Towards Data Mart Building from Social Network for Opinion Analysis

Imen Moalla and Ahlem Nabli

Sfax University, Faculty of Science, Road Sokra Km 3 BP 1171, 3000 Tunisia
MIRACL Laboratory, Sfax, Tunisia
imen.moalla@hotmail.fr,
ahlem.nabli@fsegs.rnu.tn

Abstract. In the recent years, social networks have played a strategic role in the lives of many companies. Therefore several decision makers have worked on these networks for making better decisions. Furthermore, the increased interaction between social networks and web users has lead many companies to use a data warehouse to collect information about their fans. This paper deals with a multidimensional schema construction from unstructured data extracted from social network. This construction is carried out from Facebook page in order to analyze customers' opinions. A real case study has been developed to illustrate the proposed method and confirming that he social network analysis can predict chance of the success of products.

Keywords: Data warehouse, Data Mart, Social Networks analysis, Opinion Analysis, Schema Design.

1 Introduction

Recently, social networks and online communities, such as Twitter and Facebook, have become a powerful source of knowledge being daily accessed by millions of people [2]. Social networks help people to follow breaking news, to keep up with friends or colleagues and to contribute to online debates. Seen the emergence of these social networks, several decision maker's aim to use the abundant information generated by these networks to improve their decisions. Indeed, the main advantage of social networks is to enable companies to operate a new visibility shape on the Internet at a lower cost. Its a way to make them known to different publics and collect information on customers prospective. However, business intelligence provides a solution for companies, which allows to collect, consolidate, model and restore the data of a company offering help to decision makers. The popularity of social networks and high volume of user generated content, especially subjective content caused the heavy demand to adopt sentiment analysis in business applications. Sentiment Analysis helps decision makers to get customer opinion in real-time [3]. This real-time information helps them to design new marketing strategies, improve product features and can predict chances of product failure.

E. Corchado et al. (Eds.): IDEAL 2014, LNCS 8669, pp. 295–302, 2014.

The social network analysis problem took a big importance from the scientific community. In the literature, the data analysis approaches from social networks highlight two areas of research: community analysis and opinions analysis [10]. So, data warehouse can be used to analyze the opinion of customers based on many dimensions. In this context and in order to benefit from existing information in social networks, we are interested in using these data as a source feeding a data Mart schema. In fact, we propose a heuristic to design data Mart schema from Facebook page in order to analyze customers' opinions.

The remainder of this paper is organized as follows: section 2 details the related works. Section 3 describes our method Data Mart construction from the social network Facebook. Opinion analysis is performed in section 4. Finally, section 5 draws conclusion and future works.

2 Related Works

The importance of social networks for decision making process is highlighted in several studies such as [9,10,4,1,7]. For example, Rehman and al. [9] proposed an architecture to extract tweets from Twitter and load them to a data warehouse. Multidimensional cubes resulting can be used for analyzing user's behavior on twitter during event earthquake in Indonesia. But in this work, the authors present schema without detailing how this schema is determined. However, in [10] the authors proposed a Business Intelligence architecture, called OSNBIA (Online Social Networks Business Intelligence Architecture). Therefore, the authors did not explain how the data warehouse schema is performed. In 2011 Kazienko and al. [4] proposed a multidimensional model for social network that allows to capture data about activities and interactions between web users. The multidimensional model of the social network can be used To make a new relationships and to analyze different communication ways. Also, in this study the researchers did not perform an experimental phase to validate their process. A data warehouse schema is proposed by [1] to analyze the big volume of data tweets. The authors suggested using information retrieval approaches to classify the most significant words in the hierarchy level of the dimensions. Moya and al.[7] presented an approach to integrate sentiment data extracted from the web into the corporate data warehouse. In 2013 Mansmann and al.[6] proposed to model data warehouse elements from the dynamic and semi-structured data of Twitter. In addition, they propose to extend the resulting model by including dynamic categories and hierarchies discovered from DM and semantic enrichment methods. In [2] the researchers presented a data analysis framework to discover groups of similar twitter messages posted by users about an event.

The comparative study shows that the data warehouse design methods [9,10,1] are based on the social network twitter except the work of [7] that used an approach from Web. In addition, the multidimensional concepts are generally defined by the designer without presenting how these concepts are determined [7,10,1]. Furthermore all methods operate on the texts except [4] operates on the link. Based on these lakes, we propose a method composed of set steps, applied to Facebook page, which based on heuristics to generate the data Mart schema.

