Soc. Netw. Anal. Min. (2017)7:5
DOI 10.1007/s13278-017-0423-8

REVIEW ARTICLE

Data warehouse design approaches from social media: review and comparison

Imen Moalla[1,2] · Ahlem Nabli[1,3] · Lotfi Bouzguenda[1,4] · Mohamed Hammami[1,5]

Received: 17 May 2016 / Revised: 31 December 2016 / Accepted: 22 January 2017

Abstract With the rise of social media in our life, several decision makers have worked on these networks to make better decisions. In order to benefit from the data issued from these media, many researchers focused on helping companies understand how to perform a social media competitive analysis and transform these data into knowledge for decision makers. A high number of users interact at any time on different ways in social media such as by expressing their opinions about products, services or transaction related to the organization which can prove very helpful for making better projections. In this paper, we provide a literature review on data warehouse design approaches from social media. More precisely, we start by introducing the main concepts of data warehouse and social media. We also propose two classes of data warehouse design approaches from social media (behavior analysis and integration of sentiment analysis in data warehouse schema) and expose for each one the most representative existing works. Afterward, we propose a comparative study of the existing works.

Keywords Business intelligence · Data warehouse · OLAP · Social media

1 Introduction

Business intelligence refers to the set of devices, the querying tools and the data analysis used to drive a company and help it in the decision making. The establishment of decision-making databases has several objectives, such as the collection, reliability, storage, transformation and the distribution of information in support of decisional analysis.

Therefore, in order to take advantage of these new technologies, it becomes fundamental to transform the information of the company into a real capital and give it the ways to exploit it in order to facilitate the decision-making process. Indeed, the decision maker should chose between several options to provide solution to a given problem or execute an action or a project with the objectives pursued by the organization. In this context, a novel area known under the name of Social Business Intelligence (SBI) appeared, which refers to the discipline that aims at combining corporate data with user generated content (UGC) to let decision makers analyze and improve their business based on the trends and moods perceived from the environment (Gallinucci et al. 2015).

In the recent years, social media have played a strategic role in the life of many companies. It has created a set of links that connect customers to the companies. Each link

✉ Imen Moalla
imen.moalla@hotmail.fr

Ahlem Nabli
ahlem.nabli@fsegs.rnu.tn

Lotfi Bouzguenda
Lotfi.Bouzguenda@isims.rnu.tn

Mohamed Hammami
mohamed.hammami@fss.rnu.tn

[1] Multimedia, InfoRmation systems and Advanced Computing Laboratory, MIR@CL, Sfax, Tunisia

[2] University of Tunis El Manar, Faculty of Sciences of Tunis, Tunis, Tunisia

[3] Al-Baha University, Faculty of Computer Science and Information Technology, Al-Baha, Kingdom of Saudi Arabia

[4] University of Sfax, Higher Institute of Computer Science and Multimedia of Sfax, Sfax, Tunisia

[5] University of Sfax, Faculty of Sciences of Sfax, Sfax, Tunisia

Published online: 07 February 2017

provides large amounts of data that enable companies to gain a massive competitive edge.

In a study found by statista in November 2015,[1] the focus was on social media audience and on the most popular networks worldwide: "The network leader Facebook was the first social medium to surpass 1 billion registered accounts. The tenth-ranked micro blogging network Twitter has over 288 million monthly active accounts."

Social media are becoming an integral part of life as a way help us to find out about our family, keep up with friends or colleagues and contribute to online debates. Currently, with our mobile devices, it is easy to quickly check up social media one or more times per day. In addition, social media have considerably increased the participation, interactions and sharing between the Web users.

Moreover, social media are characterized by vast, unstructured and dynamic data. These characteristics represent challenges for company to use, analyze and store these data. The large volume of such data can be hard to manage and store without advanced platforms. For this reason, many companies decided to use the efficient platforms of data warehouse to deal with data from social media which can prove very helpful in making projections that lead to successful strategic decisions. This latter have become very important for the management of modern enterprises.

The volume of the data created everywhere, every day and every second on social media should be analyzed and manipulated by systems that can provide reliable and fast access for analyzing the large amounts of data. Among these systems is the online analytical processing (OLAP) system which provides means for decision makers to execute complex queries and involve aggregations through interactive and consistent interfaces to display a wide variety of information. The advantages of OLAP for the analysis in social media are expressed in the storage of the amount of data in the cube, in easy access to relevant data and in executing complex query at different hierarchies' levels. The traditional OLAP analysis cannot be used as they are to analyze social networks. In this context, many research works are interested to this problematic of research and proposed new operators to manage these new kinds of data (Zaho et al. 2011; Rehman 2015).

The data from social media can be analyzed in various ways and have given birth to a novel area of data analysis, namely social media analysis (Kaplan and Haenlein 2010). This area had a great importance for the scientific community. Collecting these data and performing analyses on social media data streams are one of the main challenges of big data because very interesting business goals could be achieved, such as: addressing marketing strategies, profiling people tastes, targeting advertisements (Cuzzocrea et al. 2015).

Given the emergence of these social media, several decision makers aim at using the abundant information generated by these media to improve their decisions. In this context, many companies use the data warehouse to collect their own data as well as that of their competitors. Several decision makers have worked on these networks to get extra information about the company to make better decision. Therefore, social media are an ideal ground for the adjustment of the decision makers with more precise forecasts.

The importance of social media for the decision-making process has attracted the attention of several researchers. Nevertheless, we found few studies that deal with the problem of data warehouse from social media. Therefore, we focused on these approaches and we classified them into two major classes: those addressing user's behavior analysis and those treating the integration of sentiment analysis in data warehouse schema. For this purpose, the main goal of this paper is to provide a literature review on the existing approaches addressing the data warehouse design from social media.

This paper is organized as follows. Section 2 presents the main concepts of data warehouse and social media. Section 3 exposes the existing approaches to deal with behavior analysis. Section 4 presents approaches that integrate the sentiment analysis in data warehouse schema. Section 5 gives a comparative study of the presented approaches and underlines their limitation. Finally, Sect. 6 concludes the paper with an outline of the main perspective in this work.

2 Basic concepts

Recently, the data warehouse has gone toward the exploitation of data from the Web, in particular the social media. Therefore, in this section, we start with the description of the fundamental concepts for both data warehouse and social media.

2.1 Data warehouse concepts

Data warehouse is a centrally managed and integrated database containing data from the operational sources in an organization. A data warehouse is designed for the query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. The data

[1] http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users.

warehouse can be created or updated at any time, with minimum disruption to operational systems.

