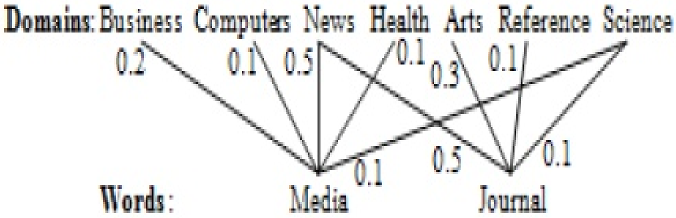

Domains: Business Computers News Health Arts Reference Science
0.2
0.1
0.5
0.1
0.3
0.1
0.1
0.5
0.1
Words:
Media
Journal

Fig. 6.5: An example of word-Domain hierarchy

domain discussed by users in a location for a certain period.

We present in Figure 6.6 the star schema of a *Microblogging Domain Word Cube.* In this model, *Domain_Word* is the fact type, and *Location, User, Time, Community, Word* and *Domain* are the dimension types.

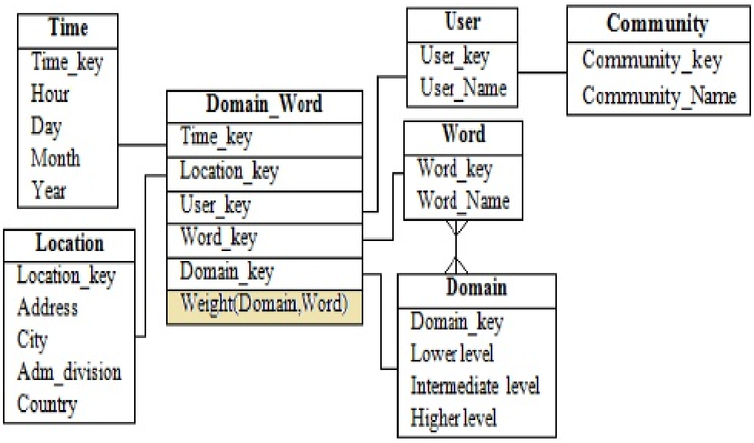


Fig. 6.6: Star Schema of a Microblogging Domain-Word Cube

Different from *Microblogging Word Cube* where the list of selected measures considers only words, the selected measure in this model is based on both words and domains. It calculates the weight of a domain according to a selected word in order to define the importance of this domain according to this word. It is represented by the following expression:

$$Weight(domain_j, word_k) = n_{j,k}/N \tag{6.5}$$

where, $n_{j,k}$: occurrence number of *domain j* with *word k*. $N$: The total number of domains associated with *word k*.

To answer a queries like, *"what is the weight of domain j associated with word k according to user s?"*, we use the following formula:

$$Weight_s(domain_j, word_k) = TF\text{-}IDF_s(w_k) \times n_{j,k}/N \tag{6.6}$$

### 6.3.3 Microblogging Domain Cube

The domain dimension presents a great interest to have comprehensive and global vision about the data contained in microblogs. From the list of possible analysis with this semantic dimension, we can cite: the top-K interesting domains, the domain most representative of a location or a user for a certain period, the distance and the closeness between users or locations according to their treated domains, etc.

As described in the previous subsection, we define the domain dimension by matching between the top-K most representative words of each user or location and the ODP taxonomy.

We present in Figure 6.7 the star schema of *Microblogging Domain Cube*. In this model, *Microblog_Domain* is the fact type, and *Location, User, Time, Community* and *Domain* are the dimension types.

Different from the *Microblogging Word Cube* where we have three types of measures, the basic measure in *Microblogging Domain Cube* is the domain weight. It is based on the matching between the  characterizes the relevance of selected words for a user or a location and the weights values characterize the importance relationship between the selected words and selected Domains.

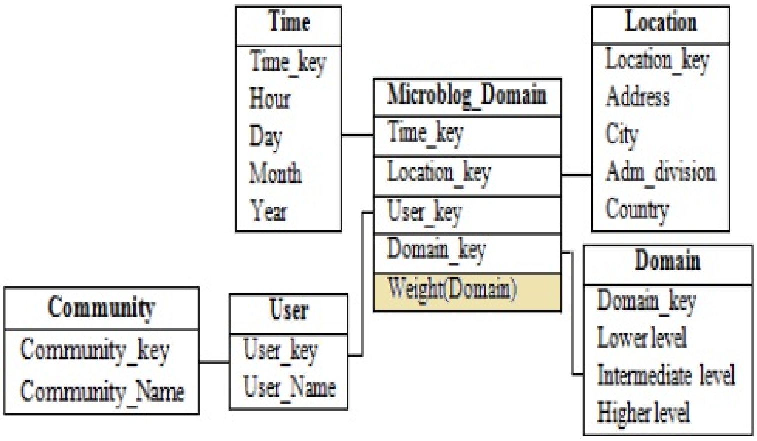
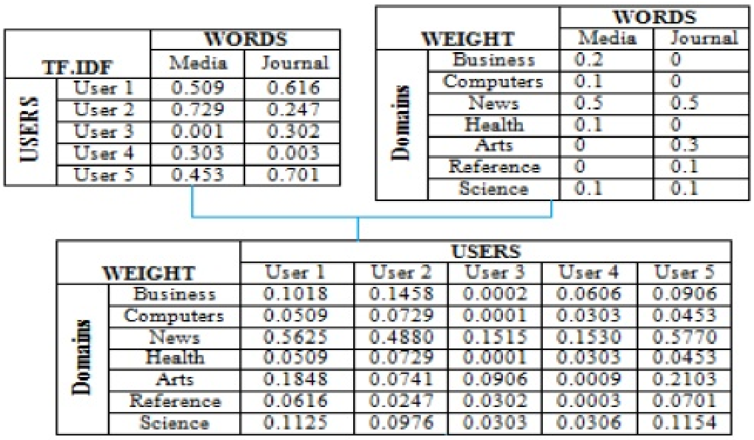


Fig. 6.7: Star Schema of a Microblogging Domain Cube

The following formula calculates the weight of *domain j* according to *user s*. It can also be used to calculate the weight of this domain according to locations.

$$Weight(domain_j, user_s) = \sum_{k=0}^{K} TF\text{-}IDF_s(w_k) \times Weight(domain_j, w_k) \tag{6.7}$$

In Figure 6.8, we present an example of the top-level domains generated according to the five experimental users presented previously. Each domain is associated with a weight value that describes the importance of this domain for the selected user.



| TF.IDF | | WORDS | |
|---|---|---|---|
| | | Media | Journal |
| USERS | User 1 | 0.509 | 0.616 |
| | User 2 | 0.729 | 0.247 |
| | User 3 | 0.001 | 0.302 |
| | User 4 | 0.303 | 0.003 |
| | User 5 | 0.453 | 0.701 |

| WEIGHT | | WORDS | |
|---|---|---|---|
| | | Media | Journal |
| Domains | Business | 0.2 | 0 |
| | Computers | 0.1 | 0 |
| | News | 0.5 | 0.5 |
| | Health | 0.1 | 0 |
| | Arts | 0 | 0.3 |
| | Reference | 0 | 0.1 |
| | Science | 0.1 | 0.1 |

| WEIGHT | | USERS | | | | |
|---|---|---|---|---|---|---|
| | | User 1 | User 2 | User 3 | User 4 | User 5 |
| Domains | Business | 0.1018 | 0.1458 | 0.0002 | 0.0606 | 0.0906 |
| | Computers | 0.0509 | 0.0729 | 0.0001 | 0.0303 | 0.0453 |
| | News | 0.5625 | 0.4880 | 0.1515 | 0.1530 | 0.5770 |
| | Health | 0.0509 | 0.0729 | 0.0001 | 0.0303 | 0.0453 |
| | Arts | 0.1848 | 0.0741 | 0.0906 | 0.0009 | 0.2103 |
| | Reference | 0.0616 | 0.0247 | 0.0302 | 0.0003 | 0.0701 |
| | Science | 0.1125 | 0.0976 | 0.0303 | 0.0306 | 0.1154 |

Fig. 6.8: An example of the domains weights computation

### Domain Community Extraction

The Community dimension is defined according to much potential information represented by the similar treated domains and the communication relationships between users. Thus, we use the same method presented previously. However, in this cube we replace the semantic word distance by a new expression, describes the distance between users according to treated domain. It is represented as follows:

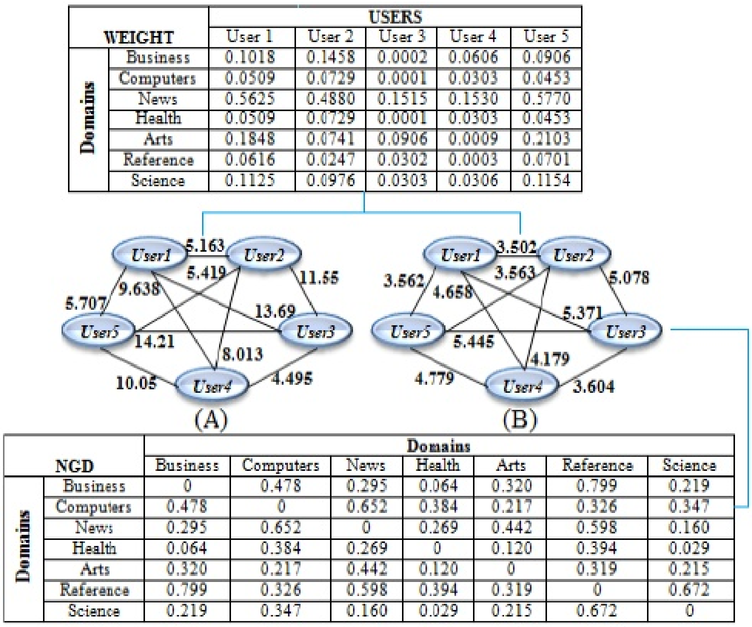
$$dist_s(user_s, user_l) = \alpha \times \left( \sum_{j=0}^{J} Weight_s(j) log \frac{Weight_s(j)}{Weight_l(j)} + Weight_l(j) log \frac{Weight_l(j)}{Weight_s(j)} \right)$$
$$+\beta \times \frac{1}{2} \times \sum_{j=0}^{J} \sum_{q \in \{J-j\}}^{J} \left( Weight_s(j) log \frac{Weight_s(j)}{Weight_l(q)} \right) +$$
$$\left( Weight_l(q) log \frac{Weight_l(q)}{Weight_s(j)} + NGD(D(j), D(q)) \right)$$

where, $J$: top-J most representative domains. $D(j), D(q)$: domain $j$ and domain $q$, respectively.

Figure 6.9 (A) and (B) present the semantic domain closeness between the five experimental users with and without the use of NGD distance in the equation 11, respectively. In this example, we assign alpha the value 0.65 and beta the value 0.35.

In the phases of community construction and evaluation, we use the same process presented in subsection 3.1.1.

Figure 6.10 shows an example of *Microblogging Cube*. The left cuboid presents the importance of a list of words according to some users in different locations, while the right cuboid presents the importance of a list of domains according to some communities in different locations. It is considered as the aggregation of the left cuboid.



| WEIGHT | | USERS | | | | |
|---|---|---|---|---|---|---|
| | | User 1 | User 2 | User 3 | User 4 | User 5 |
| Domains | Business | 0.1018 | 0.1458 | 0.0002 | 0.0606 | 0.0906 |
| | Computers | 0.0509 | 0.0729 | 0.0001 | 0.0303 | 0.0453 |
| | News | 0.5625 | 0.4880 | 0.1515 | 0.1530 | 0.5770 |
| | Health | 0.0509 | 0.0729 | 0.0001 | 0.0303 | 0.0453 |
| | Arts | 0.1848 | 0.0741 | 0.0906 | 0.0009 | 0.2103 |
| | Reference | 0.0616 | 0.0247 | 0.0302 | 0.0003 | 0.0701 |
| | Science | 0.1125 | 0.0976 | 0.0303 | 0.0306 | 0.1154 |

| NGD | | Domains | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Business | Computers | News | Health | Arts | Reference | Science |
| Domains | Business | 0 | 0.478 | 0.295 | 0.064 | 0.320 | 0.799 | 0.219 |
| | Computers | 0.478 | 0 | 0.652 | 0.384 | 0.217 | 0.326 | 0.347 |
| | News | 0.295 | 0.652 | 0 | 0.269 | 0.442 | 0.598 | 0.160 |
| | Health | 0.064 | 0.384 | 0.269 | 0 | 0.120 | 0.394 | 0.029 |
| | Arts | 0.320 | 0.217 | 0.442 | 0.120 | 0 | 0.319 | 0.215 |
| | Reference | 0.799 | 0.326 | 0.598 | 0.394 | 0.319 | 0 | 0.672 |
| | Science | 0.219 | 0.347 | 0.160 | 0.029 | 0.215 | 0.672 | 0 |

Fig. 6.9: An example of the semantic domain closeness

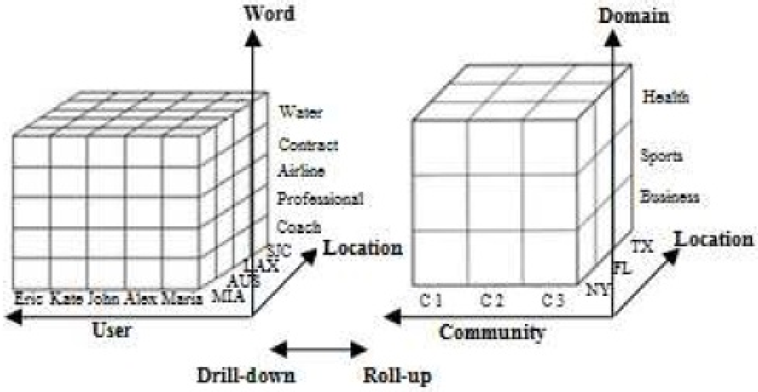


Fig. 6.10: An example of Microblogging Cube

## 6.4 Conclusion

In this Chapter, we proposed a new multidimensional data model, namely *Microblogging Cube*. This model presents the ability to analyze and understand the information behind microblogs, according to semantic, geographic and temporal axes by taking into account the concept of summarizability. Our proposed model is divided into three parts; each of them is intended to present an analysis according to specific views. In this study, we proposed a new semantic hierarchy represented by the list of words in the lowest level and the different treated domains in the highest level. The list of treated domains is defined by using the ODP taxonomy. Further, *Microblogging Cube* proposes new method to navigate through user hierarchy. Different from existing models, the measures in our model may vary depending on the level of aggregation.