3 Schema Design from Facebook Page

This paper deals with a method to generate a data Mart schema from Facebook page. As depicted in Fig.1, our method encloses four steps: (1) *Data extraction* which involves collecting information about Facebook page, fans and posts. (2) *Data analysis* that includes specific operations to select relevant data and to eliminate inconsistency. (3) *Schema modeling* which suggests multidimensional concepts. (4) *Loading step* that incorporates procedures for charging modeled schema from the second step. Since the data mart is constructed, the decision maker can analyze the opinions of fan page based on MDX query. Before performing our method we start by introducing Facebook page.

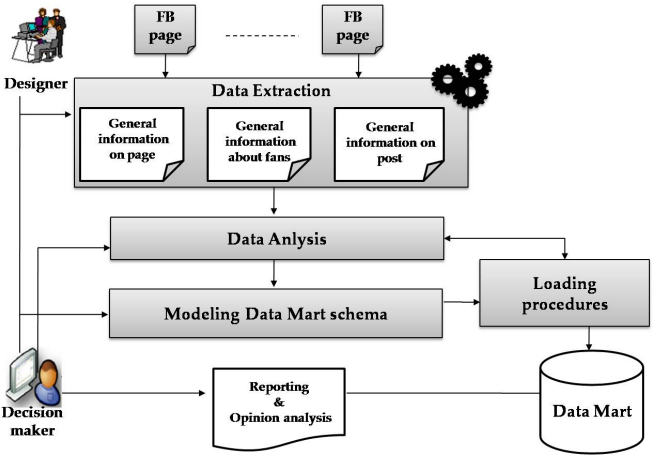

Fig. 1. The proposed method Process

Facebook Page. (FP) is a sort of website integrated into the Facebook community which allows the companies to interact with customers by starting discussions within comments and photos. It has three types of information: general information of the page, information about fans and about posts. In the following sections we give an idea about data extraction and data analysis. After that, we explain the extraction of multidimensional concepts in section 3. Finally, we expose reporting based on opinion analysis query.

3.1 Data Extraction

The first step of our method is to extract data from a FP, so we collect general information of FP, information about fans and information related to posts. In fact, this extraction consists of creating an application enabling the access to the API graph in order to select the access permissions to the data. The general information of FP are numerous, here we elicit some of them: Id (page

identifier), name (page name), user name (FP username), category (FP type), number of fans (number of fans in FP), site (link website), etc. We recognize that a post can be photos, links, videos or statutes. So, each one is concerned by some information. For example, a post is defined by the ID (post identifier), name (post name), type(the type of post), created date (date of creation of post), number of likes (number of fans liked post), number of shares (number of fans shared post) etc. We noticed a difference between the data collected on the fans who have a friendship with the admin of FP and who're without friendship with the admin. The data available for the first category are the data that we select their access tokens and which are visible to fans friends. However, for fans not friends we have only public data.

3.2 Data Analysis

Since data are collected we proceed to analyze these data. The analysis concerns the selection of relevant data, the elimination of duplicate and inconsistency data. The decision maker is implied during this step in order to contribute in the selection of relevant data.

Select Relevant Data: The main needs of the decision maker are to determine the client's opinions' and satisfaction towards their products. In fact, client's opinions provide valuable information for companies. First, they help to understand how their products and services are perceived [5]. They yield clues about costumers satisfaction and expectations that can be used to determine their current and future needs and preferences. Second, they may be helpfull in understanding what product dimensions or attributes are important to each clients. Third, client's opinions on the products offered by competitors provide essential information to accomplish a successful competitor analysis.

Eliminating Duplicates and Inconsistency Data: From studied case, we found that, when a post is published many times in the page, Facebook assigns for each post an ID. Therefore, the same post has different ID. In this case, analyzes on the same post are distributed over multiple IDs instead of the same ID. To solve this problem, we decided to group the different ID of the same post published several times and assigned to them the same identifier. Another problem was detected when computing the number of positive comments. Indeed, the information extracted from the comments can be relative of fans or admin; this made us a need to delete the comments of the admin page.

3.3 Modeling Schema

This phase consists in defining the multidimensional schema of the data Mart from the extracted data. This definition consists in determining: facts, measures, dimensions and hierarchies. This definition is based on heuristics adapted from [8].

Heuristics of Fact and Measure Determination: Fact to be analyzed describes the daily activity of events performed in a Facebook page which allow capturing the opinions of fans. Thats why we define the **Fact Opinion**.

The opinions expressed by fans on posts can be positive or negative. The positive opinions on the posts are determined through the following heuristics:

Share-Measure: if the post has an important number of shares then we conclude that there are a number of fans impressed by the product. This allows us to define the measure Number of shares called *Number-shares*.