According to (Inmon 1996), *a data warehouse is a* ***subject-oriented, integrated, time-variant, and non volatile*** *collection of data in support of management's decision making process.*

The data in a data warehouse are organized according to the multidimensional model. The objective is to enable analysts to have an idea about the data that support and assist them in their decision making. In this context, the multidimensional modeling gave rise to fundamental concepts: such as fact, measure, dimension and hierarchy (Fig. 1) (Ravat et al. 2001; Nabli et al. 2005).

- Fact and measure concepts

The fact models the subjects, the events or the phenomena that the decision makers of the organization wish to analyze. Every fact is characterized by one or several measures representing the analyzed indicators. A measure is a numerical property of a fact which describes a quantitative attribute relevant to the analysis (Kimball 1996; Ravat et al. 2001).

- Dimension and Hierarchy concepts

Dimension is described by the attributes corresponding to the information that makes the measures of activity vary, i.e., it is the axis of the analysis.

The dimension is organized in hierarchies to enable analysis of the measures at various levels of detail. The hierarchies of dimension define the structures for aggregating measures. Every hierarchy in a dimension specifies a particular way of grouping the members of the dimension (Abello et al. 2001).

- Multidimensional Schema

A DW is characterized by its multidimensional schema (MS) which can be either a star schema analyzing a single fact examined according to the axis of analysis (dimensions) or a constellation schema gathering several facts with shared dimensions (Kimball 1996; Ravat et al. 2001). Each schema belongs to one specific application domain. We use multidimensional schema as a generic term for both star and constellation.

In the literature, several multidimensional models were proposed. In what follows, we present a classification of these models based on the conceptual and logical design models.

2.2 Conceptual design models

The main goal of a conceptual design modeling is to develop a formal, complete, abstract design based on the user's requirements (Phipps and Davis 2002). In this context, several models have been proposed to model the conceptual schema of DW. Generally, the models proposed in the literature are classified into three types: extended entity relationship model, object-oriented multidimensional model and ad hoc formalism.

- *Extended Entity Relationship Model* Several extensions of the model E/R were proposed to model the DW schema. We cite, among others, StarER (Tryfona et al. 1999) and MultiDimER (Malinowski and Zimányi 2005). Every extension recommends a graphical representation to model the concepts from the multidimensional modeling.
- *Object-Oriented Multidimensional Model (OOMD)* Unified modeling Language (UML) has been widely accepted as a standard object-oriented modeling language for software design. OOMD modeling approach is based on UML. In OOMD model, dimensions and facts are represented by dimension classes and fact classes (Trujillo and Palomar 1998; Lujan-Mora et al. 2002).
- *Ad hoc formalism* The personalized models represent the second types of modeling methods proposed in the literature. The dimensional fact model (DFM) was proposed by Golfarelli et al. (1998) and then extended by Rizzi (2007). The dimensional fact model is a collection of tree structured fact schemas whose elements are facts, attributes, dimensions and hierarchies.

These formalisms can be used to construct at conceptual-level star, snowflake and constellation schemas.

- *Star schema* The most common modeling paradigm is the star schema in which the data warehouse contains a fact table that refers to a set of dimension tables. Each dimension table contains attributes (strong or weak), including its primary key that establishes the link with the fact table.
- *Snowflake schema* The snowflake schema is the variant of the star schema model, which consists in normalizing some dimensions of the star model in hierarchies. The fact is preserved and the dimensions are fragmented according to the hierarchy of the parameters.
- *Constellation schema* This schema consists in merging several star schemas that use common dimension. A constellation model thus includes many facts and shared dimensions.

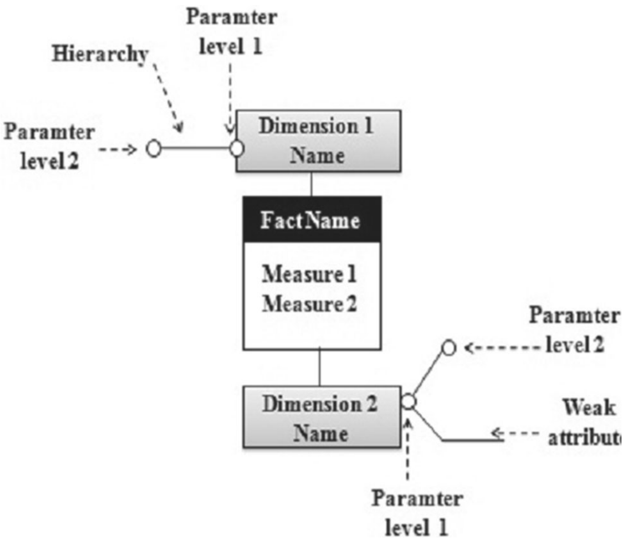

Fig. 1 Multidimensional concepts

2.3 Logical design models

A logical data model describes the data in as much detail, without regard to how they will be physical implemented in the database. These models involve the definition of structures that enable an efficient access to information. The implementation and manipulation of DW are directly obtained from cube. The concepts defined in a multidimensional structure were presented by relational schema adapted to OLAP applications. In the literature, these models are classified into three fields.

- *Relational OLAP (ROLAP)* performs dynamic multidimensional analysis of data stored in a relational database, where the base data and dimension tables are stored as relational tables.
- *Multidimensional OLAP (MOLAP)* stores data in a multidimensional cube. This enables to optimize data access time and to reduce query response time.
- *Hybrid OLAP (HOLAP)* combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP

2.4 Social media concepts

The Web 2.0 has brought the evolution of social media. It is increasingly used by Internet users and affects a very large audience. This development makes social media a leading market, which cannot be ignored by companies.

2.4.1 Social media definition

In the last decade, social media have become one of the most powerful sources of news updating, online collaboration, networking, viral marketing and entertainment. Social media are interactive platforms where content is created, distributed and shared by Web users. They include popular networking Web sites, like Facebook and Twitter. They involve blogging and forums and any aspect of an interactive presence which enables users to create and exchange content where people talk, share information and discuss over a particular blog post, news article or events.

