Towards Data Warehouse Schema Design from Social Networks
Dynamic Discovery of Multidimensional Concepts

Rania Yangui[1], Ahlem Nabli[2] and Faiez Gargouri[1]
[1]*Institute of Computer Science and Multimedia, Sfax University, BP 1030, Sfax, Tunisia*
[2]*Faculty of Sciences, Sfax University, BP 1171, Sfax, Tunisia*

Keywords: Data Warehouse, Social Network, Ontology, Flexibility, Scalability, Dynamicity, Clustering.

Abstract: This research work is conducting as part of the project BWEC (Business for Women in Women of Emerging Country) that aims to improve the socio-economic situation of handicraft women by providing true technological means. In fact, since few years, the Web has been transformed into an exchange platform where users have become the main suppliers of information through social media. User-generated data are usually rich and thus need to be analyzed to enhance decision. The storage and the centralization of these data in a data warehouse (DW) are highly required. Nevertheless, the growing complexity and volumes of the data to be analyzed impose new requirements on DW.
In order to address these issues, in this paper, we propose four stages methodology to define a DW schema from social networks. Firstly we design the initial DW schema based on the existing approaches. Secondly, we apply a set of transformation rules to prepare the creation of the NOSQL(Not Only SQL) data warehouse. Then, based on user's requirement, clustering of social networks profiling data will be performed which allows the dynamic discovery of multidimensional concepts. Finally, the enrichment of the NoSQL DW schema by the discovered MC will be realized to ensure the DW schema evolution.

1 INTRODUCTION

Because of the rapid development of the Internet, the availability of various types of data has increased tremendously. In fact, the creation of many sites (such as Facebook, LinkedIn, Twitter, etc.) and forums has made the users perceive the Web as a place in which they exchange ideas, opinions as well as contents. However, if these tools make the sharing and collaboration between the users easy, they may cause new challenges concerning the relevant exploitation of the produced data. Analyzing, understanding, as well as managing the huge volumes of complex data produced from the social networks (SN) broach a paramount importance and draw the attention of many researchers. In fact, the companies expect to acquire important information from this data so as to improve their marketing. This is the case of handicraft women in the BWEC[1] project. This latter aims at improving the socio-economic situation of these women by providing true technological means in concordance with the women habits and the technical context of their countries. The stages proposed to accomplish the purpose of this project are summarized in (Figure 1).

The use of online SN can play a very important role in the social and economic development of this population. For instance, SN can be used not only to promote their products but also to enhance the brand value. It can also be used to strengthen consumer relations as well as to improve the quality of the services and products through receiving feedback from the market itself. Thus, the establishment of a decision-making process has proved to be necessary.

Originally developed for the needs of support decision, data warehouses (DW) have proven to be an adequate solution to a variety of applications and fields. DW contains all the information integrated from heterogeneous sources into multidimensional schema to enhance data access for both analysis and decision making.

Many methodologies can be used to create a DW (Nabli, 2013) which are demand-driven methodologies, data-driven methodologies and Mixed (demand/data driven) methodologies. These warehousing methodologies have shown their efficiency when

[1]Towards a new Manner to use Affordable Technologies and Social Networks to Improve Business for Women in Emerging Countries (http://projetat.cerist.dz/)

Yangui R., Nabli A. and Gargouri F..
Towards Data Warehouse Schema Design from Social Networks - Dynamic Discovery of Multidimensional Concepts.
DOI: 10.5220/0005383903380345
In *Proceedings of the 17th International Conference on Enterprise Information Systems* (ICEIS-2015), pages 338-345
ISBN: 978-989-758-096-3

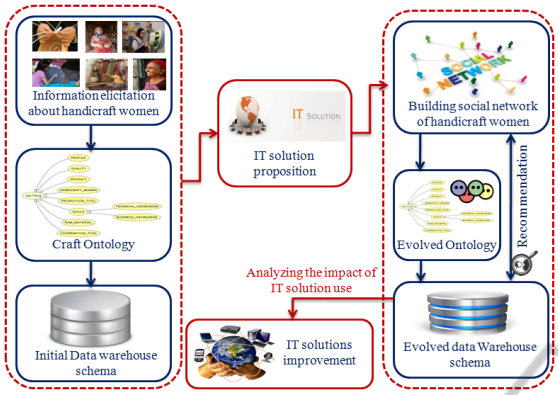

Figure 1: Project Steps.

dealing with structured data. Its require measures and dimensions of a DW to be known at the design stage (N. U. Rehman and Scholl, 2012). However, the growing complexity and volumes of the data to be analyzed impose new requirements on data warehousing systems.