# Chapter 7

# Community Cube: A Semantic Framework for Analyzing Social Network Data

## 7.1 ABSTRACT

In this Chapter we study the use of data warehousing and OLAP technologies with such new multidimensional social network by proposing a Community Cube architecture. Our design aims to support OLAP-style analysis on information-enhanced multidimensional social network data for efficient information extraction. Going beyond traditional OLAP operations, which are based on simple multidimensional attributes associated with the relational data, our Community Cube architecture proposes a new method that combines data mining and OLAP tools to navigate through the user hierarchy. This hierarchy consists on the representation of users at the highest level by defining the granularity community. Different from most of the community extraction methods focused on the connections between users, our proposed method is based on the aggregation of users into a set of clusters according to the topological structure and the semantic relationship between them. Besides traditional OLAP queries, our approach introduces a new class of queries, which we named *NetCuboid.* These queries take into account the multidimensional attributes associated with networks entities, the user-generated content, and the topological structure of the networks.

## 7.2 Introduction

Business Intelligence (BI) represents a set of technologies and systems that play a major role in delivering the right information mined from large amounts of data to decision-making and planning systems. On-Line Analytical Processing (OLAP), a major technology for BI, represents a powerful and flexible tool for mining and performing deep analyses on large amounts of data. In OLAP, a data cube represents a way to organize data in a multidimensional model to support user defined data views. This data cube is based on two main concepts: the concept of 'fact', or 'measure' (which describes the events of interest for an analyst), and the concept of 'dimensions' (which specifies different axes the data can be viewed and presented). To facilitate navigating the cube, the dimension values are typically organized along hierarchies of one or more levels. Using operations such as roll-up, drill-down, slice-and-dice and pivot, the result of on-line analysis is viewed as sets of points in a multidimensional space, which enables analysts to quickly study and navigate through the data from different perspectives and with multiple granularities.

Recently, social network, like Twitter and Facebook, have attracted millions of users. They allow people to share photos, videos and messages, thus enabling them to exchange information about their private lives, news, opinions, etc. The combination of large amount of users and the sheer volume of information provided by each one of them has led to the accumulation of enormous amounts of both structured and unstructured data. Typically, the topological structures of the social networks can be modeled as large underlying graphs, in which the vertices representing entities and edges depicting relationship between these entities [AW10]. Multidimensional attributes are usually defined and assigned to the network entities, forming multidimensional networks. Community extraction methodologies for social networks are increasingly raising interest from researchers, as they allow for the identification of latent semantic clusters by grouping users that share same characteristics and properties (i.e.: interests). Existing research on community extraction aims, mainly, to infer a list of groups according to the topological structure. However, in resulting user graphs, the relations do not present the dynamics between users according to their common views or preferences. For instance, although a set of users may have a friendship relationship, we cannot extract from this relation information about what interests they have in common. For this reason, research on user-generated content selected from the social network sites has attracted significant attention in recent years. We think that the relation between users, as defined in classic methods such as [For10] and [DMn04] is not enough when aiming to identify user groups with similar interests.

By applying OLAP techniques to analyze communities extracted from social network, we are able to flexibly explore the communities' data and get a fresh and timely perception of the online social channel. Unfortunately, standard OLAP implementations cannot handle the kind of complex multidimensional data generated from these communities, due to traditional OLAP technology not considering different types of relationships between individual data tuples and lackluster capabilities when analyzing unstructured data such as free natural language transcripts.

Studies on modern networks have been carried out for decades [New10], and as a result, multiple applications and algorithms have been proposed to aid decision makers using relation databases ([CD97], [GCB$^+$07]), and even OLAP queries for multidimensional networks ([ZHPL12], [ZLXH11a], [ZCY$^+$08a]). But none of the existing works have considered both the unstructured user-generated content and the topological structure into account in the multidimensional social network scenario. Moreover, none of such contributions have considered the combined use of OLAP-style multidimensional analysis and community extraction methodologies, in order to provide the analysts with an unprecedently comprehensive view of communities' data.

Our study focuses on extending the decision support facilities for social network services by using a dynamic data cubing and mining system, called Community Cube, for efficient and effective exploration the data contained in these services. Going beyond traditional data cubes, which are based on simple multidimensional attributes associated with the relational data, the Community Cube is an advanced data cube architecture that allows decision makers to summarize and navigate through the user hierarchy by proposing a new community extraction method for efficient query and analysis. This hierarchy consists on the representation of users at the highest level by defining the list of communities they belong to. Both topological and semantic relationships among users are combined into one integrated framework for the definition of communities.

We propose to use the set of words selected from user-generated content to compute the semantic relationships among users. However, as a word is a very specific unit and connected to different areas, we propose to go further in this study by using the Open Directory Project (ODP) as an external resource.

Additionally, besides traditional OLAP queries, our approach introduces a new type of queries, called NetCuboid. These queries take into account the multidimensional attributes associated with networks entities, the user-generated content and the topological structure of the networks. To the best of our knowledge, this type of queries has not been studied before. An instance of NetCuboid query would be *"What is the list of communities extracted from the social network based on the semantic and topological relationships at time level?"*. This type of queries breaks the boundaries created in the classic OLAP-style in that it straddles different aggregations simultaneously.

The contributions of our approach can be summarized as follows:

1. We propose a dynamic data cubing and mining framework, called Community

Cube, to extend decision support services on social data with complex features. Community Cube allows analysts to interactively analyze and navigate structured data together with network structure and unstructured text according to different perspectives and with multiple granularities.

2. We provision Community Cube with advanced unstructured text analysis capabilities for defining aggregations using new clustering methods. This is a significantly change over traditional data cubes where OLAP aggregations are directly computed by using the list of attributes associated with dimensions.

3. We present new solutions to answer different OLAP queries in the multidimensional social network scenario. Besides traditional OLAP queries, our approach introduces a new class of queries, NetCuboid, which takes into account the multidimensional attributes associated with network entities, the user-generated content and the topological structure of the networks. We show that NetCuboid is effective and useful to navigate through unstructured text and topological structures.

4. We evaluate our approach with real data, collected by crawling public tweets and network structures of following-follower relationships. The experimental results demonstrate the power and efficiency of our proposed Community Cube framework.

## 7.3 Community Cube Architectural Overview

The system architecture we propose to support information summarization and querying to social network data is depicted in Figure 2.1.

The Community Cube architecture presents two main advantages: (1) it uses a new multidimensional model to provide OLAP functionality on unstructured text data con-

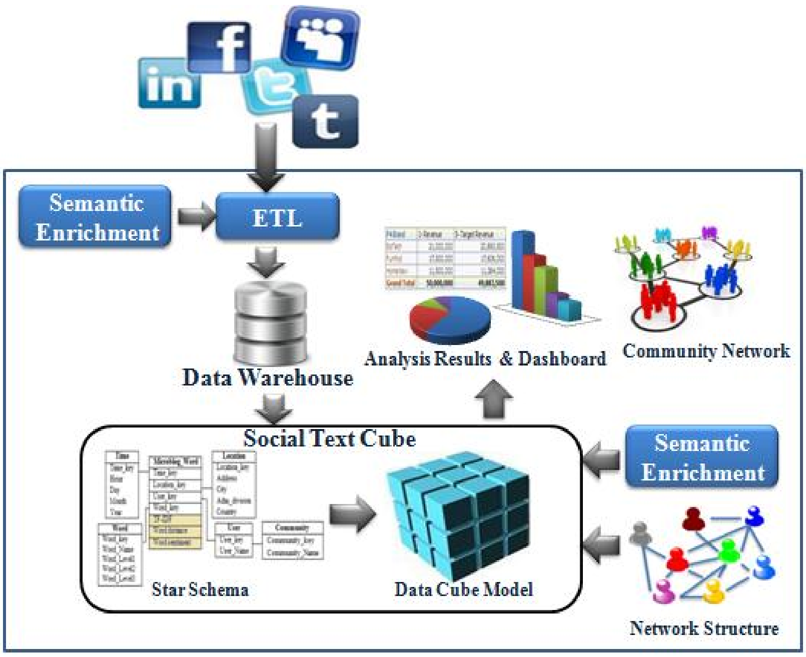


Fig. 7.1: Community Cube architecture

tained in social services, and (2) it improves OLAP hierarchies with a new levels by suggesting new method that combines data mining clustering and OLAP operators. In the following subsection, we describe each component in this architecture.

### 7.3.1 Extract-Transform-Load process

A sample of actual tweets from a random set of users is shown in Table 7.1. Each tweet selected from the Twitter site is associated with additional data, such as user identifier, time, longitude, latitude, etc. As we can see, the text contained in these tweets is a very rich data set, where users put significant amounts of information about their current activities or opinions. However, this text is generally noisy and unstructured data, and therefore, a preprocessing step is required. The Extract-Transform-Load (ETL) process is one of the backbone elements of data warehousing. It focuses on the extraction of data from specific source databases and stores the results into the data warehouse. ETL comprises to three separate phases: extraction, transformation, and loading, all of which are next described.

Table 7.1: Real-world tweets

| **Users** | **Example of tweets content** |
|---|---|
| user0 | ..From environment to economics and world affairs.. |
| user1 | ..The Product Creation System are such powerful.. |
| user2 | ..My company is the only marketing open in the.. |
| user3 | ..The Today show is the best morning.. |
| user4 | ..Just heard some devastating events, my prayers.. |
| user5 | ..Four questions drive IBM innovation.. |
| user6 | ..Waiting on our new Business cards to.. |
| user7 | ..The year when the Chinese economy will.. |

First, in the extraction phase we pull data from different social networking services by using available tools and techniques. The main tool in the most popular social networks such as Twitter and Facebook is the provided APIs (Application Programming Interface). These APIs allow for data retrieval in different formats, which usually includes an Extensible Markup Language (XML) or JavaScript Object Notation (JSON). In this phase, we clean our dataset by removing stop words, URLs, noisy words, etc.

In the transformation phase we convert the extracted data from its previous state into the target state by using different semantic enrichment tools. To achieve the syntactic transformation of textual data, we use linguistic knowledge based on different techniques such as stemming, spelling correction with WordNet dictionary, etc. While the temporal specifications are captured explicitly in the meta-information of social services, the geographical specifications can be defined implicitly. In our study we use the geographic database *"Geonames"*[1], to enhance the information about locations. This database is available for free under a Creative Commons Attribution license.

Finally, in the load phase we transmit and write the obtained data to the target data

[1] http://www.geonames.org/

warehouse for analysis.

### 7.3.2 Social Text Cube

The data cube architecture offers analysts the ability to analyze quickly and navigate through data collections at multiple granularities and from different visions in order to define exceptions and interesting parts. It categorizes data as being either facts, which describe the measures of interest for an analyst, or as being dimensions, which specify different axes the data can be presented. In order to facilitate analysis and visualization in Social Text Cube, we start by presenting our proposed multidimensional model. Then, and based on this model, we design the data cube architecture. To provide Social Text Cube with accurate, actionable and fast answers to 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.

#### Star schema of Social Text Cube

According to [PJ99] the basic multidimensional fact schema is a two-tuple *S= (F, D)*, where $F$ is a fact type and $D = T_i; i = 1...n$ is its corresponding dimension types. The dimension type $T$ is divided into a set of categories $C = C_j; j = 1...k$. Each category represents the values associated with a level of granularity. $T$ is also presented with a partial order $(\leq_t)$ on the $C_j$ 's, with $T_T \in C$ and $\perp_T \in C$ being the higher and finer level of the ordering, respectively. We note $e \in D$ where $D$ is a dimension of a type $T$, to indicate that $e$ is a dimensional value of $D$, if there is a category $Cj \subseteq D$ such that $e \in \cup_j C_j$. In [PJD99], the authors present a multidimensional object (MO) as a four-tuple $M = (S, F, D, R)$, where $S = (F, D = T_i)$ is the fact schema, $F$ is a set of facts $f, D = D_i; i = 1...n$ is a set of dimensions, and $R = R_i; i = 1...n$ is a set of fact-dimension relations.

The star schema is a kind of multidimensional model for the OLAP data cube. In this schema there is one or more fact tables connected to multiple dimension tables to allow for the different possible analysis over words contained in our dataset, such as the most representative words of a location or a user over a period of time, the semantic distance between a specific word and a list of users, the classification of users, or location according to the sentiment they express, etc. We present in Figure 7.2 the star schema of Social Text Cube as an extension of the classical multidimensional model. In this model, Microblog_Word is the fact type, while Location, User, Time, Community and Word are the dimension types.