The terms of social media and social network are used everyday. We saw to be more or less the same. In fact, social media include social networking, blogs, forums and platforms. In other words, social network is only a part of the social media.[2]

2.5 Types of social media

There are several types of social media, where each one has features and different purposes. However, many researchers have different classification about types of social media sites (Shankar and Hollinger 2007; Kaplan and Haenlein 2010; Akar and Topçu 2011) that emerges as a problem when realizing a study on social media. In this context, (Kaplan et Haenlein 2010) classified social media sites into these six types based on social presence, media richness and self-presentation, and self-disclosure. Thus, they have got six forms of social media, which are: (a) collaborative projects (e.g., Wikipedia), (b) blogs (e.g., Wordpress.com, Blogger.com), (c) content communities (e.g., YouTube), (d) social networking sites (e.g., Facebook), (e) virtual game worlds (e.g., World of Warcraft), and (f) virtual communities (e.g., SecondLife)). In 2011, Akar et Topçu add to this classification the microbloggings type such as Twitter. Shankar et Hollinger (2007) classified these new media into three groups: importun (Internet advertising, product placement in video games or advergames et m-commerce), non-importun (Internet advertising, social networking sites, podcasting, buzz or viral marketing) and user-generated (blogs, video site, ratings/recommendations and summary).

Based on the different types of social media proposed in the literature, we have recapitulated the classification of the most common forms of social media as follows:

- *Social network* refers to all the Web sites enabling to build a network of friends or professional knowledge and providing their member with some tools and interfaces, such as interaction, presentation and communication. The most known social networks are Facebook, *MySpace and* LinkedIn

[2] http://blog.kinoa.com/2013/08/05/reseaux-sociaux-et-medias-sociaux-quelle-difference/#comments. http://www.reseaux-professionnels.fr/comprendre-ce-sont-les-medias-sociaux/.

- *Microblogs* Microblogging is a form of blogging where the amount of information that can be shared per author is either enforced to be very short, or just typically very short. It allows users to publish it by a variety of means, including text messaging, instant messaging, email, digital audio or the Web. The most common examples of microblogs are Twitter, Google Buzz, Jaiku and Tumblr.
- *Blogs* A blog (short for weblog) is a personal online journal frequently updated and intended for general public consumption. Blogs generally represent the personality, experience, observations and opinions of the author. The most common examples of blogging platforms are OverBlog, canalblog and blogspot.
- *Forum* Online discussion groups in which participants with common interests can exchange open messages. Examples of forums are Doctissimo and macrumors.
- *Media sharing sites* a Web site that enables users to uploaded and store photographs, videos and audio that can be accessed from anywhere in the world. The most known sites are YouTube, Snapchat, dailymotion and flicker.

In the following, we will present a thorough survey of approaches to design a data warehouse from social media. We classified these approaches into two classes: *(1) behavior analysis and (2) integration of sentiment analysis in data warehouse schema.* Section 3 details approaches for behavior analysis. Then, Sect. 4 reviews approaches that propose the integration of sentiment analysis in data warehouse.

3 Behavior analysis approaches

The behavior analysis approaches focused on analyzing Web user's activities that can be made on social media (such as friendship creation, group creation, profile browsing and messaging) and also to view how Web users interact in front of a political or sports event in order to help the decision makers to discover new knowledge. Moreover, the behavior analysis is an ideal ground, to analyze the behavior of people using social media based on entered text and to develop innovative ideas fostering the discovery of the new knowledge of social media.

The purpose of this section is to study in depth the proposed existing approaches dealing with behavior analysis. More precisely, we examine these works of (Bringay et al. 2011; Liu et al. 2012; Rehman et al. 2012; Ben Kraiem et al. 2015; Cuzzocrea et al. 2015, 2016) which are the most representative of the state of the art that treat data warehouse.

3.1 Approach proposed by Bringay et al. (2011)

A multidimensional star model is proposed by Bringay et al. (2011) to analyze the big volume of Tweets. The authors suggested using information retrieval approaches; an adapted measure called TF-IDF adaptive (Grabs and Schek 2002) is used to classify the most significant words in the hierarchy level of the dimensions. The first step of this approach is to instantiate the multidimensional model of tweets to take into account the new dimensions which change from one context to another. In the second step, the authors proposed the measure TF-IDF adaptive which is used to identify the most significant words according to hierarchy level of the cube. The last step is to identify the context of a tweet; the researchers opted for word hierarchy to extract the context of the tweets. Finally, they opted for the use of the thesaurus MeSH (medical subject headings) in their case study that permits the search to be carried out at various levels of specificity. The objective of decision makers through this study is

to define and manipulate cubes of tweets. The advantage of this proposal is the classification of the most significant words in the hierarchy level of the dimensions by using TF-IDF adapted measure, but it is a partial approach at the modeling level (fixed multidimensional schema). Additionally, the schema definition rules of data warehouse are not present and the proposed model of a data warehouse is not adapted to the huge amount of social data. However, the authors focus their analysis on one type of social media.

3.2 Approach proposed by Liu et al. (2012)

In the same context, Liu et al. (2012) presented a text cube architecture designed to analyze the human social cultural behavior (HSCB) in the social media. Therefore, they demonstrated that the text cube architecture also supports the development of prediction models, which can help to develop more effective strategies for making decisions based on social media data. In the first step, the authors proposed a framework for HSCB feature analysis. This framework is organized around three layers: the generic language function layer used to automatically generate the linguistic features, the feature selection layer exploited to select the linguistic features with reference to psychological and sociological expectations and the HSCB dimension layer employed to evaluate a given dimension HSCB. In the second step, the authors designed a star schema to store the linguistic features extracted from different HSCB dimensions. Finally, in the last step, they used the techniques of data mining to select the important linguistic features and build predictive models for each HSCB dimension using the selected features. The resulting social cube was used in an attempt to predict violence during the uprising in Egypt 2011. The goal of the decision makers in this work is to define an architecture of text cube designed to organize social media data in multiple dimensions and hierarchies. The benefits of this approach are to perform several types of analyses and build prediction models, but the authors did not propose a theoretical approach for opinion analysis and focused their study on one type of social media. However, the proposed model of a data warehouse is not adapted to the huge amount of social data.

3.3 Approach proposed by Rehman et al. (2012)

Rehman et al. (2012) presented a multidimensional model to analyze the streams for twitter. An architecture was proposed to extract tweets from Twitter and load them into a data warehouse. As a consequence, the data warehouse was created to analyze the user's behavior on Twitter during the event earthquake in Indonesia. In order to do this, the authors opted for the extended dimensional fact model (x-DFM) (Mansmann 2008) to represent the multidimensional model. Finally, the authors extended the cube data by integrating the language and the spam detection to analyze the contents of the text generated by user. After that, they transformed the cube model into a graph (set of nodes and edges) to treat all the elements of cube serving as potential input fields for the new knowledge discovery. In this paper, the objectives of the decision makers are to present a multilayered architecture for Twitter data warehouse system and define a multidimensional model to enable comprehensive analysis of massive data volumes generated by the social network Twitter. This paper tries to build an exhaustive cube for an OLAP analysis of tweets. Nevertheless, the schema definition rules of data warehouse are ignored. Besides, the proposed model of a data warehouse is not adapted to the huge volume of social data. Thus, the authors focused on one type of social media. We also noticed that the proposed approach is used only to analyze a specific event.