As argued by many works (N. U. Rehman and Scholl, 2012) (E. Gallinucci and Rizzi, 2013), existing warehousing methodologies cannot be successfully applied to handle the above-mentioned challenges. So, a remarkable effort must be made to integrate the huge amount of complex data from SN and make them accessible to OLAP (On-Line Analytical Processing) and analyzing tools.

Our challenges can be summarized by the following questions: how to determine multidimensional concepts (MC) from social data? How to adjust the DW to take into account these new concepts? How to overcome the problems caused by the heterogeneity of the data? How to take into account the dynamicity of user-generated data as well as the users needs in an existing DW?

This paper tends to answer the mentioned questions by proposing a methodology to design a DW schema via content-based dynamic discovery of MC from social networks generated data. The proposed methodology will be applied on real case study. This case study looks at handicraft women social networks.

Our contributions focus on the DW schema evolving to manage data about handicraft women from SN. In this context, DW schema modeling is a complex task which involves knowledge of SN structure and familiarity with DW technologies. What makes this task even more challenging is the fact that social data continue to grow rapidly and analytical requirements change over time. Given these challenges, the semi-automatic modeling of the warehouse schema is required.

This paper is organized as follows. Section 2 reviews some related works concerning the DW schema creation from SN and the semi-automatic generation of DW schema. Section 3 overviews our methodology. Sections 4 describes the initial DW schema creation. Sections 5 details the dynamic discovery of MC. Section 6 concludes the paper and draws future research directions.

2 RELATED WORKS

Data generated from SN are usually rich and need to be analyzed to support decision. The storage and the centralization of these data in a DW are highly required. However as mentioned above, warehousing methods have shown some limitations when dealing with SN data. Challenged by these limitations, DW researchers put tremendous efforts into extending its.

2.1 Multidimensional Modeling from Social Networks

In the literature, many researchers have proposed approaches for semi-automatic modeling of DW.

For example, a data warehousing architecture for analyzing large data sets at Facebook, used for friend recommendation, is described by (A. Thusoo and Liu, 2010). The authors describe the challenges of implementing a DW for data intensive Facebook applications and present a number of contributed open-source technologies for warehousing petabytes of data. These include Scribe[2], Hadoop[3] and Hive[4] which together form the cornerstones of the log collection, storage and analytic infrastructure at Facebook.

The paper mainly focuses on flexibility and scalability issues. However no insight on the underlying models is given and no dynamic discovery of MC is mentioned.

(P. Kazienko and Brdka, 2011) focus on developing a conceptual generic model for multidimensional SN that allows capturing information about different types of activities and interactions between users. It also represents the dynamics of users behavior. The proposed model encompasses information about the different relations and groups that exist within a given relation layer and in a specific time window. The proposed model covers three main dimensions: relation layers, time windows and groups. Social groups are extracted by means of clustering methods. However,

[2]http://wiki.github.com/facebook/scribe
[3]http://wiki.apache.org/hadoop
[4]http://wiki.apache.org/hadoop/Hive

the proposed model doesnt manage the flexibility and the scalability of social networks data.

(N. U. Rehman and Scholl, 2012) provide a DW solution for hosting the public data stream of Twitter messaging. The authors enrich the multidimensional analysis of such data via content-driven discovery of dimensions and classifying hierarchies. In the first step, data mining algorithms are applied to cluster dimensional data. In the second step, the acquired classification is added as a new aggregation path to the respective dimension, leading to the third step of enabling this new aggregation path in OLAP queries. Nevertheless, this work is limited to the granularity level addition and ignores the other MC such as facts and dimensions. Moreover, the proposed model is inflexible and no scalable.