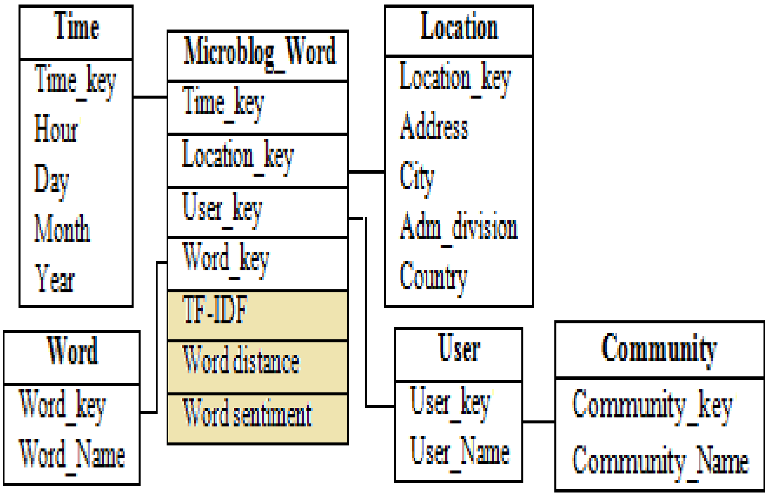


Fig. 7.2: Star Schema of a Social Text Cube

The traditional OLAP Data Cube only enables covering hierarchies where only immediate parent and child values can be connected as follows: Given three categories, $C_1$, $C_2$, $C_3$, such that $C_3$ is one of immediate predecessors of category $C_2$, and $C_2$ is one of immediate predecessors of category $C_1$, the mapping from $C_2$ to $C_3$ is covering with respect to $C_1$ iff $\forall e_1 \in C_1(\forall e_3 \in C_3(e_1 \leq e_3 \Rightarrow \exists e_2 \in C_2(e_1 \leq e_2 \wedge e_2 \leq e_3)))$. Otherwise, it is non-covering with respect to $C_1$. However, the location dimension hierarchy in the social network data can be non-convering. For example, there are some addresses without associated cities (they are thus directly part of administrative divisions). Therefore, if we compute aggregates at the city level, we will have no value for these addresses. Thus, the facts mapped to these addresses will not be considered in aggregations by administrative division through the computed aggregations by city. To overcome this type of summarizability problem, we use the algorithm presented in [PJD99] which proposes to insert an intermediate value at the city level and link this value with the address and the administrative division level.

From the several possible metrics that can be used in the fact type, we choose the following ones:

- **Term Frequency Inverse Document Frequency ()**: in order to determine the top-K most representative words of a location or a user, we represent the aggregation of all social content transmitted by each user or transmitted from each location as a normalized  vector (i.e., in the range [0, 1]), which reflects how important a word is to a collection of social messages.
- **Semantic word distance**: Usually, users utilize different words to describe the same idea, or they use same words but not necessary with the same meaning. Addressing these issues require acquiring information regarding the way these words are used or about the different personal background of users. Thus, in order to support queries like: *"what is the semantic distance between a specified location and a selected word?", "what is the most closest or distant word to a specified user in a period of time?"*, etc, we use the well-founded measure "Normalized Google Distance (NGD)" introduced by Cilibrasi and Vitanyi's in [CV07]. This measure does not depend on a particular dictionary or corpus, as it takes advantage of the

vast knowledge available on the web where all possible interpretations for a word are considered. The NGD measure calculates the distance between two words $w_i, w_k$ as follows:

$$NGD(w_i, w_k) = \frac{max\{\log f(w_i), \log f(w_k)\} - \log f(w_i, w_k)}{\log P - min\{\log f(w_i), \log f(w_k)\}} \tag{7.1}$$

$P$: 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 word $w_i$ 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 Normalized Google Distance by using different web search engines as the frequency source. Thus, the semantic word distance is calculated by computing the average of the NGD measure between the word studied and the top-K most representative words characterizing users or locations obtained from  vector. As an example, in the following formula, we calculate the semantic distance between *user s* and *word w*:

$$dist\_word(user_s, w) = \frac{\sum_{k=1}^{K} NGD(w_k, w)}{K} \tag{7.2}$$

where $K$: the number of top-K most representative words selected to represent users.

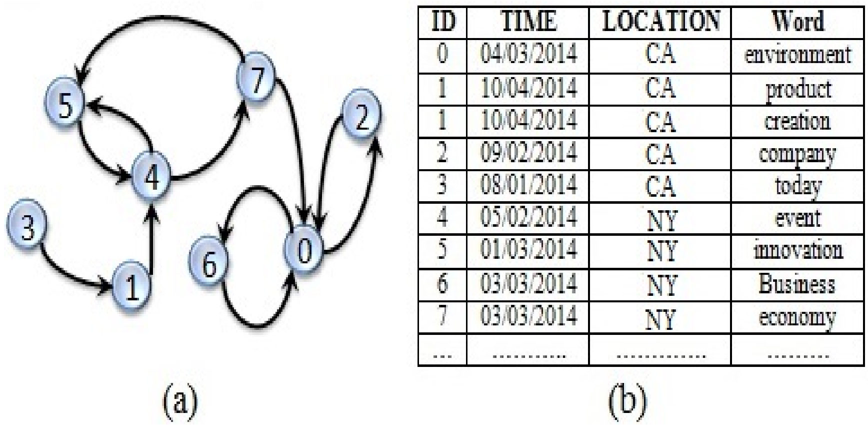
- **Sentiment measure**: the content of social data is a rich source of sentiment analysis. In order to automatically define the sentiment of social content, we use the top-K most representative words characterizing each user or location as candidate sentiment words. For each word in this collection we calculate the two binary polarization weights (positive, negative) by using Naive Bayes classifiers. These weights are summed to represent the word sentiment. If the sum of the top-K most representative words sentiment is greater than 0, then this collection is regarded as positive. In the other side, if the sum is less than 0, then this collection is regarded as negative. As we can see, we use the simplest method to calculate

> the user's sentiment or the location's sentiment. However, our aim is to present a quick overview of the opportunities of the mapping between the sentiment analysis field and the OLAP techniques in the analysis of social networking services.

The Location dimension is associated with the following order in its category types: $\perp_{Location} = Address < City < Administrativedivision < Country < T_{Location}$. Additionally, the Time dimension type has the following order in its category types: $\perp_{Time} = Hour < Day < Month < Year < T_{Time}$. These two dimensions are utilized for two purposes: First, they are used in the selection of user's words. For example, in case that the analyst's interest is to study the reaction of users in a specific location towards event over a specific time period, she shall express her request through the following query *"What is the semantic distance between tweets transmitted over the time period t and users of location l?"*. The two dimensions (time, location) are utilized to define the list of words mentioned over the period $t$ and define the list of words mentioned by users of the location $l$. Only the top-K most representative words characterizing period $t$ and users of location $l$ are used to answer the previous query. Second, they are used to carry out the different types of OLAP operations, such as the grouping of users and words at a desired level of detail, which can be represented by the list of users' communities based on users' location. However, in the case of the word dimension, classic OLAP system cannot handle the grouping of users according to the words' closeness. Thus, in the following section, we propose a new method to map the community extraction area to OLAP query processing in order to define the aggregation of users according to word dimension.

## 7.4 Word Community Extraction

As mentioned before, we use the list of words in user messages to infer the list of communities. We present in Figure 7.3 an example of a simple social network consisting of the set of users presented previously in Table 7.1. The users in this network are interconnected with a 'follower' relationship. There are eight nodes (identified with an ID) and eleven edges in the underlying graph, as shown in Figure 7.3(a). We process the list of additional data associated with real-world tweets in Table 7.1 (such as user identifier, time, longitude, latitude, etc.), and the semantic information which characterizes the set of words identified in those, in order to define the set of multidimensional attributes characterizing each user in the social network. These attributes include user ID, time, location (in state) and word, which are represented as a tuple in a vertex attribute table, as shown in Figure 7.3(b). The topological structure of the graph, together with the multidimensional attributes associated with vertex, forms a multidimensional network.



| ID | TIME | LOCATION | Word |
|---|---|---|---|
| 0 | 04/03/2014 | CA | environment |
| 1 | 10/04/2014 | CA | product |
| 1 | 10/04/2014 | CA | creation |
| 2 | 09/02/2014 | CA | company |
| 3 | 08/01/2014 | CA | today |
| 4 | 05/02/2014 | NY | event |
| 5 | 01/03/2014 | NY | innovation |
| 6 | 03/03/2014 | NY | Business |
| 7 | 03/03/2014 | NY | economy |
| ... | ............ | .............. | .......... |

Fig. 7.3: Example of multidimensional social network

On the one hand, traditional data cube and OLAP techniques use the structured attributes dimensions to support the navigation along any dimension hierarchies. For example, these techniques group users in Figure 7.3 based on dimension location, i.e., the users which have the same value on dimension location are grouped together in the coarser levels of granularity. However, this approach makes it a great challenge to answer some complex queries of the form *"What is the list of communities extracted from the social network based on the semantic relationships at word level?"*

On the other hand, most existing works on community extraction in data mining and social network analysis areas generally aim to define the list of users communities mainly according to the topological structure which represents the connection or the communication between users. This topological structure is represented in Figure 7.3(a) using the 'follower' relationship between users. However, in the network structure, the relationships between users do not display the dynamics between them according to their common views. For example, it is not possible to infer common opinions or interests from the aforementioned 'follower' relationships. Therefore, it is not possible to provide the users with relevant recommendations.

To overcome the limitations presented previously, we propose in our model a new method that combines data mining mechanisms and OLAP techniques to navigate the user hierarchy by introducing the granularity community. This granularity represents the highest level in this hierarchy. The suggested method consists on the aggregation of users into a set of clusters according to the topological structure of the social network and the semantic relationship between users at word level. For instance, if there is a friendship between two users, and these users always talk about health in their social network exchanges, they will be assigned in a community related to health orientation even if they mention different words. To achieve this goal, we propose the following process: First, we calculate the semantic distance between users according to the set of words they mention; second, we construct the semantic graph which represents the different closeness relations between users; third, we detect the list of user communities based on the semantic graph constructed previously; fourth, we evaluate the obtained communities, in order to get the best result. In the following subsection, we study in detail each phase in this method.

### 7.4.1 Semantic distance between users

In [RZGSS04] Rosen-Zvi et al. present the distance between individuals as the symmetric Kullback-Leibler divergence between the topics distribution conditioned on each of the individuals, as follows:

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

Where $i, j$: represent $useri$ and $userj$. $T$: is the number of topics. $\theta_{it}, \theta_{jt}$: 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 words distribution rather than topics distribution. This distribution is computed by using the normalized  metric 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  weights. The distance between $users$ and $userl$ is then represented by the following formula:

$$dist_{S_1}(user_s, user_l) = \sum_{k=0}^{K} TF\text{-}IDF_s(w_k)log\frac{TF\text{-}IDF_s(w_k)}{TF\text{-}IDF_l(w_k)} + TF\text{-}IDF_l(w_k)log\frac{TF\text{-}IDF_l(w_k)}{TF\text{-}IDF_s(w_k)}$$

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

However, this expression compares only the weights associated with the same words without considering the semantic relatedness among the different words. For example, consider the following users with their top 3 most representative words according to normalized 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 *userl* and *users* is calculated depending only on the T weights associated with word *media.* It does not take into account the semantic relationship between 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 weights associated with different words distribution with another metric like NGD distance. The result of this combination is represented in the following formula:

$$dist_{S_2}(user_s, user_l) = \frac{1}{2} \times \left( \sum_{k=0}^{K} \sum_{i \in \{K-k\}}^{K} TF\text{-}IDF_s(w_k) log \frac{TF\text{-}IDF_s(w_k)}{TF\text{-}IDF_l(w_i)}) + \right.$$
$$TF\text{-}IDF_l(w_i) log \frac{TF\text{-}IDF_l(w_i)}{TF\text{-}IDF_s(w_k)} + NGD(w_k, w_i))$$

Based on the two previous equations, we suggest a global formula divided into two parts: The first part represents the distance between the selected users depending on the weight of same words, multiplied by the weighting factor $\alpha$. The second part represents the distance between users based on the weight and the semantic relatedness of different words, multiplied by the weighting factor $\beta$. These two weighting factors $(\alpha, \beta)$ indicate the relative importance or impact of each part in the formula, taken in the range $[0, 1]$. The result of this formula is represented as follows:

$$dist_S(user_s, user_l) = \alpha \times dist_{S1}(user_s, user_l) + \beta \times dist_{S2}(user_s, user_l) \tag{7.3}$$

### 7.4.2 Construction of the semantic graph

Due to their significant strengths, graphs have been vastly utilized for modeling interrelated and multi-typed datasets, such as large scale social networks, spatiotemporal data, the World Wide Web, etc. The components of these graphs are vertices and edges. In the social network case, the vertices represent a set of social entities such as users or social organizations, while the edges just stand for straightforward connections, i.e.: they indicate that a direct relationship has been created during social interactions. However, there is abundant significant information between social entities. In order to maximize the potential value when using this representation and visualization method, we propose to model social networks by a semantic graph. This graph not only presents the communication relationship between social entities but also enriches the modelisation with another type of relationship, which vividly describes the closeness of interests and views between social entities. This relationship is obtained by using the semantic distance equation defined in the previous subsection. To construct the semantic 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 $i$ to the vertex $j$ if the semantic distance between these two users is less than or equal to the mean associated with *user i*, which is calculated by the following formula:

$$mean_i = \sum_{u=1}^{U} dist_s(user_i, user_u)/U \tag{7.4}$$

where $U$ represents the total number of selected users.

In the case where we get a large number of semantic edges compared to the number of vertices, we repeat this process.

By considering the list of words associated with the eight users presented in Figure 7.3(b), we present in Figure 7.4 the semantic relationships between these users. In that graph we can see the bidirectional relationships between $(user_0, user_7)$, $(user_0, user_1)$, $(user_5, user_2)$...etc, which reveal that there is a very high degree of similarity between these users. On the other hand, the unidirectional relationships between: $(use_6, user_0)$, $(user_3, user_7)$...etc, let us know that even though there is certain similarity between these users, the weight associated with it does not permit to create the relationship in both directions.