3.4 Approach proposed by Kraiem et al. (2015)

Kraiem et al. (2015) proposed a multidimensional model dedicated to the online analytical processing (OLAP) of data from tweets. Therefore, the authors focused their study on the content of a tweet. Consequently, they developed a model that takes into account the specificity of tweets data and the link between tweets and tweets responses. The researchers focused their analysis on the content of a tweet to model a multidimensional schema of tweets. In fact, they opted for the extended constellation model. After that, they transformed the constellation model into logical model R-OLAP by applying a set of rules. Finally, the authors developed a software prototype called tweet OLAP to study the evolution of Twitter accounts created per year and per language. Then, they studied the distribution of the users analyzed per source and date. The objective of the decision makers is to provide a generic multidimensional model dedicated to the OLAP of data generated by Twitter social medium. Nevertheless, the proposed model is not adaptable to the huge amount of social data. Since, the authors have focused only on one type of social media. Therefore, the proposed approach can be used only for ensuring a special treatment.

3.5 Approach proposed by Cuzzocrea et al.

Cuzzocrea et al. (2015) presented an initial study where they arranged the microblogs in a hierarchical structure of concepts. In other words, the authors integrated time-aware fuzzy formal concept analysis with OLAP. Then, Cuzzocrea et al. (2016) presented an extension of their work where they focus on the definition of multidimensional data model for the storage of tweet data to deal with the OLAP analysis. In order to do this, the authors started with the definition of the concepts of the multidimensional model. In this context, the dimensions are classified into two types, namely semantic dimension which was determined from Wikipedia knowledge base and the meta-data dimension which was derived from the information of the tweet (time and location). Moreover, the authors proposed a new measure which exploited the wikification service to analyze the textual content of tweets. After that, they presented the model of fuzzy extension of formal concept analysis (Ganter and Wille 1997) to address aggregation of textual data in the tweet cube. Then, the authors proposed a summarization algorithm to select the best tweets that represent the data of each cube according to OLAP dimensional fact model. The proposed approach is exemplified with a real case study of tweet streams collected during the Italian regional election. The goal of the decision makers is the integration of knowledge mining approaches (i.e., FFCA) and OLAP technology analyzing unstructured data exchanged in social media and enabling advanced analytics services.

4 Integration of sentiment analysis in data warehouse schema

The approaches that treat the integration of sentiment analysis in data warehouse schema analyzed the social, psychological, philosophical, behavior and perception of an individual person or a group of people about a product, policy, services and specific situations. This study can prove very helpful for the decision makers to design new marketing strategies, improve product features and identify which customers like or dislike particular product features.

The plus value of the sentiment analysis integration in the analysis process is to discover how people feel about a particular topic, events or products. Extracting and analyzing text generated by users in social media for sentiment will answer the following question: what do users say about this product?

The multidimensional data warehouse support a solution for the opinion mining for analyzing products, services, user's feeling or transaction related to the organization at any time which can prove very helpful for company in making projections that lead to successful strategic decisions and in improving every day the decision making.

Given the importance of public sentiments, the researchers are working to build a system to capture and evaluate opinions, emotions and feelings expressed in social media. The purpose of this section is to study in depth the proposed existing approaches of the integration of sentiment analysis in data warehouse schema. More precisely, we examine the studies of (Moya et al. 2011; Costa et al. 2012; Gallinucci et al. 2013; Rehman et al. 2013; Liu et al. 2013; Moalla and Nabli 2014; Walha et al. 2015) which are the most representative of the state of the art.

4.1 Approach proposed by Moya et al. (2011)

In 2011, Moya et al. presented an approach to integrate the sentiment data extracted from the Web into the corporate data warehouse. Consequently, a multidimensional model is trained to highlight this approach which is found in two parts: The corporate part of the data warehouse contains the database and the internal documents of company, and the sentiment part of the data warehouse was extracted from the Web by the use of opinion analysis techniques. In the first step of the proposed approach, (Moya et al. 2011) collected the opinions of customers from the Web in order to identify the products about which customers expressed their opinions. After that, they selected the potential features that have been modified by the opinion words. Then, they built a list of opinion words by the intersection of the adjectives of the two lists (Eguchi and Lavrenko 2006; Stone et al. 1966). The resulting list is verified manually, by adding some adverbs and verbs with context-independent polarity. So, they got a list of classified words either as a positive or negative opinion. The second step consists in classifying the potential features according to their importance, which depend on two factors: the function relevance and the feature frequency (Zhang et al. 2010). Finally, the authors proposed to calculate synonym group characteristics by using Jaccard distance function to take into account both the lexicon and overlapping word synonyms. However, in this work the schema definition rules of data warehouse are ignored. Additionally, the proposed model is not adaptable to the huge amount of social data.

4.2 Approach proposed by Costa et al. (2012)

Costa et al. (2012) proposed a business intelligence architecture, called OSNBIA (online social networks business intelligence architecture). It is based on the construction of a corporate data warehouse in order to integrate it into a business intelligence tool. This work focuses mainly on extracting data from Twitter and applies sentiment analysis to produce data warehouse. The proposed architecture is established as follows: Data extraction step consists in determining the necessary information for the decision makers. In fact, the authors used Twitter API to retrieve tweets containing the text "lenovo thinkpad." Then, in the cleaning step, they performed some operations to handle with the missing attributes and the appearance of NULL identifiers of the extracted data. After that, the researchers used the API suggested by Go et al. (2009) to identify the feelings of 58, 906 tweets. Moreover, they used the metrics degree to analyze the links. After that, they stored the tweets, which are correctly treated by the previous components of the architecture, in flat files. Finally, the researchers inserted these files into a data warehouse to analyze the feelings of tweets relative to sales performance. Once the data warehouse is generated, the authors used QlikView to develop a BI analysis application offering a bigger flexibility for data analysis. The disadvantage of this approach is the lack of data warehouse schema. However, the authors focused on one type of social media and presented partial approach for opinion analysis.