(E. Gallinucci and Rizzi, 2013) propose a methodology called meta-stars to model topic hierarchies in ROLAP systems. Its basic idea is to use meta-modeling coupled with navigation table and with traditional dimension tables. The navigation tables support hierarchy instances with different lengths and with non-leaf facts, and allow different roll-up semantics to be explicitly annotated. The meta-modeling enables hierarchy heterogeneity and dynamics to be accommodated. However, this work is based on a relational approach which presents limitations regarding to schema scalability.

(Moalla and Nabli, 2014) present a method to multidimensional schema construction from unstructured data extracted from SN. This construction is carried out from Facebook page in order to analyze the customers opinions. A real case study has been developed to illustrate the proposed method and to confirm that the SN analysis can predict the success prospects of the products. Nevertheless, the dynamic discovery of MC is not supported. The proposed model is not flexible and not adaptable to the huge amount of social data.

Based on the previous study, most of the works show no indication of the dynamic determination of MC seen the velocity of SN data. Also the DW schemas are generally fixed at design stage.

2.2 Dynamic Discovery of Multidimensional Concepts

Nowadays, we are experiencing a rapid growth of social structures supported by communication technologies and various Web-based services. Due to scale, complexity and dynamicity, user-generated data from SN are very difficult to store and analyze in terms of traditional data warehousing methods (N. U. Rehman and Scholl, 2012). To overcome these problems, many authors have worked on dynamic discovery of MC and have used data mining to build a DW.

In this context, (Usman and Pears, 2011) provide a methodology to design semi-automatically DWs schema with hierarchical clustering. This latter is used to perform a pre-processing on the data. After that, the system identifies both facts and dimensions into the clustered data.

Rehman proposes a system to dynamically build hierarchies based on data from Twitter (N. U. Rehman and Scholl, 2012). This paper has two Interests: a) The cube is built on original data which are the messages of users on a SN. b) Data mining is used to dynamically build hierarchies. Thanks to data mining, the categories of network users described in hierarchies are updated automatically. On the other hand, Ceci uses a hierarchical clustering to integrate continuous variables as dimensions in a DW schema (M. Ceci and Malerba, 2011). It discretizes a continuous dimension so that the user can perform operations on existing querying a cube: Roll-up and DrillDown.

As for the current work, (L. Sautot and Molin, 2014) propose using hierarchical agglomerative clustering with a metric that comes from ecological studies to build semi-automatically hierarchical dimensions in an OLAP cube. The authors perform a hierarchical clustering on heterogeneous data sets that contains qualitative and quantitative variables. They offer a prototypical automatic system which builds dimension for an OLAP cube and measure the performances of this system according to the number of clustered individuals and according to the number of variables used for clustering.

Table 1 highlights a summary of the literature review which is based on seven criteria (Concept M.: Conceptual Model, D. MC: multidimensional concepts, Methodology, SN: Social Network, Ontology, Flexibility, Scalability).

All the mentioned works present several interesting mining. It has been recognized that mining techniques such as Clustering can help in designing DW schema. That is why we adopt this orientation for the dynamic discovery of MC. However, no work has ever dealt with the semantic heterogeneity. Moreover, no work has ever followed a mixed approach (data/demand driven approach). Furthermore, it is worth noting that just one work has provided the scalability of the schema. At the same time, the heterogeneity and the growth of the social data need to be considered in order to properly retrieve needed data. The frequent arrival of new needs requires that the system should be adaptable to changes.

Based on the above discussion, there is a strong need of a significant methodology that allows a dy-

Table 1: Summary of the literature review.

Work	Concept M.	D. MC	Methodology	SN	Ontology	Flexibility	Scalability
(Thusoo, 2010)	-	No	Data driven	Yes	No	Yes	Yes
(Ceci, 2011)	-	Yes	Demand driven	No	No	No	No
(Usman, 2011)	-	Yes	Data driven	No	No	No	No
(Kazienko, 2011)	-	Yes	Data driven	Yes	No	No	No
(Rehman, 2012)	x-DFM	Yes	Data driven	Yes	No	No	No
(Gallinucci, 2013)	Meta-Star	Yes	Data driven	Yes	No	No	No
(Moalla, 2014)	Star Schema	No	Mixed	Yes	No	No	No
(Sautot, 2014)	Constellation	Yes	Data driven	No	No	No	No
Our Proposition	x-DFM	Yes	Mixed	Yes	Yes	Yes	Yes

namic discovery of MC based on ontology. The modeling of a flexible and scalable DW schema is also required to deal with data from SN.