Likes-Measure: when the post has an important number of likes this means that there are a significant number of fans who liked the product. This allows us to define the measure named Number of likes called *Number-likes*.

Comments-Measure: every post can have a set of comments. Comment can be positive or negative. It is qualified as positive if it contains positive words, which means that the fan is interested by the product. In our context, we determined all the positive words through which we can conclude that the comment is positive. These latter give us to define the measure *Number-comment-positive*. The negative opinions on posts are captured by actions realized by fans on the posts. These actions are deducted when the fans click mask or signal on a post of the page in their newsfeed. So, the measure *Number-comment-negative* is proposed. All the determined measures are depicted in Fig.3.

Heuristics of Dimensions Determination : The extraction of the dimensions is based on a type of object named base object BO which completes the definition of fact. A base object answers the questions: "who", "what", "when" and "where". Every BO defined an axis of analysis which can interest the decision maker. From a FP, the objects which answer these questions for the fact opinion are:

Who declare the opinion ? – Fans declare their opinions.
Where are posted the opinions ? – The opinions are published in page.
What are the opinions ? – Posts encapsulate the opinions.
When the posts have been shared ? – The post has its publication date.

Based on these four questions we obtain four dimensions: the dimensions page, post, fans and date. Due to the lack of space we present only the parameter extraction of Page dimension.

Table 1. Parameters of the page dimension

Tags describing the page	Type concepts	Name of multidimensional concept
⟨`id`⟩	identifier	Page-ID
⟨`name`⟩	weak attribute	name
⟨`username`⟩	weak attribute	user name
⟨`website`⟩	weak attribute	website
⟨`phone`⟩	weak attribute	phone
⟨`category`⟩	Level 2 parameter	category
⟨`location`⟩	Level 2 parameter	location

Extraction of Dimension Parameters': The dimension Page is determined from the general information of Facebook page. Table 1 presents all the attributes composing the dimension Page (column 1). Based on this information we determine the type of each attribute (weak attribute or parameter) in column 2. Then we define the equivalent multidimensional concept in column 3. Fig.2 depicts a graphical representation of the page and date dimensions.

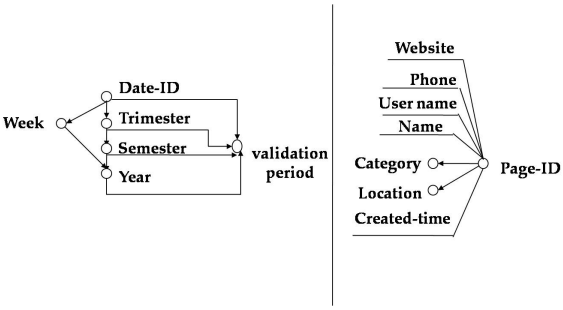

Fig. 2. Hierarchies of Page and Date dimensions

Notes that, a FP can be removed from the Net, so we define a parameter called validation period in which a page is active as shown in Fig 2.

From the previous steps, we generate the data Mart schema modeled in X-DFM structure as shown in Fig 3.

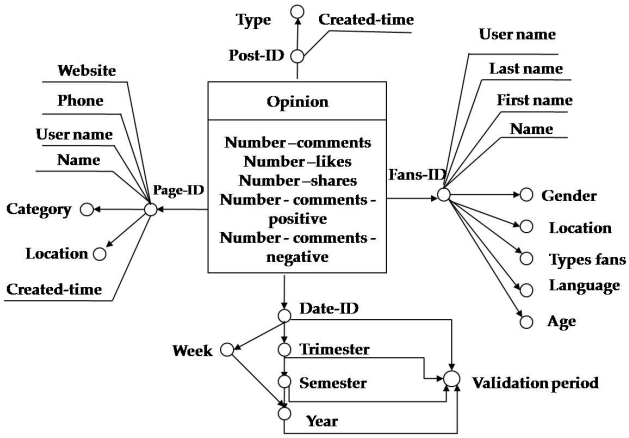

Fig. 3. Data Mart schema generated from Facebook page

The structure of the schema is a graph centered at the fact type node (opinion) which includes all measures (Number-comments, Number-likes, Number-comments-positive, and Number-comments-negative). The fact is relied by four

dimensions (Page, Fans, Date, and Post) each of them contents a node key attribute and others node represents dimension parameters (user name, first name, last name, phone, website, created-time).

4 Opinion Analysis

Opinion analysis is a crucial step for communities. The goal is to help decision makers to model strategies and access to relevant information. From our Data Mart constructed we analyzed the number of likes, shares, and the number of positive and negative comments by week, trimester and semester. The decision makers can use of the following queries:

- Analyze the number of comments for posts by week, trimester and semester.
- Analyze the number of likes for posts by week, trimester and semester.
- Analyze the number of shares for posts by week, trimester and semester.
- Analyze the number of positive and negative comments for posts by week, trimester and semester.