4.3 Approach proposed by Rehman et al. (2013)

Rehman et al. (2013) presented an extension of their previous work. This approach is based on the integration of opinion mining methods and knowledge discovery techniques into the data warehousing system, in order to deal with semi-structured and unstructured data from social media. In the first step, the authors identified the facts, measures, hierarchies and dimensions of the twitter data warehouse. In fact, they extracted the hierarchical structure of dimensions as aggregation paths for drill-down and roll-up operations. Moreover, for the other concepts, they used a specific algorithm to extract named entities, events and other semantic knowledge. In the second step, the authors exploited the opinion mining methods to make an exploratory and predictive analysis. In fact, they used two API for semantic enrichment Alchemy API services: AlchemyAPI[3] and OpenCalais.[4] First API is used for sentiment analysis, while the second is used for topic extraction and concept tagging for uniformity of results. In

[3] http://www.alchemyapi.com.

[4] http://www.opencalais.com.

addition, they developed a classifier to predict the popularity category of a tweet. The result of classifier takes the form of rule-based decision tree. In the last step, the researchers concentrated on the change in dimensions and hierarchies of the data warehouse. The ETL process (extraction, transformation and loading) is repeated each time when new data are uploaded in the data warehouse. For this purpose, they modified some concepts based on the proposals of slowly changing dimensions presented in (Kimball 1996) and changed the name and screen name of attributes by replacing the existing data with new ones. Nevertheless, this work is limited to the use of API available for analyzing opinions. However, the authors focused on one type of social media and proposed a model of data warehouse that is not adapted to the huge volume of social data.

4.4 Approach proposed by Gallinucci et al.

In 2013, Gallinucci et al. focused in their study on the social business intelligence, which enables to combine corporate data with the user-generated content (UGC) to help the decision makers to improve their company and aggregate subjects at different levels. Therefore, they proposed to model topic hierarchies in ROLAP systems called meta-stars. This approach is based on the combination of the traditional dimension tables and the navigation tables to deal with the dynamics of the subject area. Gollinucci et al. introduced the architecture for SBI (social business intelligence) that integrates both the corporate data and the sentiment data of the Web users (UGC). In the implementation of this architecture, the researchers manually defined the topics and roll-up relationships. After that, they presented a cube to analyze the sentiments expressed by the Web users. Furthermore, they defined a set of relationships between the topics in the hierarchy roll-up. In fact, they proposed to model topic hierarchies on ROLAP platforms combined with classical dimension tables and with recursive navigation tables. Then, they extended the obtained result by using meta-modeling called meta-stars. Therefore, the authors tested some OLAP queries to evaluate the performance of meta-stars against star schema. In fact, they noted that meta-stars are better in terms of space efficiency and query expressiveness and lower in terms of time. In Gallinucci et al. (2015) improved their work by extending their meta-stars model. Firstly, they took non-covering and non-strict hierarchies. Secondly, they dealt with the techniques of slowly changing topics and levels. Thirdly, they supported the semantics queries on topic hierarchies. Finally, they evaluated the meta-star approach proposed on wider set of tests. Nevertheless, these works are limited to the use of one type of social media. Additionally, the proposed model is not adaptable to the huge amount of social data.

4.5 Approach proposed by Liu et al. (2013)

Liu et al. (2013) presented an extension of their previous work (Liu et al. 2012) where they focused on integrating the sentiment analysis in the text cube. The particularity of this approach is to display the contents of the cube on a heat map and develop data mining models using the data taken from cubes. In the first stage, the authors organized the linguistic feature of sentiment in a data cube model. Accordingly, they designed a comprehensive framework to analyze the HSCB (human social cultural behavior) linguistic feature in order to extract emotional features. In addition, they chose to model a star schema to store the extracted linguistic features for different HSCB dimensions, including sentiments which are stored as measures in the fact table. In the second stage, the researchers collected tweets from the Washington D.C. region. Then, they performed an HSCB linguistic analysis of the tweets to extract sentiment measures, such as “positive emotion” and “negative emotion.” In fact, they added the possibility of posting the contents of the cube on a card of heat. This type of card enables the analyst to quickly visualize the areas that have the highest negative emotions on a given day or set of days. Finally, they developed data mining methods, such as: LIBSVM (Fan et al. 2005), REPTree and IBK (Witten and Frank 2005), to detect political events (Aha et al. 1991). Nevertheless, in this work the schema definition rules of data warehouse are ignored. Besides, the proposed model is not adaptable to the huge amount of social data. Since, the authors presented a partial approach for opinion analysis; this work is limited to the use of one type of social media.

4.6 Approach proposed by Moalla and Nabli (2014)

In 2014, Moalla and Nabli proposed a multidimensional schema from Facebook page in order to analyze the customers’ opinions. A real case study was developed to illustrate the proposed method and to confirm that the social network analysis can predict the success prospects of the products. This approach is organized around four steps: The first one is the data extraction, which consists in extracting data from a Facebook page. Therefore, the authors collected general information from the Facebook page, information about fans and information related to posts. The second step is the data analysis, in which the authors defined specific operations to select the relevant data and eliminate duplicate