3 OVERVIEW OF THE PROPOSED METHODOLOGY

In this paper, we propose four stages to define data warehouse methodology from social networks (Figure 2). We will begin with the creation of the initial DW schema from structured and heterogeneous sources following a classical approach. The second stage will relate to the transformation of the DW schema into a NoSQL Data Base. After that, we will dynamically determine the MC, their types as well as their locations. These MC are used to enrich the NoSQL DW schema.

Our methodology takes advantages of the maturity of existing design approaches, the scalability of NOSQL Data Base and the capability of the dynamic discovery of multidimensional concepts through clustering techniques. Figure 2 depicts our proposed methodology.

1. Initial DW schema creation: involves the creation of the initial DW schema following a classical mixed approach;
2. NOSQL DW schema creation: it consists on the generation of NOSQL data base for the initial data warehouse schema based on a rules set. These rules allow the transformation of a DW schema concepts to specific concepts of NOSQL data base;
3. Discovery of multidimensional concepts: consists on defining the features set that meet the users needs, generating clusters from social networks based on the defined features and then determining the MC, their types as well as their locations;
4. NoSQL DW schema evolving: using the discovered MC to enrich the NoSQL DW schema.

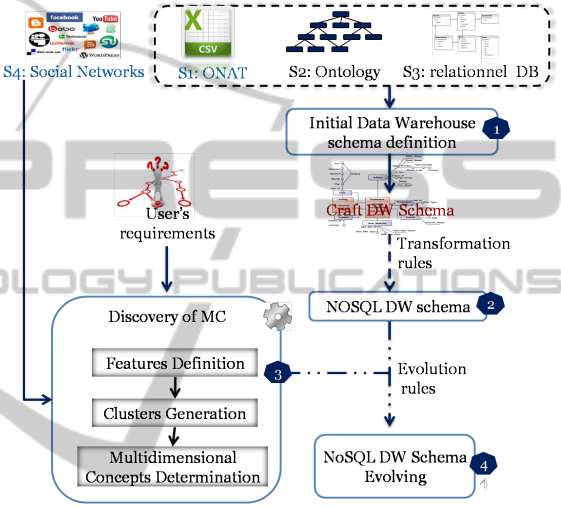

Figure 2: Overview of the proposed methodology.

In the following we expose the initial data warehouse creation, then we detail the discovery of multidimensional concepts using clustering technique.

4 INITIAL DW SCHEMA CREATION: A DW FOR CRAFT PRODUCTION ANALYSIS

In our research project, we need to create a DW to analyze the impact of using IT solutions on the situation of handicraft women. To accomplish this stage, we have data sources (both internal and external) available in the project. They are the following:

- S1: Data Base of the National Office for Tunisian Handicrafts;
- S2: Ontology defined from interviews;
- S3: Tunisians and Algerian postcodes data bases.

In our case study, ontology (S2) represents an internal source which is an essential component to

evolve a DW schema. It contains both data and metadata. This source is a consistent support in the discovery of dynamic multidimensional concepts. The used ontology presents knowledge about the profile of handicraft women. Similarly, the data Base of the National Office for Tunisian Handicrafts (ONAT) is an internal source. Otherwise, the Tunisian and Algerian postcodes data bases s an external source which is important to enrich the craft DW. The craft DW is built from these three sources (S1, S2 and S3) using the ETL (Extract Transform Load) process based on data driven approach using Talend Open Source[5]. An overall view of the initial craft DW schema is depicted in Figure 3.