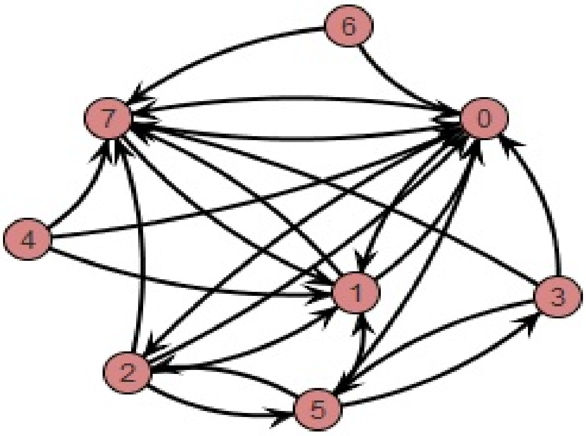


Fig. 7.4: Semantic graph without topological relationships

Third, we create an edge from the vertex $i$ to the vertex $j$, if there is no edge connecting these vertices and there is a direct communication relationship from *user* $i$ to the *user* $j$. If there is already an edge connecting the vertex $i$ to the vertex $j$, we increase the weight associated with the edge between these two vertices. Figure 7.5 presents an example of semantic graph generated from the list of words presented in Figure 7.3(b) and the topological structure of graph presented in Figure 7.3(a). The vertices in these graphs represent the eight users while the edges represent the communication relationships and the semantic closeness between these users according to the list of words. As we can see, the difference between Figure 7.4 and Figure 7.5 is the addition of the following six edges: $(use_0, user_6), (user_1, user_4), (user_3, user_1), (user_4, user_5),$

$(user_5, user_4), (user_7, user_5)$.

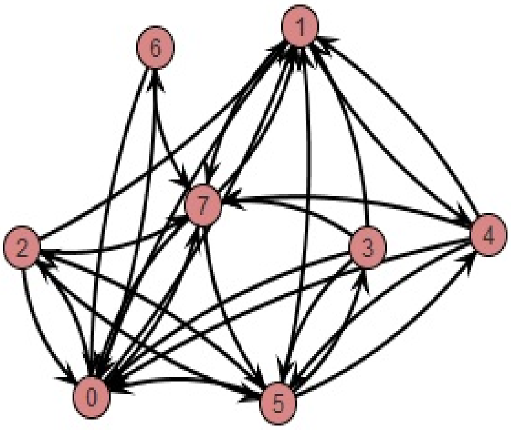


Fig. 7.5: Semantic graph with topological relationships

### 7.4.3 Extract Users' Communities

Using the obtained semantic graph from the previous subsection we can extract the users' communities by adapting the algorithm of Newman [GN02b], which is one of the most successful and well-known algorithms proposed so far. This algorithm is built around the idea of utilizing the concept of edge betweenness centrality to detect the list of communities. Newman's algorithm is a divisive classification, which represents a top down approach. It starts with all vertices as a single community and applies the division method. The algorithm proposed in [GN02b] is based on the following steps: first calculate the betweenness centrality values for all edges in the semantic graph. Then, find the edge with the highest value and remove it from the semantic graph. Next, recalculate the betweenness centrality for all for all edges affected by the removal. Finally, repeat the process for all remaining edges until we obtain the set of users' communities.

As we can notice, the classic approach uses the traditional graphs to calculate the betweenness centrality, where edges just stand for straightforward connection between users. The betweenness centrality measures the influence of a vertex over the flow of information between other vertices by using the simple relationships presented in the graph. However, in our approach, we improve this algorithm to extract the set of communities by using the semantic graphs shown in the previous subsection. The betweenness centrality in this case reveals the influence of a vertex not only according to his social connections, but also according to his semantic closeness to all other vertices.

### 7.4.4 Evaluation of obtained Communities

In this subsection we evaluate the extracted communities based on the closeness of interests and views between social entities. The question asked here is *how to get the best results?* Girvan and Newman provide an answer in [GN02b] by defining their popular modularity metric that evaluates the extracted communities. This metric is calculated by comparing the number of edges within a community minus the expected number in an equivalent network with edges placed at random. Building on this idea, Arenas et al. present in [ADFG07] an extension of modularity for directed graphs. The adapted formula is:

$$Q = \frac{1}{m}\sum_{i,j\in V}(A_{ij} - \frac{k_i^{out}, k_j^{in}}{m})\delta(C_i, C_j) \tag{7.5}$$

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

We think that this metric is more suitable for community extraction methods based on the topological properties contained in classic graphs, as it computes only the link between users contained in classic graph without considering the semantic relationship between them. Thus, we will compute the modularity by using the semantic graphs constructed in subsection 4.2. After the modularity computation for the resulted communities over the experimented users, we note that the division in three communities (where the first one contains user2, the second one is comprised of user0, user1, user3, user5, user6 and user7, and the final one contains user7), has the maximum value of modularity.

## 7.5 OLAP queries on Social Text Cube

Traditional OLAP techniques were initially designed to manage straightforward numeric aggregations on relational databases. For this reason, they are not well suited to handle the multidimensional data arising in real-world situations, such as the data present in social network services. For example, decision makers are interested in getting an unprecedently comprehensive view of ongoing events or similar interests through Twitter. They turn to a large social network and a big collection of tweets to study the interest and preference models of interconnected users within several multidimensional areas, such as words, locations, times, topological structures and possible combinations of these properties. These reports let analysts study and analyze the underlying network in a summarized way within different multidimensional areas, which is effective and of high value in the majority of data warehousing systems.

For example, we can consider a semantic clustering tasked with answering queries such as *"What is the structure of the aggregation network as grouped by the users' words?"*

We propose to calculate the distance between users according to the similarity between their most representative words, and then use the results to define the communities. For example, using the data presented in Figure 7.3(b) generates the results shown in Figure 7.6. The vertices represent users (identified with user ID). The links represent the list of pertinent semantic relationships between users, regardless of whether they belong to the same community or not.

users belong to the same community and users belong to different communities. The pertinence of semantic relationships is defined according to the high values of closeness between users at dimension "word".

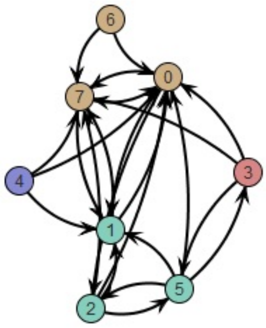


Fig. 7.6: List of communities according to word closeness

In the Figure 7.6 we can see how we used different colors to characterize the list of extracted communities, which consist of the users from the original network that belong to the same community. As we can see, *user0, user6* and *user7* talk about economic news in their tweets, thus they are grouped in the same community (related to business). While, *user1, user2* and *user5* talk about companies and products, thus they are grouped in a different community (focused on marketing). Finally, *User3* and *user4* describe their Personal Status in their tweets, so they are grouped in different communities.

Figure 7.7 presents another example of extracted communities by summarizing the multidimensional network shown in Figure 7.3 according to the structure and the semantic dimension relationships. In this example, we do not consider the list of selected words as the only inputs, but also the network structure. Different colors are used to present the characteristics of each extracted community in the aggregate network. As we can see, based on the communication relationship, *user0, user1, user3, user5, user6* and *user7* are fused together within the same community. However, this is not the case with *user2*, where the insufficiency of shared communication relationships between this user and the list of the other users makes changes to his community.

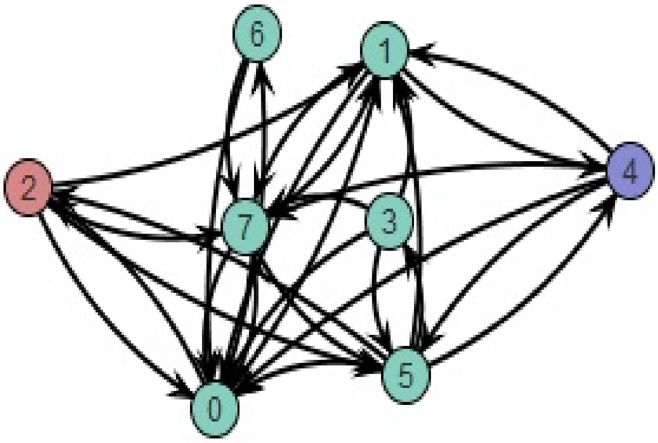


Fig. 7.7: List of communities according to semantic and topological relationships

### 7.5.1 NetCuboid Query

In our framework, we propose a Social Text Cube which combines properties of underlying multi-dimensional networks with existing data cube technologies and data mining clustering in a unified approach. An important type of OLAP query on network scenarios is to represent the result of a particular aggregation of the multidimensional network as an aggregate network. In [ZLXH11a] Zhao et al. determine two types of queries across network data: a cuboid query, producing a particular aggregation of the multidimensional network space, and a crossboid query (cross-cuboid), which is a new class of OLAP queries that crosses and straddles multiple cuboid queries simultaneously. A list of complex queries for OLAP cubes can be defined on social networks, such as *"What is the list of communities extracted from the social network based on the semantic relationships at location level?", "What is the set of communities detected from the social network based on the semantic and topological relationships at location level?"*. These queries may be very illuminating and an efficient way to reveal interesting information and behaviors, which are very difficult, if not impossible, to extract from the initial social network. In this contribution we suggest to go beyond existing cuboid queries by introducing a new class of OLAP queries not previously studied, called NetCuboid, which represents a special kind of network cuboid queries. This type of queries breaks the boundaries created in the classic OLAP paradigm in that it straddles different aggregations simultaneously and produces further insights through an aggregate network across different multidimensional network scenario. The multidimensional attributes associated with networks entities, the user-generated content (text), and the topological structure of the networks are considered into a single integrated framework for network summarization. To the best of our knowledge, this type of queries has not been studied before.

In order to obtain the results of the first query presented previously, we would use the following process: first, we summarize the multidimensional network in coarser levels of granularity by grouping vertices according to the location dimension, i.e., the vertices that have the same value on the location dimension are grouped together in the same group. In our case, we have two groups Figure 7.8(a) and Figure 7.8(b). In the first group, we have *user0, user1, user2,* and *user3* belonging to the same state (CA). We can also see how *user4, user5, user6,* and *user7* belong to the second group (located in NY). Then, we construct the semantic graph of each group. The edges within each group represent the semantic closeness between vertices according to shared views and interest. Thirdly, based on semantic graphs resulting from previous step, we define the list of extracted communities within each group by characterizing them with different colors. As we can see in Figure 7.8, the two groups include a division of two communities. The first community in the first group contains *user1*. The second community in the first group contains *user0, user2,* and *user3*. In the second group, we have *user5* as community one and *user4, user6,* and *user7* as community two.

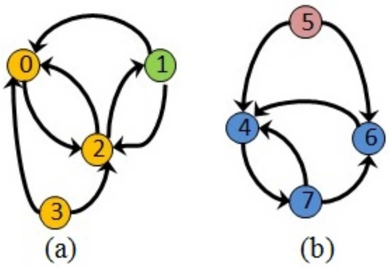


Fig. 7.8: aggregation of users according to the location and the semantic closeness

To obtain the outcome of the second query presented previously, we would use the following process: first, we aggregate the social network in coarser levels of granularity by grouping vertices according to dimension location. As discussed previously, we have in our case two groups. Then, we construct the semantic graph of each group. The edges within each group not only represent the topological relationships but also the semantic closeness between vertices according to shared views and interest. Finally, based on semantic graphs resulted from previous step, we define the list of extracted communities within each group. The result of this process is presented in Figure 7.9. The dashed lines between groups represent the topological relationships between vertices belonging to different groups.

## 7.6 Conclusion

In this Chapter we propose a new dynamic data cubing and mining framework, called Community Cube, for analyzing social network data. This framework is designed to

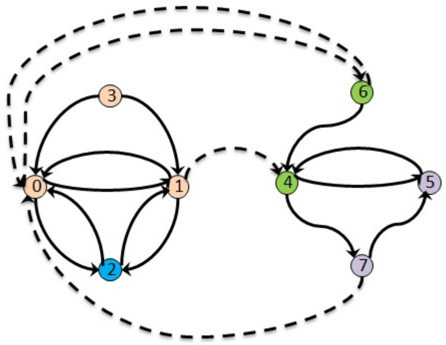


Fig. 7.9: aggregation of users according to location, semantic and topological relationships

support OLAP-style multidimensional analysis on information-enhanced multidimensional social network. Community Cube provides pertinent answers to analyst queries that have not been addressed until now.

Going beyond traditional OLAP operations, where OLAP aggregations are directly computed by using the list of attributes associated with the relational data, the Community Cube architecture proposes a new method that combines data mining and OLAP operators to navigate through hierarchies. This method consists on grouping individuals into different communities according to similar characteristics and views, which proves to be much more meaningful and comprehensive than classic aggregation in the traditional OLAP techniques. Different from the most community extraction methods focused on the relations between users, our proposed method is based on the aggregation of users into a set of clusters according to the topological structure of the social network and the semantic relationship between users. By using our proposed community extraction method we can detect the emerging interests or orientations in each community.

Besides traditional OLAP queries, our approach introduces a new class of queries, NetCuboid. This queries take into account the multidimensional attributes associated with networks entities, the user-generated content, and the topological structure of the networks. NetCuboid has shown to be effective and useful to navigate through unstructured text and topological structures.

# Chapter 8

# Experiments

## 8.1 ABSTRACT

This Chapter illustrates the obtained experimental results of the different approaches and models proposed in this thesis. They permit the decision-makers to analyze quickly and navigate through the social network data from different perspectives and with multiple granularities. These results are generated by using three types of data: The first one is the structured data illustrated in the list of multidimensional attributes associated with the social entities. The second one is the unstructured user-generated content produced in the social network services, and the third one is the topological structure data which represents the social entities with the set of relationships among them.

## 8.2 Introduction

In this Chapter, we present some brief experimental studies evaluating the that the different approaches presented in this thesis can provide for analyzing social networking data. We trained the proposed approaches 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 approaches and experimental methods are implemented in Java and tested on a Windows PC with dual

Intel Xeon processors hexacores at 3.06 GHz and 12 GB of RAM.

## 8.3 Experimental Results

We initially evaluate the efficiency of the proposed Community Cube framework as a robust decision-support system in the social network data. We will show several remarkable results by using OLAP queries on the networks generated from our Twitter data acquisition process. In the experiments, we are concerned with the features of the users from different perspectives, such as semantic, geographic, and temporal.