We performed a comparison between the measurements taken from the page and actual sales of a company. As shown in Fig .4, we note that the curve for the sharing follows the same shape as the real company sales. So, we may infer that social networks can predict sales trends of a company based on the number of shares.

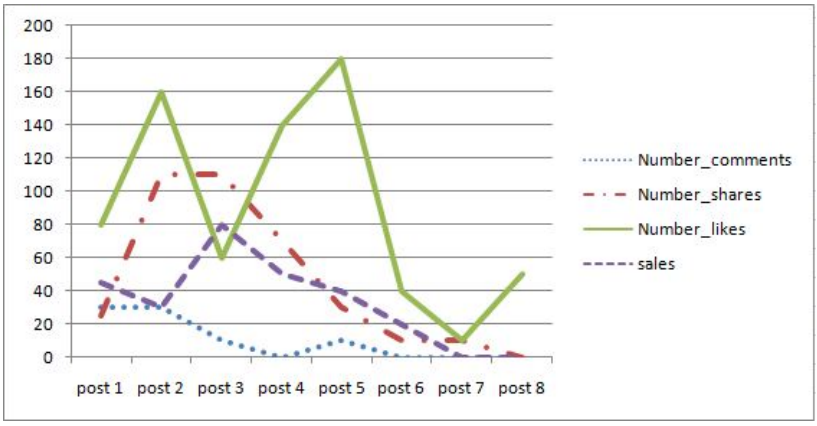

Fig. 4. Best posts for the first semester of 2013

5 Conclusions and Future Work

We have presented in this paper a method to build a data Mart from a social network for opinion analysis. This method uses Facebook page as data source. it's invloved on four steps: the first is data extraction. The second step is the

data analysis that consists of making specific operations to select the relevant data and the third step is schema definition that defines the multidimensional concepts of the data Mart. The analysis that we made, we give us an overview on the actual case of the company's sales. Our future orientations consist of studying the possibility of using ontology to better analyses fans opinions. Thus, we will enhance our method to support other types of social networks.

References

1. Bringay, S., Béchet, N., Bouillot, F., Poncelet, P., Roche, M., Teisseire, M.: Towards an On-Line Analysis of Tweets Processing. In: Hameurlain, A., Liddle, S.W., Schewe, K.-D., Zhou, X. (eds.) DEXA 2011, Part II. LNCS, vol. 6861, pp. 154–161. Springer, Heidelberg (2011)
2. Baralis, E., Cerquitelli, T., Chiusano, S., Grimaudo, L., Xiao, X.: Analysis of Twitter Data Using a Multiple-level Clustering Strategy. In: Cuzzocrea, A., Maabout, S. (eds.) MEDI 2013. LNCS, vol. 8216, pp. 13–24. Springer, Heidelberg (2013)
3. Jaganadh, G.: Opinion Mining and Sentiment Analysis, CSI Communications (2012)
4. Kazienko, P., Kukla, E., Musial, K., Kajdanowicz, T., Bródka, P., Gaworecki, J.: A generic model for a multidimensional temporal social network. In: Yonazi, J.J., Sedoyeka, E., Ariwa, E., El-Qawasmeh, E. (eds.) ICeND 2011. CCIS, vol. 171, pp. 1–14. Springer, Heidelberg (2011)
5. Laura, P., Jorge, C.A.: Sentiment Analysis in Business Intelligence: A survey. In: Customer Relationship Management and the Social and Semantic Web. IGI-Global (2011)
6. Mansmann, S., Rehman, N.U., Weiler, A., Scholl, M.H.: Discovering OLAP dimensions in semi-structured data. Information Systems (2013)
7. Moya, L.G., Kudama, S., Aramburu, J., Llavori, R.B.: Integrating Web Feed Opinions into a Corporate Data Warehouse. In: Proceedings of the 2nd International Workshop on Business Intelligence and the WEB BEWEB, New York (2011)
8. Nabli, A.: Thesis: Proposal of an approach of help to the design automated schema of data warehouse (2010)
9. Rehman, N.U., Mansmann, S., Weiler, A., Scholl, M.H.: Building a data warehouse for twitter stream exploration. IEEE/ACM (2012)
10. Santos, P., Souza, F., Times, V., Benevenuto, F.: Towards integrating Online Social Networks and Business Intelligence. In: Proceedings of the IADIS International Conference on Web Based Communities and Social Media (2012)