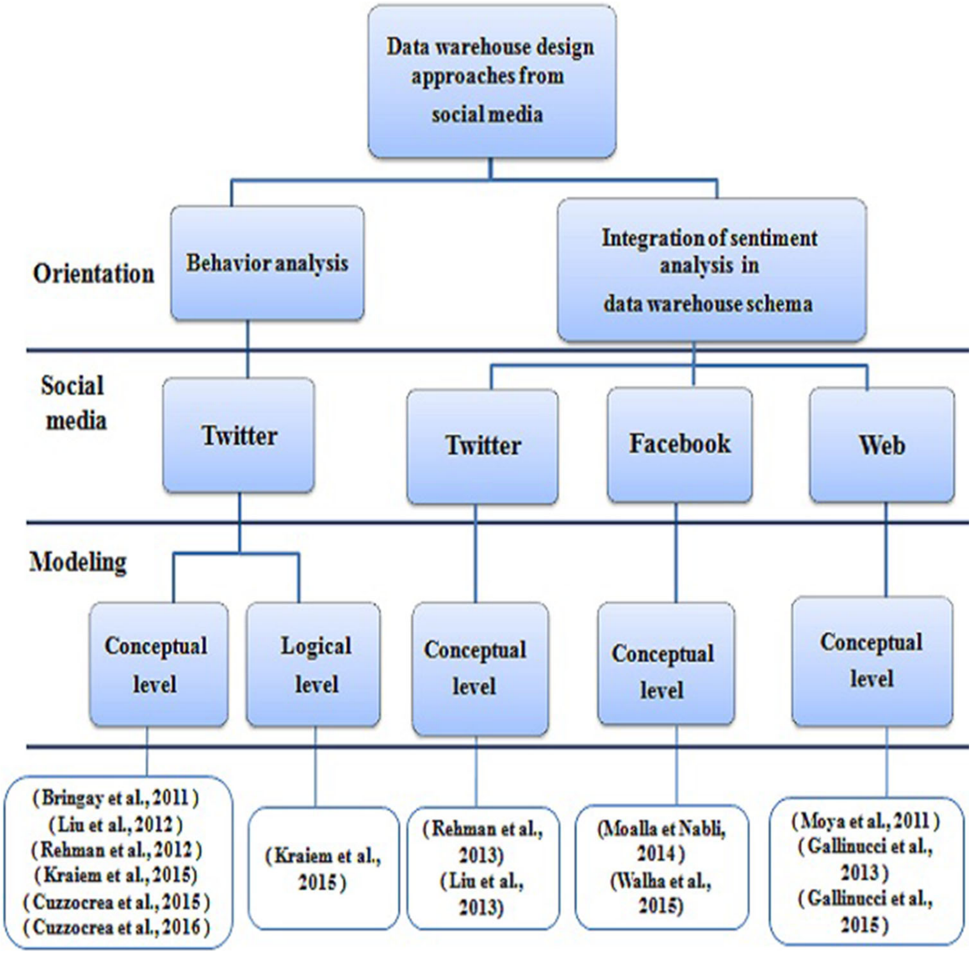

Fig. 2 Classification of data warehouse design approaches from social media

and inconsistency data. The third step implied a schema definition that consists in defining the multidimensional schema of the data mart from the extracted data. It consists in determining: facts, measures, dimensions and hierarchies. In fact, the authors opted for the X-DFM schema to model the data mart schema. Finally, the last step is an opinion analysis, which consists in helping decision makers to model strategies and access relevant information based on MDX query. However, the authors made a modest proposal for opinion analysis. Since the proposed model is not adaptable to the huge amount of social data, this work is limited to the use of one type of social media.

4.7 Approach proposed by Walha et al. (2015)

A rather similar approach is presented in the work of (Walha et al. 2015) where the authors proposed a new ETL processes modeling approach using BPMN standard. This approach integrates user opinions expressed by comments shared on the social network Facebook. It consists in the extraction of opinion data on Facebook pages, its preprocessing and loading into the DataWeB-house (DWB). The goal of this ETL is to detect both positive and negative comment polarities. For this purpose, the authors defined a new ETL design approach that integrates people's opinions to model extraction, transformation and loading processes. In the first step, called extraction step, the authors collected data about the general information of Facebook page, the post information and the actions associated with each post published in Facebook page. Then, they performed a transformation step that consists of cleaning and conforming the extracted data. In fact, they identified two dictionaries: opinion dictionary composed of opinion words and emoticons dictionary. Therefore, they proposed a lexicon-based method to compute for each comment message its sentiment score. In the last step, called loading step, the authors defined a DWB star schema in order to load the data resulted from transformation step into DWB multidimensional schema. The disadvantage of this approach is the use of one type of social media. Additionally, the proposed model is not adaptable to the huge amount of social data.

Table 1 A comparative study of data warehouse design approaches of behavior analysis

	Multilevel modeling			ETL process		Social media	Objective	Analysis process	Advantages	Drawbacks
	Conceptual	Logical	Modeling process (rules)	Tools used	Language					
Bringay et al. (2011)	Star schema	Not mentioned	Not Presented	Postgre SQL and Pentaho Mondrian	Not mentioned	Twitter	The definition of a multidimensional star model to analyze tweets by proposing measures relevant in the context of knowledge discovery	Classic OLAP operators	The use of hierarchies to manage words in the tweets to allow a contextualization in order to better understand the content	Partial approach at the modeling level (fixed DW schema)
Liu et al. (2012)	Star schema	Not mentioned	Presented	Not mentioned	Not mentioned	Twitter	The presentation of text cube architecture for analyzing the human social cultural behavior (HSCB)	Classic OLAP operators	Capability to perform several types of analysis and to build prediction models	Lack of a theoretical approach for opinion analysis
Rehman et al. (2012)	X-DFM	Not mentioned	Not Presented	BaseX and Microsoft SQL Serve	Not mentioned	Twitter	The proposition of an exhaustive cube for OLAP analysis of tweets	Classic OLAP operators	The proposed model can be used for solving any tasks based on aggregating or mining of data	Lack of schema definition rules
Kraiem et al. (2015)	Extended conceptual constellation schema	ROLAP	Presented	JAVA and ORACLE 10 g	Not mentioned	Twitter	The proposition of a multidimensional model dedicated to the OLAP of Twitter	Classic OLAP operators	The conceptual model takes into account the specificity of tweet and tweet response	The proposed model is not adaptable to the huge amount of social data
Cuzzocrea et al. (2016)	DFM	Not mentioned	Presented	Not mentioned	Not mentioned	Twitter	The integration of knowledge mining approaches (i.e., FFCA) with OLAP technology over multidimensional tweet streams for analyzing unstructured data exchanged in social media	Classic OLAP operators	Considering the temporal dimension via FFCA in the aggregation task	The proposed model is not adaptable to the huge amount of social data

Table 2 A comparative study for the approaches that integrate the sentiment analysis in data warehouse schema