As an instance of results obtained in our experiments, in Figure 8.1, we present the distance between five experimental users selected from our database and a list of words. Different colors are used to characterize users. As we can see, the word distance differs depending on the selected user. This representation can answer many queries like, *"what is the most closest or distant word to user1?", "how the distance of a specified word is changed from one user to another?".* The results of these queries can represent a rich source of information in various areas such as data mining, where we can use this distance to group words in different clusters.

In Figure 8.2, we present different communities with their most representative domains detected by using ODP taxonomy. From this figure, we notice homogeneity within each community where all its users treat related domains with height values. For example, the top two domains associated with *community 1* are respectively, *Internet* with a weight equal to *0.79* and *Technology* with a weight equal to *0.71*. From this information we can derive that the orientation of *community 1* is mostly related to computers area. This information can be exploited in different fields, such as in the influence analysis, we can study how and when one community is influenced by the orientation of another community. In the prediction field, we can predict future trends based on the analysis

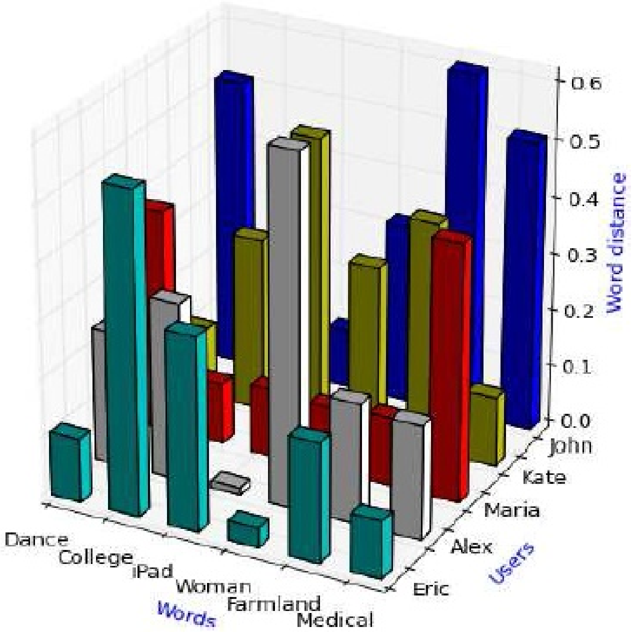


Fig. 8.1: Word distance according to users

of past treated domains.

In our study, the aggregated user networks may be considered to show compressed visions of the hidden information. This is achieved by partitioning users in a huge network into various clusters, where similar users are grouped together into the same cluster, while distinct users are separated into different clusters. These similarity and dissimilarity are evaluated according to different criteria and perspectives.

With these results, we can now obtain the aggregation of one hundred users according to dimension "location". The summarized aggregate network is shown in Figure 8.3.

This representation displays a higher-level summary of the twitter follower network by applying the roll-up operation. This aggregation is obtained by using the process proposed in [ZLXH11b]. In this network vertices represent the list of locations and the weight of the vertex illustrates the number of users that correspond to the same location values. Edges represent the overall relationships between locations and the edge

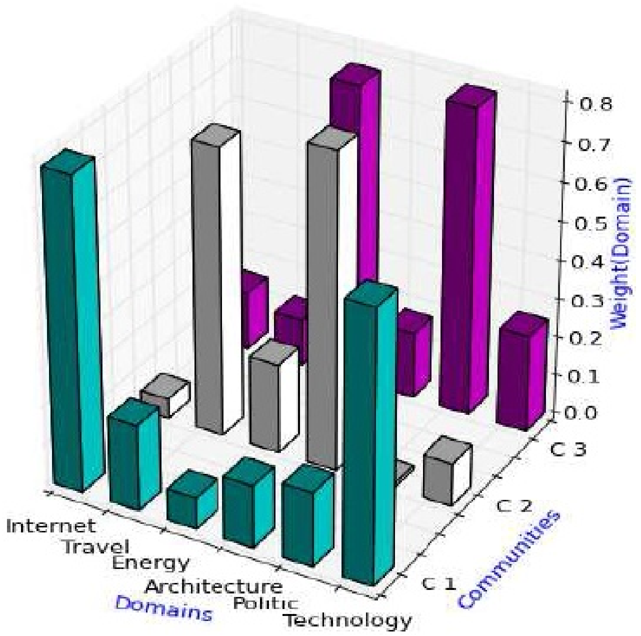


Fig. 8.2: Domain weight according to communities

weights illustrate the number of ’follower’ relationships connecting users belonging to two different locations. One conclusion we can draw from this graph is that the bidirectional relationships between locations (New York, California), (Montana, California), and (Florida, Michigan) in Figure 8.3 reveal that there is very a high degree of follower connections occurring between users belonging to these locations. On the other hand, there is no relationship between (New York, Virginia), (Florida, California), or (Florida, Montana) which indicates that there is no existing direct communications path between these locations, which translates as a lack of flow of information between users of these locations.

If decision makers are concerned with zooming into more lower-level granularity for additional details, a drill-down operation may be used. Figure 8.4 presents the lower level snapshot obtained from Figure 8.3 by applying the drill-down operation according to dimension “location” and dimension “user”.

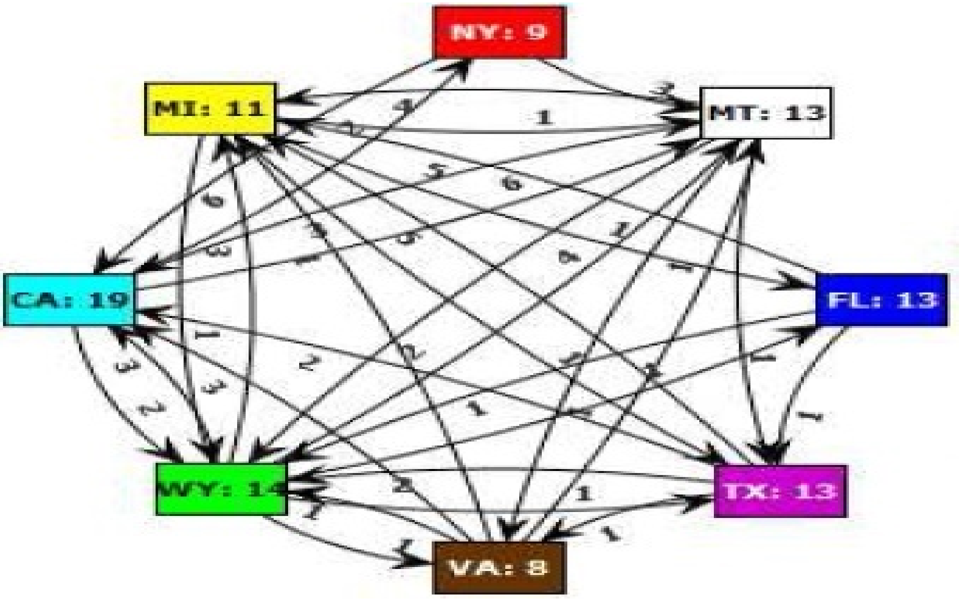


Fig. 8.3: Roll-Up according to location dimension

As we can notice, we are navigating the data from a more general view towards a highly specialized analysis. In this network, different colors are used to characterize each location. Vertices represent the list of users and edges represent the follower relationship between users. By selecting one vertex from this network, we can get the list of specific properties associated with it, such as user ID, name, url, description...etc.

In Figure 8.5 we can see an example of the distance between an experimental user selected from location California in Figure 8.4 and a list of words according to various intervals of time.

Different colors are used to characterize the selected periods of time. As we can see, the word distance differs depending on the selected periods. This representation can answer many queries like, *“what is the closest or most distant word to user1?”*, *“how the distance of a specified word changes from one user to another?”*.

The results of these queries can represent a rich source of information in various areas such as data mining, where we can use this distance to group words in different clusters.

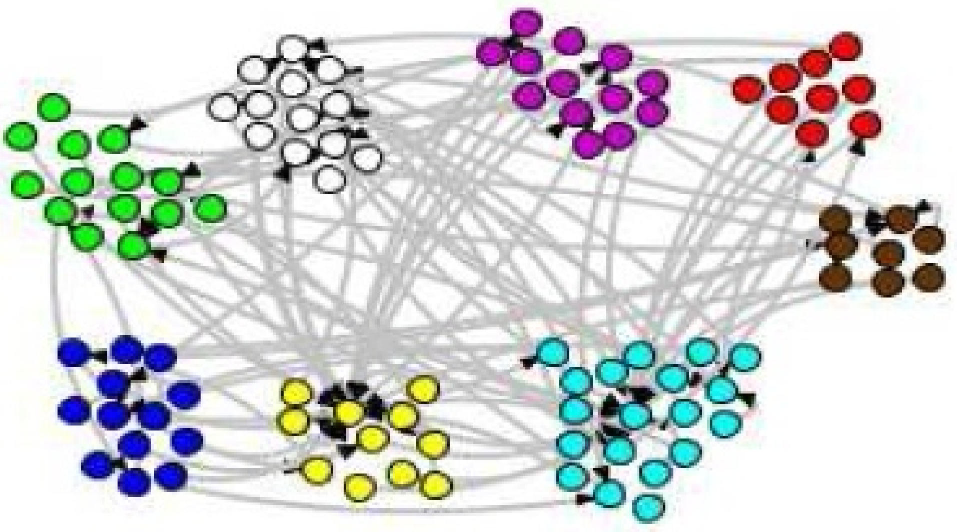


Fig. 8.4: Drill-Down according to location and user dimensions

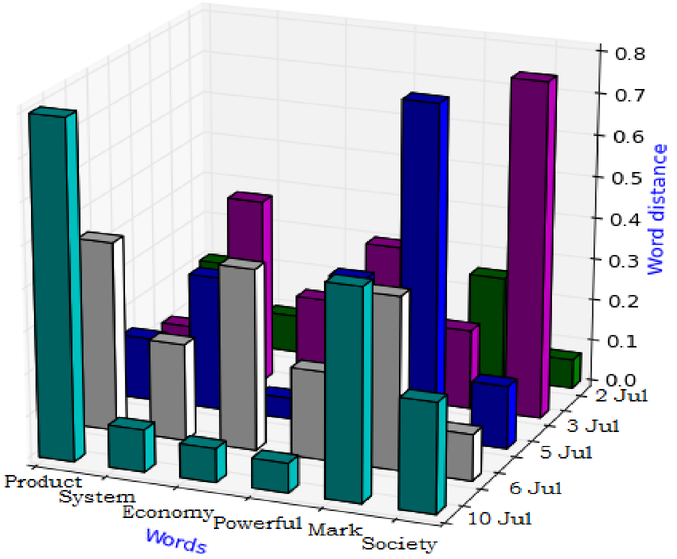


Fig. 8.5: Word distance according to different periods of time

In the case where the analyst's interest is to study the semantic closeness between locations over a specific time period, most existing studies on OLAP and multidimensional networks (such as [CYZ$^{+}$09], [QZY$^{+}$11], [ZLXH11b], or [THP08]), cannot handle this type of complex analysis. In these proposals, only multidimensional attributes associated with network entities are used without considering the user-generated content (text), which could be very illuminating and an efficient source to reveal meaningful information between social entities. To answer this analysis query, we present in Figure 8.6 the resulting semantic aggregation network.

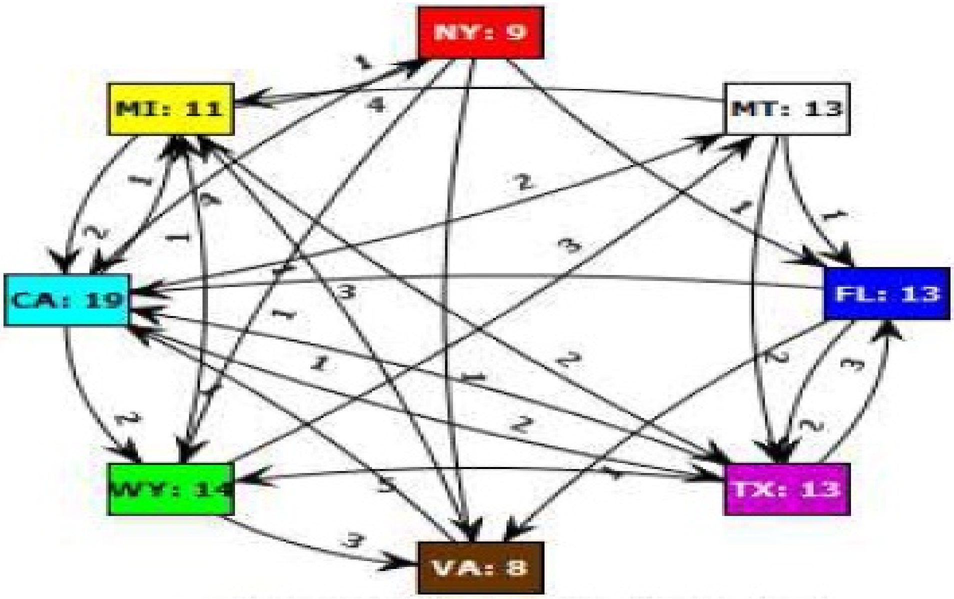


Fig. 8.6: Semantic Roll-Up according to location dimension

Only the most representative words characterizing users belonging to disparate locations are used to calculate the semantic closeness between locations. In this semantic network, we enrich the modelisation with a new type of relationship, which vividly describes the closeness of interests and views between locations. The vertices represent the list of locations and the weight of vertex illustrates the number of users that correspond to the same location values. Edges represent the overall semantic relationships between locations and edge weights illustrate the number of pertinent semantic closeness between users belonging to two different locations. As we can see, although there is no directed communication between New York and the list of locations (Wyoming, Virginia, Florida) in Figure 8.3, the semantic network presented in Figure 8.6 reveals that this list represent the most closest location for New York. On the other hand, despite existing follower relationship between Montana and Virginia, there is no semantic closeness between these two locations. From these representations, we find that direct communication does not reflect the presence of common opinions or interesting information between locations.