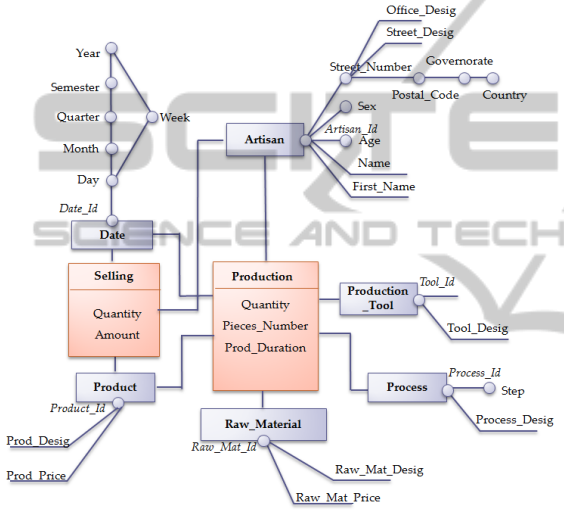

Figure 3: Data warehouse for craft production analysis.

The structure of the initial DW schema is a graph centered at two facts. First, the ***Production*** fact: includes (***Quantity***, ***Pieces_Number***, ***Prod_Duration***) as measures and is linked to (***Artisan***, ***Production_Tool***, ***Process***, ***Raw_Material***, ***Product*** and ***Date***) as dimensions. Second, the ***Selling*** fact which includes (***Quantity***, ***Amount***) as measures and is linked to (***Artisan***, ***Product*** and ***Date***) as dimensions. In this stage, we have followed a traditional approach for DW schema design which is not suitable for the latest data source (S4). Therefor, we propose to use Clustering method.

5 DYNAMIC DISCOVERY OF MULTIDIMENSIONAL CONCEPTS

Classical decision support is used in analyzing simple data. However, these systems are not adapted for SN analysis which highlights the need of creating new models. In fact, data are heterogeneous and changeable over time. Thus, a comprehensive schema for craft DW cannot be fixed at the time of design and must be dynamically modified. In a nutshell, a DW schema can be extended by adding new elements of type measure, dimension, or hierarchy level. These extracted values are to be fed into the semiautomatic schema evolving to dynamically model the DW schema. This stage is divided into three main steps: features definition, clusters generation and multidimensional concepts determination.

5.1 Features Definition

In DW lifecycle, user requirements definition is one of the most important tasks which ensure a successful DW project. The main objective is to identify analyst goals in order to reduce the risk of failure. Since the expert has prior knowledge of the analysis goal, our methodology allows to evolve a DW schema under the designers guidance. At first, the expert is required to select the multidimensional concepts that should be dynamically evolved according to his objectives. He is then required to select one or a set of features that meet his needs. These features are the basis for grouping objects in the next step.

Example. To analyze the business of handicraft women, we should cluster ***Products*** values based on the ***Product_Designation*** and then based on the used ***Raw_Material*** and the used ***Production_Tool*** values. In fact, if two products have the same designation, they belong, therefore, to the same ***Group_Activity***. Otherwise, if they use a set of ***Raw_Material*** and ***Production_Tool*** in common, they probably belong to the same ***Group_Activity***. Consequently, we derive three features which are respectively ***Product_Designation***, ***Raw_Material*** and ***Production_Tool***.

5.2 Clusters Generation

The hierarchical clustering technique is applied to the data set to generate clusters based on a similarity measure. As most of the clustering algorithms are unsupervised, in this step, we target the semi supervised hierarchical clustering in order to get the optimal results that meet the analysts needs. To do that, we have used the SHICARO (Semi-supervised Hierarchical Clustering based on Ranking features using Ontology) (R. Yangui and Gargouri, 2014a) method with a profiling ontology. This method consists of two important components. The first one consists in defining effective ontology-based similar-

[5]http://www.talend.com/

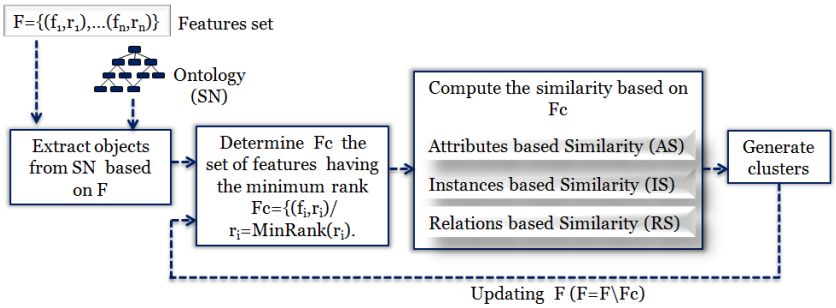

Figure 4: SHICARO method.

ity measures that combine both numerical and nominal variables along different dimensions (instances, attributes, and relation-ships) (R. Yangui and Gargouri, 2014b), while the second consists in providing a performable clustering algorithm based on ranking features. SHICARO method aims to cluster objects based on scheduled features.