	Multilevel modeling			ETL process		Social media	Objective	Analysis process	Advantages	Drawbacks
	Conceptual	Logical	Schema definition rules	Tools used	Language					
Moya et al. (2011)	Constellation schema	–	Not presented	SQL Server Business Intelligence Studio	Not mentioned	Web	The presentation of a multidimensional data model that integrates sentiment data extracted from customer opinion forums into the corporate data warehouse	Classic OLAP operators	The process of extracting sentiment data produces a semantically rich data set that enables complex queries	Lack of schema definition rules
Costa et al. (2012)	–	–	Not presented	Data Manager and ORACLE	Not mentioned	Twitter	Development of business intelligence application which integrates social network and sentiment analysis with user's decision-making processes	Classic OLAP operators	The feasibility of a business intelligence environment to support organizational and social media for a better interpretation of topics recorded in social networks	Lack of DW schema
Gallinucci et al. (2013)	Meta-star schema	–	Presented	Talend	Not mentioned	Web	Proposition of model that takes into account topic hierarchies on ROLAP platforms	–	The support of OLAP queries with increasing expressiveness and complexity, starting from queries using only static levels to end up with semantics-aware queries	Manual definition of topic and roll-up relationships
Liu et al. (2013)	Star schema	–	Not presented	Not Mentioned	Not mentioned	Twitter	The presentation of text cube for social media analysis, especially sentiment analysis	Classic OLAP operators	Display the contents of the cube on a heat map	Lack of schema definition rules
Rehman et al. (2013)	X-DFM	–	Not presented	BaseX and Microsoft SQL Server	Not mentioned	Twitter	The integration of opinion mining methods and knowledge discovery techniques into the data warehousing system, in order to perform multidimensional social media analysis	Classic OLAP operators	Capability of enabling OLAP to keep up with volatile data using the concepts of slowly changing dimensions to enable analysis of both the recent state of data and any of its previous states	Lack of a theoretical approach for opinion analysis
Moalla and Nabli (2014)	X-DFM	–	Presented	Microsoft SQL Server and XML	Not mentioned	Facebook	Proposition of multidimensional schema from Facebook page in order to analyze the customers' opinions	–	The conceptual model takes into account the specificity of Facebook page and opinions analysis	Modest proposal for opinion analysis

Table 2 continued

	Multilevel modeling			ETL process		Social media	Objective	Analysis process	Advantages	Drawbacks
	Conceptual	Logical	Schema definition rules	Tools used	Language					
Walha et al. (2015)	Star schema	–	–	Not mentioned	BPMN	Facebook	Proposition of an ETL design approach integrating user's opinion analysis, expressed on the popular social network Facebook	–	The definition of all the components of the ETL process	The proposed model is not adaptable to the huge amount of social data

5 A comparative study of the existing approaches

In the light of the above, we notice that social media analysis has drawn the attention of many researchers in the literature. The classification of presented approaches is shown in Fig. 2. However, we identified a few studies focusing on the multidimensional data modeling of data warehouse from the social media. Noteworthily, Table 1 reports a comparison between the approaches of behavior analysis. Table 2 presents a comparison between the approaches that integrate the sentiment analysis in data warehouse schema.

- *Multilevel modeling: also* called modeling level, this criterion indicates the model used for level modeling conceptual (star model, constellation model, snowflake model, DFM, x-DFM, etc.) and logical models (ROLAP, MOLAP, HOLAP). This criterion is also interested in the modeling process (rules) of multidimensional concepts if they are presented by the author.
- *ETL process* (extract–transform–load): This criterion mentions the tools used during the ETL process and the language used during the modeling process: BPMN or UML.
- *Social media* It presents the social media used.
- *Objective* This criterion presents the general idea of the work.
- *Analysis process* It shows whether the approach uses classic OLAP operators (drill down, roll up) or defines specific operators.
- *Particularities* This criterion describes the advantages and the drawbacks of each approach.

Given, the *multilevel modeling*, we notice that most of the approaches described above explicitly addressed only one modeling level. (Rehman et al. 2012, 2013; Gallinucci et al. 2013; Moalla and Nabli 2014; Cuzzocrea et al. 2016;….) specify only conceptual modeling. Except (Kraiem et al. 2015) presented both logical and conceptual modeling. Moreover, we find that the *modeling process (rules)* is proposed only by Liu et al. (2012), Kraiem et al. (2015), Gallinucci et al. (2013), Moalla and Nabli (2014), Cuzzocrea et al. (2016).

Reaching the *social media* criterion, in this context, we note that most approaches are based on a single social media Twitter or Web as a data source. Except (Moalla and Nabli 2014; Walha et al. 2015) are used Facebook social medium as a data source.

On the other hand, the *ETL process*, which is an essential stage in the construction of data warehouse, was modeled only by (Walha et al. 2015) using the BPMN language. Moreover, the other researchers did not explicitly define the different function of the ETL process. We may also conclude that all the proposed approaches used different tools in the ETL process.

Regarding the *analysis process* criterion, we notice that most of the approaches used the classical OLAP operators. We also find lack of definition of specific operators.

Consequently in this study, we can conclude that very few works have focused on the data warehouse design from social media. These studies used commercial tools or the result of some algorithms to solve this problem. However, we notice that there is a lack of a theoretical approach for the opinion analysis. However, we find that all the proposed approaches did not support the big data tools, such as Hadoop, Mapreduce, NoSQL databases to deal with the large amount of data.

Finally, in order to overcome the shortcomings seen, we will propose a new approach for the construction of a data warehouse from social media in order to deal with the opinion analysis. As a first step, we try to find a multidimensional model (conceptual, logical) that should take into account the aspects of heterogeneous social media. In the second step, we will define an approach for the sentiments analysis which will be integrated into the multidimensional model to analyze the user's opinions expressed in social media.

6 Conclusion

We presented in this paper a literature review on data warehouse design approaches from social media. More precisely, we presented the main concepts of data warehouse and social media; we also identified two classes of approaches to deal with social data warehouse design approaches. These approaches are classified into two categories, namely behavior analysis and integration of sentiment analysis in data warehouse schema. Subsequently, we proposed a comparative study of the existing approaches according to criteria that we have identified. These criteria cover the multilevel modeling, the inclusion of ETL process and the usage of different social media. Although the presented contributions are powerful, they suffer of many drawbacks. These can be resumed in the lack of schema definition rules, lack of a theoretical approach for opinion analysis, the limit in the use of one type of social media and the only use of a relational for storage. As future work, we aim to define a multidimensional model (conceptual and logical) that integrates the most popular social media. Also, we intend to present the modeling process of each model (conceptual and logical) and we plan to implement the ETL process.

Acknowledgements This work is dedicated to the soul of the author of this paper Dr. Lotfi Bouzguenda who passed away a few months ago. We are very grateful for his help, his advice and his prestigious remarks. May God have his soul in blessed mercy.