As we can notice in the proposed approach, the selection of the most representative words plays a major role in our framework. It is not only used in the majority of measures but also in the detection of communities. Thus, it is extremely important to correctly define the number of the representative words. We apply our proposed community extraction method to the social network of 1500 users with various numbers of selected words. The resulting modularity values Q are plotted in Figure 8.7.

As can be seen from the figure, the highest value of modularity that can be obtained is 0.54, achieved with the selection of the top 4 most representative words associated with each user according to  weight.

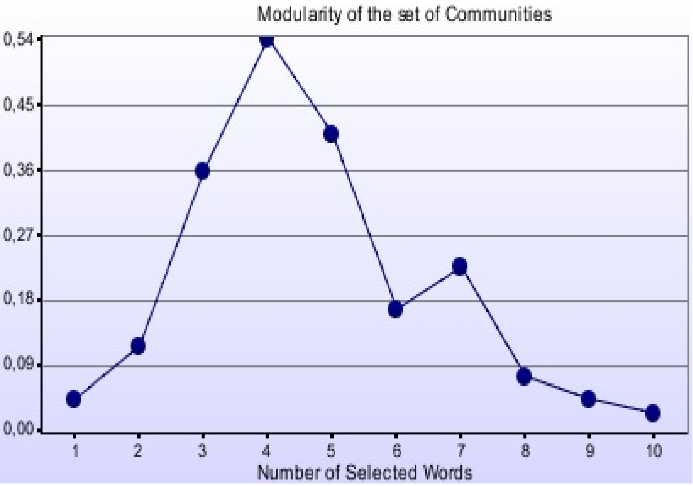


Fig. 8.7: Modularity associated with different number of selected words

Additionally, decision makers could express their inquiries through more complex queries in different forms, such as *"What is the list of communities extracted from locations based on the semantic relationships", "What is the closeness between these communities?".* The aim of these analyses is to divide the social network into different communities with high intra-community semantic similarity and homogeneous location values. In our framework we answer this type of queries by combining the data mining field and the OLAP techniques, and the resulting aggregate network is presented in Figure 8.8.

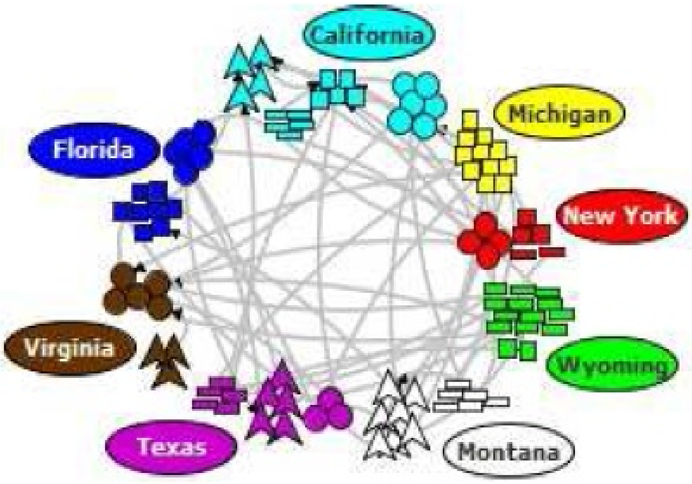


Fig. 8.8: List of communities detected in each location

To generate this output, we start by calculating the semantic closeness between all users according to the set of words they mention. Then, we divide users into different clusters according to location similarity, so that users belonging to the same location are grouped into the same cluster. Finally, we extract the list of users' communities in each cluster. In this network, we find two types of vertices: the first type is represented with ovals, and it refers to locations; the second type denotes users of the different communities. Different colors are used to identify locations, while different shapes are used in the second type of vertices to characterize the list of communities detected in each location. Edges represent the set of pertinent semantic relationships between users, regardless of where they belong to: the same community in the same location, different communities in the same location, different communities in different locations. The pertinence of the semantic relationships is defined according to the high values of closeness between users. From this aggregate network we observe that there are four communities in California, three communities in New York and Texas, two communities in Florida, Virginia, Montana, and Wyoming, and one community in Michigan

In Figure 8.9, we present the semantic aggregate network obtained as the answer for the following query: *"What is the list of communities extracted from the social network based on the semantic and topological relationships between users?"*

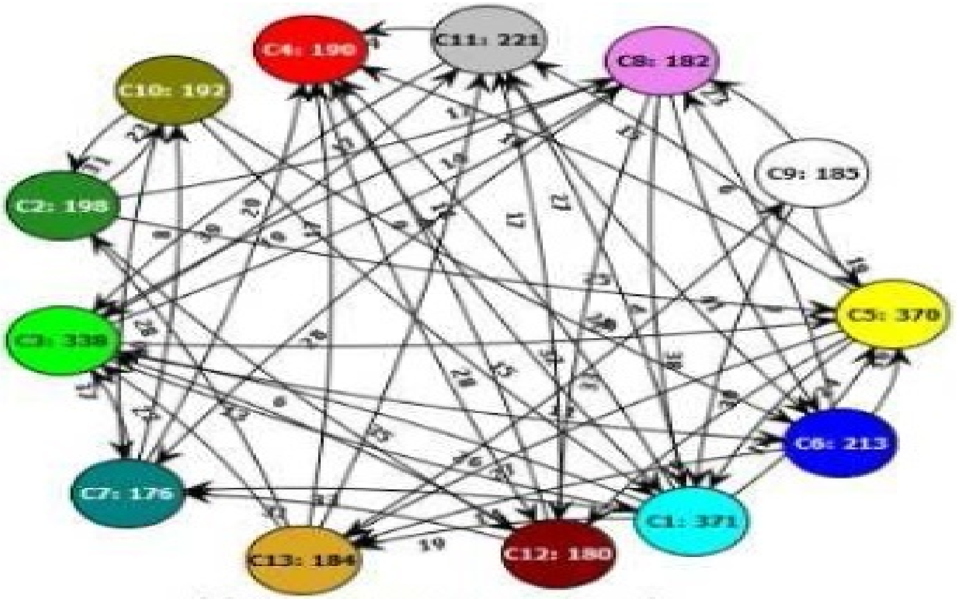


Fig. 8.9: Semantic Roll-Up according to user dimension

This representation showcases the best decomposition in communities of three thousand users selected from our dataset, and it was constructed by selecting the partition with maximum modularity Q value. The goal of this query is to identify cohesive communities of users within the social network with high intra-community topological structure and homogeneous interests. As we can see, the topological structure and the semantic closeness are seemingly independent. However, in our approach we are able to integrate these two concepts into a unified framework by using the following process: First, we define the structure of the social network where users are interconnected with 'follower' relationship. Then we enrich this structure with semantic closeness relationship. Finally, we determine the list of communities contained in the generated network. According to the results observed in our experiments, the values associated with modularity measure in the Community Cube framework could be increased or decreased depending on the pertinence and the number of selected users. However, most of the values associated with $Q$ are obtained in the range of [0.3, 0.7], which represents a strong community structure.

In Figure 8.9, vertices represent the list of extracted communities and the weight of the vertex illustrates the number of users that correspond to the same community. Edges represent the overall relationships between communities and edge weights illustrate the number of semantic closeness and follower relationships connecting users belonging to two different communities. Users in this decomposition are classified into thirteen communities, depending on their point of interest. For example, by using this aggregate network, we can answer many queries like, *"What is the closest or most distant Community to Community C5?"* or *"How the relationship between two communities is developed from one period of time to another?"*.

Table 8.1: Top words among the list of communities, sorted by weight

| | |
|---|---|
| *C1* | *android, iphone, video, design, tool* |
| *C2* | *learn, doctor, life, intelligence, service* |
| *C3* | *creative, paint, artists, computer, beautiful* |
| *C4* | *architecture, skyscraper, emotion, build, plan* |
| *C5* | *money, organization, enterprise, access, mind* |
| *C6* | *biology , game, humain, knowldge, science* |
| *C7* | *terrorist, CNN, science, obstacle, archive* |
| *C8* | *flavor, hospital, cuisine, home, online* |
| *C9* | *internet, portrait, culture, selfie, media* |
| *C10* | *rebot, device, nature, google, yahoo* |
| *C11* | *hotel, picture, trip, woman, malaysia* |
| *C12* | *news, democrat, world, search ,peace* |
| *C13* | *mall, fashion, love, day, dog* |

In Table 8.1, we use the aggregate network presented in Figure 8.9 to respond the following query *"What is the top-5 most pertinent words detected in each community?"*. The list of words is selected by using weight.

We can observe homogeneity within each community, where most of its users treat related words with height values. From this information we can derive the orientation of each community. For example, it seems that community C1 is mostly related to technology, community C2 discusses about the education field, community C3 focuses on decoration, community C4 on building, community C5 on business, community C6 on science, community C7 on news, community C8 on health, community C9 on media, community C10 on innovation, community C11 on travel, community C12 on politics, and community C13 on shopping.

Additionally, the bidirectional relationships between community C3 and community C11 indicate that there is an extremely degree of closeness between the orientation of community C3 (which is decoration), and the orientation of community C11 (travel). The unidirectional relationship between the community C1 and community C9 notifies that, even though there is a semantic closeness between these communities, the weight associated with this closeness relationship does not allow to create the relationship in both directions. This information can be exploited in different fields, such as in the influence analysis. For instance, we can study how and when one community is influenced by the orientation of another community. In the prediction field, we can forecast future trends based on the analysis of past treated domains.

We can also address the drill-down operation of the previously aggregate network according to dimension "user". The resulting network is demonstrated in Figure 8.10.

The goal of this representation is to get a vision of generated findings in a more lower-level granularity for further inspection. Each color is used to indicate users belonging to the same community. Looking at this generated network, we can use the degree and the properties of each vertex to answer many queries, such as *"Which is the most representative user in each community?"*.

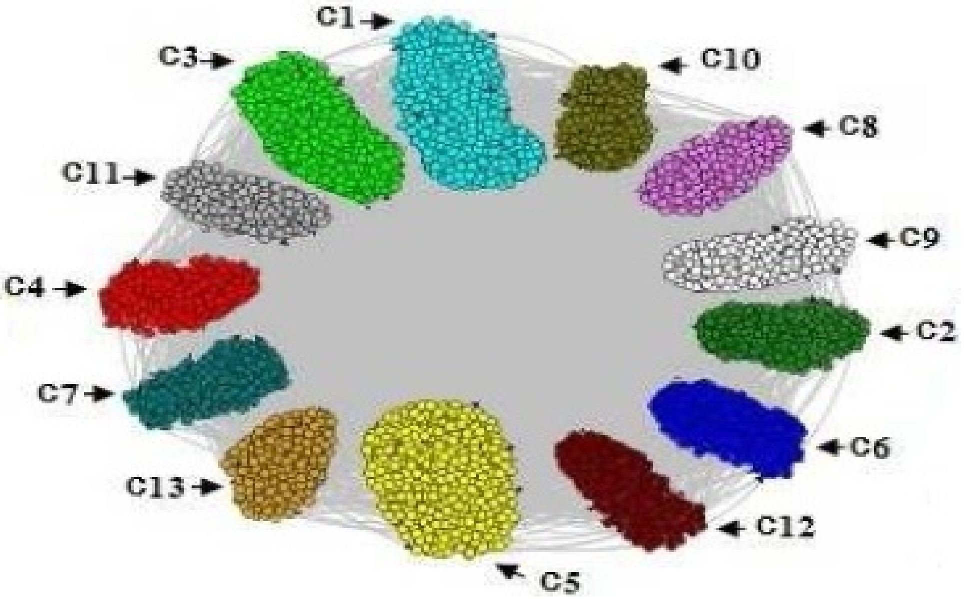


Fig. 8.10: Semantic Drill-Down according to user dimension

In Figure 8.11, we present another vision of the fine-grained network obtained from Figure 8.9.

It illustrates the drill-down operation of community C7 according to dimension “user” and dimension “location”. In this network we find two types of vertices; the first type (represented by the oval shape) refers to locations, while the second type (represented by the circular shape) denotes users. Different colors are used to identify locations. Edges represent the set of pertinent semantic closeness and ’follower’ relationships connecting users belonging to the same location or users belonging to different locations. For example, in community C7, the users belonging to Maryland communicate most with users in Nevada, while they communicate much less with users of New York.

In order to present the pertinence of our proposed Topic-Driven-model presented in Chapter 4, we determine for each user tweets its high-level-topics and their top associated domains. Based on the obtained results, we construct the user communities. Figure 8.12 illustrates the first five most treated topics in one community.

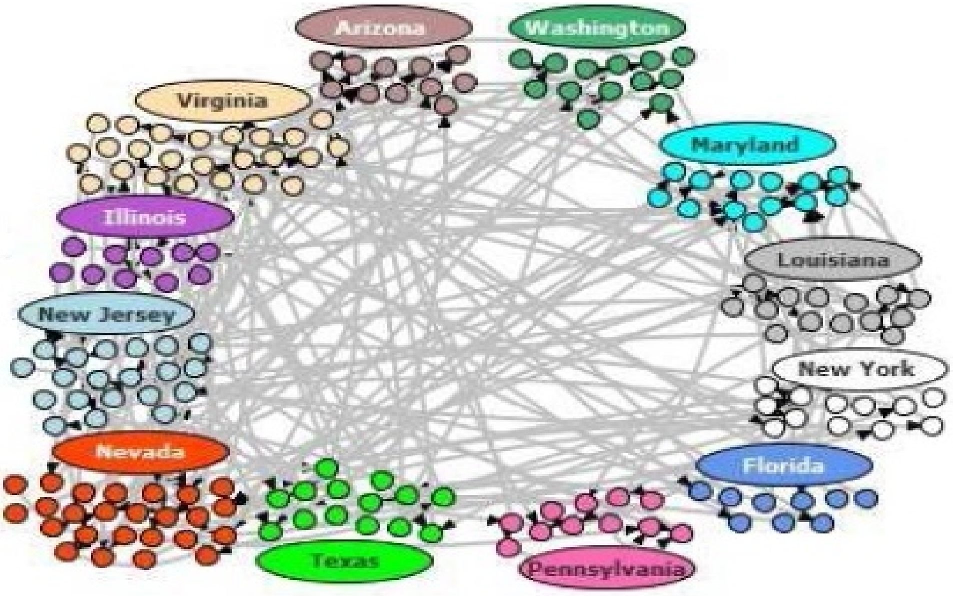


Fig. 8.11: Semantic Drill-Down according to user and location dimensions in community C7

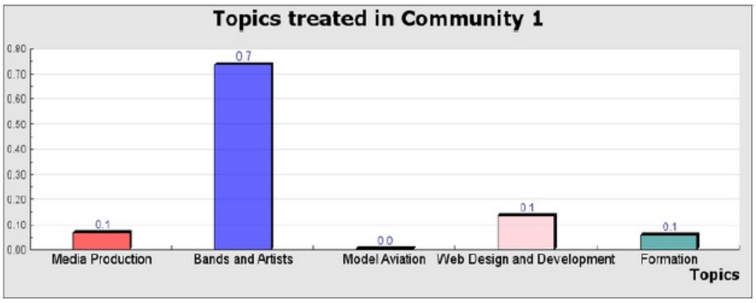


Fig. 8.12: Compressed vision of semantically related countries

We note that the majority of users in this community treat the topic 2 *"Bands and Artists"*.

We present in Figure 8.13 a compressed vision of the list of countries which are considered semantically related. As an example, by considering this representation, we can answer several queries like, *What is the most semantically closest or distant location to the country China?* or *How the relationship between two countries is developed from one period of time to another?*.

In Table 8.2, we illustrate the top six most pertinent words detected in each cluster

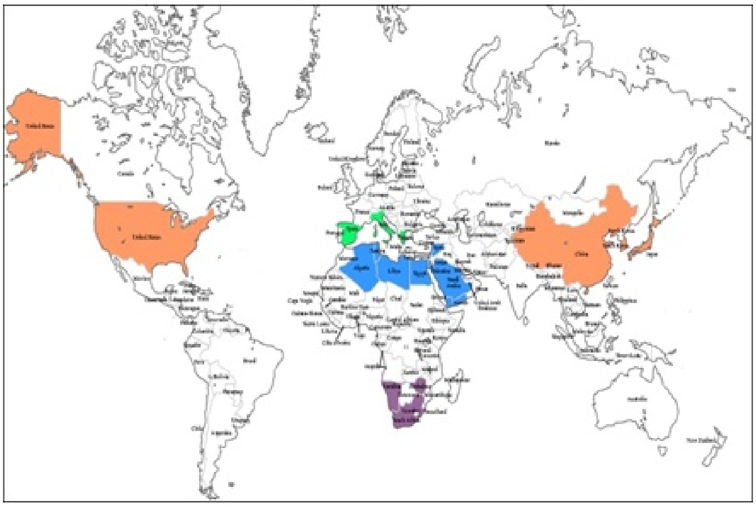


Fig. 8.13: Compressed vision of semantically related countries

presented in the previous figure. This list of words is selected by using TF-IDF weight.

From this table, we notice homogeneity within each cluster where most of its countries treat related words with height values. From this information we can derive the orientation of each cluster. As an instance, it seems that cluster C1 is mostly related to technology area, cluster C2 treats the political field; cluster C3 focuses on economic and cluster C4 on travel.

Table 8.2: Top words among the list of clusters, sorted by TF-IDF weight

| | |
|---|---|
| ***C1:*** United States, China, Japan, South Korea | intellect, assistance, china, system, digital, student |
| ***C2:*** Algeria, Tunisia, Egypt, Syria, Saudi Arabia, Yemen, Lebanon | Democrat, news, religion, internet, facebook, election |
| ***C3:*** Spain, Italy, Greece | Security, crisis, industry, company, business, woman |
| ***C4:*** Namibia, South Africa, Zimbabwe | tourism, nature, transport, photo, justice, health |

# Chapter 9

# Conclusion and perspectives

In the area of Decision support systems, on-line analytical processing (OLAP) tools have improved considerably the data analysis. The popularity of these tools has significantly increased in recent years. Mainly because they give users the ability to dynamically analyze data at different aggregation levels using operations such as roll-up and drill-down. Recently, social networks have quickly become a powerful means by which people share real-time messages. These messages provide the users with the ability to keep in touch with their contacts, using up of 140 characters in the case of Twitter sites. 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. Research in this field has focused mainly on the inference of social influence, building of user communities and prediction analysis. However, very little work has been carried out to investigate and document how OLAP tools can interactively analyze social networks data according to different perspectives and with multiple granularities. The main challenge that face analyzing Social network data with OLAP is that OLAP is designed to analyze structured data. However, it faces major issues in manipulating this complex interconnecting data.

In this thesis, we studied how to support OLAP-style analysis on information-enhanced multidimensional social networks for efficient information extraction.

We proposed a new approach called CETD: Community Extraction based on Topic-

Driven-Model to derive user's communities based on common topics from user's tweets. This approach combined our proposed model used to detect topics and domains of the user's tweets based on a semantic taxonomy together with a community extraction method based on the hierarchical clustering technique. By using CETD model, each two users talk about the same topic in their tweets, they are grouped in the same community which is related to this defined topic.

We suggested a new data warehousing model, Social Graph Cube to support OLAP technologies on multidimensional social networks. Based on the proposed model, we represented data as heterogeneous information graphs for more comprehensive illustration than the traditional OLAP technology. 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 presented a new method that combines data mining and OLAP operators to navigate through hierarchies. In this method network entities are grouped into different clusters according to similar interests, characteristics and views, which provide to be much more meaningful and comprehensive than classic aggregation in the traditional OLAP techniques.

We presented a new multidimensional model called "Microblogging Cube" to achieve OLAP techniques on unstructured data selected from social networks. It provided the possibility to analyze microblogs users and locations according to semantic, geographic and temporal axes. The semantic axe was defined by using the Open Directory Project (ODP) taxonomy. Different from existing classical multidimensional models, the measures in Microblogging Cube may vary depending on the aggregation levels. Further, in order to define the multiple granularities associated with microblogs users, we proposed a new process to extract the list of their communities.

We studied the use of data warehousing and OLAP technologies with such new multidimensional social network by proposing Community Cube architecture. This framework designed to support OLAP-style analysis on information-enhanced multidimensional social network data for efficient information extraction. Community Cube provides pertinent answers to analyst queries that have not been addressed until now.

Different from most of the community extraction methods focused on the connections between users, our proposed method is based on the aggregation of users into a set of clusters according to the topological structure and the semantic relationship between them. Besides traditional OLAP queries, our approach introduced a new class of queries, which we named 'NetCuboid'. These queries take into account the multidimensional attributes associated with networks entities, the user-generated content, and the topological structure of the networks.

In the future works, we will improve our experimentation and we will propose new evaluation protocols. We will go further in this analysis by using other types of measures proposed in social network area. We aim to extend OLAP analysis to other fields such as social sentiment, influence, etc and compare our proposed approaches with existing studies. We plan also to incorporate the strength of OLAP and IR tools in order to get a better vision about the unstructured data.

# Bibliography

[AAD$^+$96] Sameet Agarwal, Rakesh Agrawal, Prasad Deshpande, Ashish Gupta, Jeffrey F. Naughton, Raghu Ramakrishnan, and Sunita Sarawagi. On the computation of multidimensional aggregates. In *VLDB'96*, pages 506–521, 1996. 50

[ABL10] Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466:761–764, August 2010. 56

[ADFG07] Alex Arenas, J. Duch, A. Fernandez, and Sergio Gomez. Size reduction of complex networks preserving modularity. *CoRR*, abs/physics/0702015, 2007. 82, 103, 115, 144

[AW10] Charu C. Aggarwal and Haixun Wang, editors. *Managing and Mining Graph Data*, volume 40 of *Advances in Database Systems*. Springer, 2010. 13, 126

[Bar54] J. A. Barnes. [duplicate] Class and Committees in a Norwegian Island Parish. *Human Relations*, (7):39–58, 1954. 42

[Bav50] Alex Bavelas. Communication Patterns in Task?Oriented Groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, November 1950. 97

[BDRV07] Douglas Burdick, AnHai Doan, Raghu Ramakrishnan, and Shivakumar Vaithyanathan. Olap over imprecise data with domain constraints. In

Christoph Koch, Johannes Gehrke, Minos N. Garofalakis, Divesh Srivastava, Karl Aberer, Anand Deshpande, Daniela Florescu, Chee Yong Chan, Venkatesh Ganti, Carl-Christian Kanne, Wolfgang Klas, and Erich J. Neuhold, editors, *VLDB*, pages 39–50. ACM, 2007. 50

[Bea65] Murray A. Beauchamp. An improved index of centrality. *Behavioral Science*, 10(2):161–163, 1965. 97

[BGJT04] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, page 2003. MIT Press, 2004. 51, 60

[BGLM08] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, page P10008, 2008. 56

[BK07] Nilesh Bansal and Nick Koudas. Blogscope: A system for online analysis of high volume text streams. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 1410–1413. VLDB Endowment, 2007. 52

[BL06] D. Blei and J. Lafferty. Correlated Topic Models. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, 18:147, 2006. 50, 64

[Bla00] Phil Blackwood. 11 steps to success in data warehousing. *Transportation and Distribution*, 41:60, 2000. 27

[BNJ03] David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003. 13, 50, 51, 60, 64

[Bon72] Phillip Bonacich. Technique for analyzing overlapping memberships. *Sociological Methodology*, 4:176–185, 1972. 45

[Bou13] Doulkifli Boukraa. *Complex Object Data Warehouses: Multidimensional Modeling and Vertical Fragmentation*. PhD thesis, Ecole Nationale Superieure d'Informatique, ResearchGate, 2013. 8, 29, 30

[Bur80] R. S. Burt. Models of network structure. *Annual Review of Sociology*, 6:79–141, 1980. 43

[CCLR05] Bee-Chung Chen, Lei Chen, Yi Lin, and Raghu Ramakrishnan. Prediction cubes. In Klemens Bohm, Christian S. Jensen, Laura M. Haas, Martin L. Kersten, Perke Larson, and Beng Chin Ooi, editors, *VLDB*, pages 982–993. ACM, 2005. 50

[CCS93] E. F. Codd, S. B. Codd, and C. T. Salley. Providing OLAP to User-Analysts: An IT Mandate. 1993. 35

[CD97] S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. *ACM Sigmod Record*, 26(1):65–74, 1997. 15, 29, 32, 33, 50, 127

[CDH+02] Yixin Chen, Guozhu Dong, Jiawei Han, Benjamin W. Wah, and Jianyong Wang. Multi-dimensional regression analysis of time-series data streams. In *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB '02, pages 323–334. VLDB Endowment, 2002. 50

[CKKS02] W. F. Cody, J. T. Kreulen, V. Krishna, and W. S. Spangler. The integration of business intelligence and knowledge management. *IBM Syst. J.*, 41(4):697–713, October 2002. 52

[CV07] Rudi L. Cilibrasi and Paul M. B. Vitanyi. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, March 2007. 94, 111, 134

[CYZ$^+$09] Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu. Graph olap: A multi-dimensional framework for graph data analysis. *Knowl. Inf. Syst.*, 21(1):41–63, October 2009. 54, 157

[Dav06] Thomas H. Davenport. Competing on Analytics. *Harvard Business Review*, 84(1):98–107, January 2006. 21

[Dav10] Thomas H. Davenport. Business intelligence and organizational decisions. *IJBIR*, 1(1):1–12, 2010. 21

[DFLJ07] Haifeng Du, Marcus W. Feldman, Shuzhuo Li, and Xiaoyi Jin. An algorithm for detecting community structure of social networks based on prior knowledge and modularity: Research articles. *Complex.*, 12(3):53–60, January 2007. 56

[DH07] Thomas H. Davenport and Jeanne G. Harris. *Competing on Analytics: The New Science of Winning*. Harvard Business School Press, Boston, MA, USA, 1st edition, 2007. 21

[DMn04] L. Donetti and M. Muñoz. Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, page P10012, 2004. 14, 56, 62, 127

[Eck03] W. Eckerson. Smart companies in the 21st century: the secrets of creating successful business intelligent solutions. *The Data Warehousing Institute*, 2003. 22

[Fie73] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973. 56

[fJPF07] Fianny Ming fei Jiang, Jian Pei, and Ada Wai-Chee Fu. Ix-cubes: iceberg cubes for data warehousing and olap on xml data. In Mario J. Silva, Alberto H. F. Laender, Ricardo A. Baeza-Yates, Deborah L. McGuinness, Bjorn Olstad, ystein Haug Olsen, and Andre O. Falco, editors, *CIKM*, pages 905–908. ACM, 2007. 50

[For10] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010. 14, 55, 56, 62, 127

[Fra97] J.M. Franco. *Le Data warehouse, le Data mining.* Informatiques magazine. Eyrolles, 1997. 27, 31

[GCB$^+$07] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *CoRR*, abs/cs/0701155, 2007. 15, 50, 127

[GKT05] David Gibson, Ravi Kumar, and Andrew Tomkins. Discovering large dense subgraphs in massive graphs. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 721–732. VLDB Endowment, 2005. 54

[GN02a] Michelle Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Science*, 99(12):7821–7826, 2002. 56

[GN02b] Michelle Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Science*, 99(12):7821–7826, 2002. 81, 82, 86, 103, 115, 143, 144

[GR09] Matteo Golfarelli and Stefano Rizzi. *Data Warehouse Design: Modern Principles and Methodologies*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 2009. 39

[GS04] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004. 64

[GSM71] G.A. Gorry and M.S. Scott-Morton. A framework for management information systems. *Sloan Management Review*, 13(1):55–71, 1971. 20