References

Abello A, Samios J, Saltor F (2001) Understandding analysis dimensions in a multidimensional object-oriented model. In: Proceedings of the international workshop on design and management of data warehouses (DMDW), Interlaken, Switzerland, pp 4-1–4-9

Aha D, Kibler D, Albert M (1991) Instance-based learning algorithms. J. Mach Learn 6(1):37–66

Akar E, Topçu B (2011) An examination of the factors influencing consumers' attitudes toward social media marketing. J Internet Commer 10(1):35–67

Bringay S, Béchet N, Bouillot F, Poncelet P, Roche M, Teisseire M (2011) Towards an on-line analysis of tweets processing. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6861 LNCS, pp 154–161

Costa PR, Souza FF, Times VC, Benevenuto F (2012) Towards integrating online social networks and business intelligence. In: IADIS international conference on web based communities and social media

Cuzzocrea A, De Maio C, Fenza G (2015) Towards OLAP analysis of multidimensional tweet streams, 69–73. OLAP analysis of multidimensional tweet streams for supporting advanced analytics. Sac 2016, pp 992–999. doi:10.1145/2811222.2811233

Cuzzocrea A, De Maio C, Fenza G, Loia V, Parente M (2016). OLAP analysis of multidimensional tweet streams for supporting advanced analytics. In: Proceedings of the 31st annual ACM symposium on applied computing. ACM, pp. 992–999

Eguchi K, Lavrenko V (2006) Sentiment retrieval using generative models. In: Proceedings of the 2006 conference on empirical methods in natural language processing, pp 345–354

Fan R-E, Chen P-H, Lin CJ (2005) Working set selection using the second order information for training SVM. J Mach Learn Res 6:1889–1918

Gallinucci E, Golfarelli M, Rizzi S (2013) Meta-stars: multidimensional modeling for social business intelligence. In: Proceedings of the sixteenth international workshop on Data warehousing and OLAP. ACM, pp 11–18

Gallinucci E, Golfarelli M, Rizzi S (2015) Advanced topic modeling for social business intelligence. Inf Syst 53:87–106

Ganter B, Wille R (1997) Formal concept analysis: mathematical foundations, vol 1. Springer, New York

Go A, Bhayani R, Huang L (2009) Twitter Sentiment classification using distant supervision. Processing, CS224N Project Report, Stanford, vol 1, p 12

Golfarelli M, Maio D, Rizzi S (1998) The dimensional fact model: a conceptual model for data warehouses. Int J Coop Inf Syst 7:215–247

Grabs T, Schek H.-J. (2002) ETH Zürich at INEX: Flexible Information Retrieval from XML with PowerDB-XML. Xml With Powerdb-Xml. Inex Workshop, 2002, pp 141–148

Inmon WH (1996) Building the data warehouse. Wiley, New York

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

Kimball R (1996) The data warehouse toolkit practical techniques for building dimensional data warehouses. Wiley, New York

Kraiem MB, Feki J, Khrouf K, Ravat F, Teste O (2015) Modeling and OLAPing social media: the case of Twitter. J Social Netw Anal Inf Syst 5(1):47

Liu X, Tang K, Hancock J, Han J, Song M, Xu R, Pokorny B (2012). SocialCube: A text cube framework for analyzing social media data. In: International conference on IEEE social informatics (SocialInformatics), pp 252–259

Liu X, Tang K, Hancock J, Han J, Song M, Xu R, Pokorny B (2013) A text cube approach to human, social and cultural behavior in the Twitter stream. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7812 LNCS, pp 321–330

Lujan-Mora S, Trujillo J, Song I (2002) Multidimensional modeling with UML Package diagrams, In: 21st international conference on conceptual modeling (ER2002), 2002

Malinowski E, Zimányi E (2005) Spatial hierarchies and topological relationships in the spatial MultiDimER model. In: British national conference on databases. Springer, Berlin, pp. 17–28

Mansmann S (2008) Extending the OLAP technology to handle non-conventional and complex data. Ph.D. dissertation, University of Konstanz, Department of Computer and Information Science Germany

Moalla I, Nabli A (2014) Towards data mart building from social network for opinion analysis. In: International conference on intelligent data engineering and automated learning IDEAL14, pp 295–302

Moya LG, Kudama S, Cabo MJA, Llavori RB (2011) Integrating web feed opinions into a corporate data warehouse. In: Proceedings of the 2nd international workshop on business intelligence and the WEB-BEWEB'11, p 20

Nabli, A, Jamel F, Faiez G (2005) Automatic construction of multidimensional schema from olap requirements. In: 3rd ACS/IEEE international conference on computer systems and applications, 2005 2005 (Figure 1), pp 97–103. doi:10.1109/AICCSA.2005.1387025

Phipps C, Davis K (2002). Automating data warehouse conceptual schema design and evaluation, DMDW'02

Ravat F, Teste O, Zurfluh G (2001) Modélisation multidimensionnelle des systèmes décisionnels. In EGC, pp 201–212

Rehman NU (2015) Extending the OLAP technology for social media analysis (Doctoral dissertation)

Rehman NU, Mansmann S, Weiler A, Scholl MH (2012) Building a data warehouse for twitter stream exploration. In: Proceedings of the 2012 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM, pp 1341–1348

Rehman NU, Weiler A, Scholl M (2013) OLAPing social media: the case of Twitter. In: International conference on advances in social networks analysis and mining, ASONAM2013, pp 1139–1146

Rizzi S (2007) Conceptual modeling solutions for the data warehouse. Data Warehouses and OLAP: Concepts, Architectures and Solutions, pp 1–26

Shankar V, Hollinger M (2007) Online and mobile advertising: current scenario, emerging trends, and future directions. Market Sci Inst 31(3):206–207

Stone PJ, Dunphy DC, Smith MS, Ogilvie DM (1966) The general inquirer: a computer approach to content analysis. MIT Press, Cambridge

Trujillo J, Palomar M (1998) An object oriented approach to multidimensional database conceptual modeling (OOMD). In: Proceeding 1st international workshop on data warehousing and OLAP (DOLAP98)

Tryfona N, Busborg F, Borch Christiansen JG (1999) starER: a conceptual model for data warehouse design. In: Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP. ACM. pp 3–8

Walha A, Ghozzi F, Gargouri F (2015) ETL design toward social network opinion analysis. In: Computer and information science 2015. Springer International Publishing, p 235–249

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, The Morgan Kaufmann Series in Data Management Systems

Zhang L, Bing L, Lim SH, O'Brien-Strain E (2010) Extracting and ranking product features in opinion documents. In: Proceedings of the 23rd international conference on computational linguistics: association for computational linguistics. Posters, pp 1462–1470

Zhao P, Li X, Xin D, Han J (2011) Graph cube: on warehousing and OLAP multidimensional networks. In: Proceedings of the 2011 ACM SIGMOD international conference on management of data. ACM, pp 853–864