[HCD$^+$05] Jiawei Han, Yixin Chen, Guozhu Dong, Jian Pei, Benjamin W. Wah, Jianyong Wang, and Y. Dora Cai. Stream cube: An architecture for multi-dimensional analysis of data streams. *Distributed and Parallel Databases*, 18(2):173–197, 2005. 50

[HK92] Lars W. Hagen and Andrew B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 11(9):1074–1085, 1992. 56

[Hof99a] Thomas Hofmann. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In Thomas Dean, editor, *IJCAI*, pages 682–687. Morgan Kaufmann, 1999. 60

[Hof99b] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999. 13, 50, 51, 59, 64

[HSS10] Derek Hansen, Ben Shneiderman, and Marc A. Smith. *Analyzing Social Media Networks with NodeXL*. Morgan Kaufmann, 2010/08/27/ 2010. 42

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. 56

[Inm96] William H. Inmon. *Building the data warehouse*. Wiley computer publishing. Wiley, New York, NY [u.a.], 2. ed edition, 1996. 24, 27

[JMN93] Ellis L. Johnson, Anuj Mehrotra, and George L. Nemhauser. Min-cut clustering. *Math. Program.*, 62:133–151, 1993. 56

[Kel97] S. Kelly. *Data Warehousing in Action*. Wiley, 1997. 37

[Kim96] Ralph Kimball. *The Data Warehouse Toolkit*. John Wiley, 1996. 26, 31

[KL70] B.W. Kernighan and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. *The Bell Systems Technical Journal*, 49(2), 1970. 56

[KLN08] B. Karrer, E. Levina, and M. Newman. Robustness of community structure in networks. *Physical Review E*, 77(4), 2008. 56

[Kra96] David Krackhardt. Social networks and the liability of newness for managers. *Journal of Organizational Behavior*, 3:159–173, 1996. 41, 42

[LDH$^+$08] Cindy Xide Lin, Bolin Ding, Jiawei Han, Feida Zhu, and Bo Zhao. Text cube: Computing ir measures for multidimensional text database analysis. In *ICDM*, pages 905–910. IEEE Computer Society, 2008. 53

[LKH$^+$08] Eric Lo, Ben Kao, Wai-Shing Ho, Sau Dan Lee, Chun Kit Chui, and David W. Cheung. Olap on sequence data. In *Proceedings of the 2008*

*ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 649–660, New York, NY, USA, 2008. ACM. 50

[LT10] Kristen LeFevre and Evimaria Terzi. Grass: Graph structure summarization. In *SDM*, pages 454–465. SIAM, 2010. 54

[May33] E. Mayo. *The Human Problems of an Industrial Civilization*. Mac Millan, New York, 1933. 41

[MC99] Brennan Niamh O'Higgins Eleanor Mac Canna, Leo. National networks of corporate power: An irish perspective. *Journal of Management and Governance*, 2(4):355–377, 1999. 42

[MCZZ08] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web*, pages 101–110, New York, NY, USA, 2008. ACM. 51

[Mit74] J. C. Mitchell. Social networks. *Annual Review of Anthropology*, 3:279–299, 1974. 42

[MLSZ06] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web*, pages 533–542, New York, NY, USA, 2006. ACM. 51, 60

[MLW$^+$07] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, pages 171–180, New York, NY, USA, 2007. ACM. 60

[MM10] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM. 52

[Mor37] J. L. Moreno. Sociometry in relation to other social sciences. *Sociometry*, 1:206–219, 1937. 42

[MSZ07] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499, New York, NY, USA, 2007. ACM. 50, 51

[MWN$^+$09] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 880–889, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. 50

[Nag06] S. Nagabhushana. *Data Warehousing Olap And Data Mining*. New Age International (P) Limited, 2006. 33

[NAXC08] Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 542–550, New York, NY, USA, 2008. ACM. 50

[NBW06] Mark Newman, Albert-Laszlo Barabasi, and Duncan J. Watts. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, Princeton, NJ, USA, 2006. 56

[Neg04] Solomon Negash. Business intelligence. *Communications of the Association for Information Systems*, 13(1):177–195, 2004. 20

[New03] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. 56

[New06] M E Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, June 2006. 55, 56

[New10] M. Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010. 15, 127

[New11] M. E. J. Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31, December 2011. 55, 56

[NG04] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69(026113), 2004. 43, 56

[NP03] M.E.J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(026121), 2003. 55

[NRS08] Saket Navlakha, Rajeev Rastogi, and Nisheeth Shrivastava. Graph summarization with bounded error. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 419–432, New York, NY, USA, 2008. ACM. 54

[NSK09] Saket Navlakha, Michael C. Schatz, and Carl Kingsford. Revealing biological modules via graph summarization. *Journal of Computational Biology*, 16(2):253–264, 2009. 54

[NWS02] M. E. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 1(Suppl 1):2566–2572, February 2002. 45

[PDFV05] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005. 56

[PJ99] Torben Bach Pedersen and Christian S. Jensen. Multidimensional data modeling for complex data. In *ICDE*, pages 336–345, 1999. 32, 132

[PJD99] Torben B. Pedersen, Christian S. Jensen, and Curtis E. Dyreson. Extending Practical Pre-Aggregation in On-Line Analytical Processing. In *The VLDB Journal*, pages 663–674, 1999. 32, 110, 118, 132, 134

[POM09] M. A. Porter, J.-P. Onnela, and P. J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, 2009. 56

[QCT$^+$08] Yan Qi, K. Selçuk Candan, Jun'ichi Tatemura, Songting Chen, and Fenglin Liao. Supporting olap operations over imperfectly integrated taxonomies. In Jason Tsong-Li Wang, editor, *SIGMOD Conference*, pages 875–888. ACM, 2008. 50

[QZY$^+$11] Qiang Qu, Feida Zhu, Xifeng Yan, Jiawei Han, Philip S. Yu, and Hongyan Li. Efficient topological olap on information networks. In Jeffrey Xu Yu, Myoung-Ho Kim, and Rainer Unland, editors, *DASFAA (1)*, volume 6587 of *Lecture Notes in Computer Science*, pages 389–403. Springer, 2011. 54, 157

[Rap50] A. Rapoport. Outline of a mathematical theory of peck right. *Biometrics*, 6:330–341, 1950. 42

[RB08] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008. 56

[RBEP59] A.R. Radcliffe-Brown and E.E. Evans-Pritchard. *Structure and Function in Primitive Society: Essays and Addresses.* Cohen & West, 1959. 42

[RHMGM09] Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 54–63, New York, NY, USA, 2009. ACM. 51

[RS90a] Maurizio Rafanelli and Arie Shoshani. Storm: A statistical object representation model. In *In Proc. of SSDBM*, pages 14–29, 1990. 14, 107

[RS90b] Maurizio Rafanelli and Arie Shoshani. Storm: A statistical object representation model. In Zbigniew Michalewicz, editor, *Statistical and Scientific Database Management, 5th International Conference SSDBM, Charlotte, NC, USA, April 3-5, 1990, Proccedings*, volume 420 of *Lecture Notes in Computer Science*, pages 14–29. Springer, 1990. 86

[RZGSS04] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages

487–494, Arlington, Virginia, United States, 2004. AUAI Press. 95, 113, 139

[SBSR08] Alkis Simitsis, Akanksha Baid, Yannis Sismanis, and Berthold Reinwald. Multidimensional content exploration. *PVLDB*, 1(1):660–671, 2008. 52

[Sco00] John Scott. *Social Network Analysis: A Handbook*. Sage Publications, second. edition, 2000. 56

[SK10] A. Shollo and K. Kautz. *Towards an Understanding of Business Intelligence*. 2010. 21

[SKO$^+$07] J. Saramäki, M. Kivelä, J.P. Onnela, K. Kaski, and J. Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007. 45

[SSRZG04] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004. 50, 60

[Str43] Claude Levi Strauss. The social use of kinship terms among brazilian indians. *American Anthropologist*, 45(3):398–409, jul 1943. 42

[SZ10] Devavrat Shah and Tauhid Zaman. Community detection in networks: The leader-follower algorithm. *CoRR*, abs/1011.0774, 2010. 56

[TALS06] Efraim Turban, Jay E Aronson, Ting-Peng Liang, and Ramesh Sharda. *Decision Support and Business Intelligence Systems*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006. 20

[TG04a] Joana Trajkova and Susan Gauch. Improving ontology-based user profiles. In Christian Fluhr, Gregory Grefenstette, and W. Bruce Croft, editors, *RIAO*, pages 380–390. CID, 2004. 61, 70

[TG04b] Joana Trajkova and Susan Gauch. Improving ontology-based user profiles. In *Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, RIAO '04, pages 380–390, Paris, France, France, 2004. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. 118

[THP08] Yuanyuan Tian, Richard A. Hankins, and Jignesh M. Patel. Efficient aggregation for graph summarization. In Jason Tsong-Li Wang, editor, *SIGMOD Conference*, pages 567–580. ACM, 2008. 50, 54, 157

[TKMP11] Amanda L. Traud, Eric D. Kelsic, Peter J. Mucha, and Mason A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.*, 53(3):526–543, August 2011. 56

[TM08] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 111–120, New York, NY, USA, 2008. ACM. 51

[vdARS05] Wil M. P. van der Aalst, Hajo A. Reijers, and Minseok Song. Discovering social networks from event logs. *Computer Supported Collaborative Work*, 14(6):549–593, 2005. 42

[W.78] Wolfe Alvin W. The rise of network thinking in anthropology. *Social Networks*, 1(1):53–64, 1978. 42

[War37] W.L. Warner. *A black civilization: a social study of an Australian tribe*. A Black Civilization: A Social Study of an Australian Tribe. Harper & Brothers, 1937. 41

[Wat06] Martin Wattenberg. Visual exploration of multivariate graphs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 811–819, New York, NY, USA, 2006. ACM. 54

[WC06] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM. 51

[WF94] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994. 40

[WL41] W.L. Warner and P.S. Lunt. *The Social Life of a Modern Community*. Number vol. 1 in The social life of a modern community. Yale University Press, 1941. 41

[WLJH10a] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM. 51

[WLJH10b] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *WSDM*, pages 261–270. ACM, 2010. 77

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10, 1998. 45

[WSR07] Ping Wu, Yannis Sismanis, and Berthold Reinwald. Towards keyword-driven analytical processing. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 617–628, New York, NY, USA, 2007. ACM. 52

[WW90] R.J. Wilson and J.J. Watkins. *Graphs: an introductory approach : a first course in discrete mathematics*. Wiley, 1990. 40

[WZHS07a] Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 784–793, New York, NY, USA, 2007. ACM. 51

[WZHS07b] Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 784–793, New York, NY, USA, 2007. ACM. 60

[ZBZ$^+$08] Ding Zhou, Jiang Bian, Shuyi Zheng, Giles Lee, and Hongyuan Zha. Exploring social annotations for information retrieval. In *Proceedings of the 17th International World Wide Web Conference*, Beijing, Peking, 2008. 51

[ZCY$^+$08a] Feida Zhu, Chen Chen, Xifeng Yan, Jiawei Han, and Philip S Yu. Graph OLAP: Towards Online Analytical Processing on Graphs. In *Proc. 2008*

*Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008.*, December 2008. 15, 127

[ZCY$^+$08b] Feida Zhu, Chen Chen, Xifeng Yan, Jiawei Han, and Philip S Yu. Graph OLAP: Towards Online Analytical Processing on Graphs. In *Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008.*, December 2008. 53

[ZCY09] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow.*, 2(1):718–729, August 2009. 54

[ZHPL12] Jing Zhang, Xiaoguang Hong, Zhaohui Peng, and Qingzhong Li. Nested-cube: Towards online analytical processing on information-enhanced multidimensional network. In Zhifeng Bao, Yunjun Gao, Yu Gu, Longjiang Guo, Yingshu Li, Jiaheng Lu, Zujie Ren, Chaokun Wang, and Xiao Zhang, editors, *WAIM Workshops*, volume 7419 of *Lecture Notes in Computer Science*, pages 128–139. Springer, 2012. 15, 127

[ZJW$^+$11] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag. 51

[ZLXH11a] Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han. Graph cube: on warehousing and olap multidimensional networks. In *SIGMOD Conference*, pages 853–864, 2011. 15, 127, 147

[ZLXH11b] Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han. Graph cube: On warehousing and olap multidimensional networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 853–864, New York, NY, USA, 2011. ACM. 55, 154, 157

[ZLZ11] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, May 2011. 56

[ZTP10] Ning Zhang, Yuanyuan Tian, and Jignesh M. Patel. Discovery-driven graph summarization. In Feifei Li, Mirella M. Moro, Shahram Ghandeharizadeh, Jayant R. Haritsa, Gerhard Weikum, Michael J. Carey, Fabio Casati, Edward Y. Chang, Ioana Manolescu, Sharad Mehrotra, Umeshwar Dayal, and Vassilis J. Tsotras, editors, *ICDE*, pages 880–891. IEEE, 2010. 54

[ZZH09a] Duo Zhang, ChengXiang Zhai, and Jiawei Han. Topic cube: Topic modeling for olap on multidimensional text databases. In *SDM*, pages 1123–1134. SIAM, 2009. 53

[ZZH$^{+}$09b] Duo Zhang, ChengXiang Zhai, Jiawei Han, Ashok Srivastava, and Nikunj Oza. Topic modeling for olap on multidimensional text databases: Topic cube and its applications. *Stat. Anal. Data Min.*, 2(56):378–395, December 2009. 53

[ZZH11] Duo Zhang, ChengXiang Zhai, and Jiawei Han. Mitexcube: Microtextcluster cube for online analysis of text cells. In Ashok N. Srivastava, Nitesh V. Chawla, and Amal Shehan Perera, editors, *CIDU*, pages 204–218. NASA Ames Research Center, 2011. 53

[ZZH13] Duo Zhang, ChengXiang Zhai, and Jiawei Han. Mitexcube: Microtextcluster cube for online analysis of text cells and its applications. *Statistical Analysis and Data Mining*, 6(3):243–259, 2013. 53