Since the expert knows the goal behind which the clustering is performed, SHICARO method performs clustering under the users guidance. Thus, the user is required to order the features from the highest to the lowest ones. In each iteration, a set of features $F = \{(f_1, r_1), \ldots, (f_n, r_n)\}$ that have the same rank r_i are applied to cluster objects. SHICARO steps are depicted by Figure 4.

Example. By applying SHICARO method based on the set of features F=(***Product_Designation***, 1), (***Raw_Material***, 2), (***Production_Tool***, 2), we obtained at the first iteration (based on ***Product_Designation*** feature) fives clusters and at the second iteration (based on ***Raw_Material*** and ***Production_Tool*** features) two clusters (Figure 5).

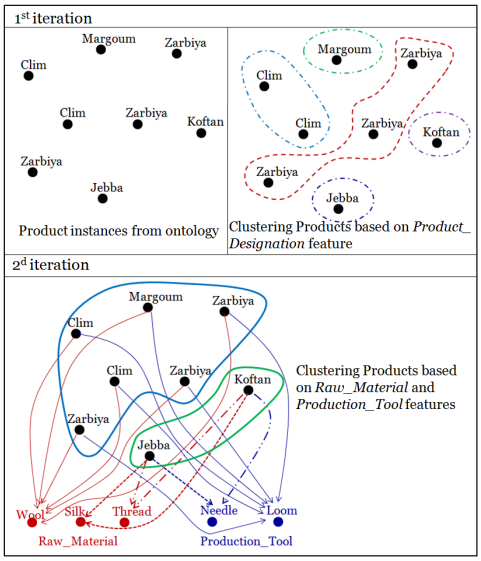

Figure 5: Clustering Product instances.

The extracted values, present in each of the generated clusters, become the input to the dynamic schema evolving in the next step.

5.3 Multidimensional Concepts Determination

This step is performed by the designer. It consists in analyzing generated clusters, determining the type of multidimensional concept, assigning names to clusters and specifying the location of insertion in the multidimensional schema.

Example1. Adding the ***Group_Activity*** level to the ***Product*** dimension hierarchy (Figure 6).

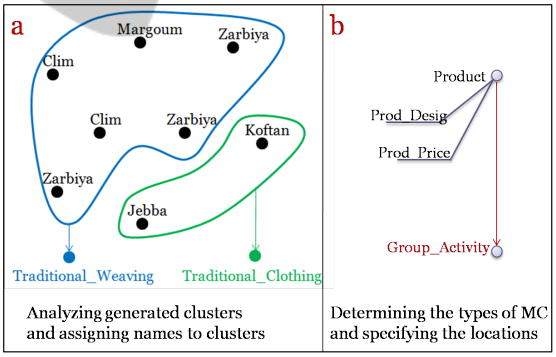

Figure 6: Adding the ***Group_Activity*** level to the ***Product*** dimension hierarchy.

In Figure 6.a, we have two clusters: the first one describes the ***Traditional_Weaving*** and the second one describes the ***Traditional_Clothing***. So we can define the concept ***Group_Activity*** as MC. This MC is considered as parameter of the ***Product*** dimension as depicted in Figure 6.b.

Example2. Adding the ***Customer***, ***Supplier*** and ***Fan*** dimensions (Figure 7) Artisan friends in SN can be divided into groups based on their job, their link and the exchanged clips. Feature set used to cluster womens friends is F= (***link***, 1), (***job***,2), ***Clip***,3). Clustering algorithm based on F determine three groups named ***Customer***, ***Supplier*** and ***Fan***.

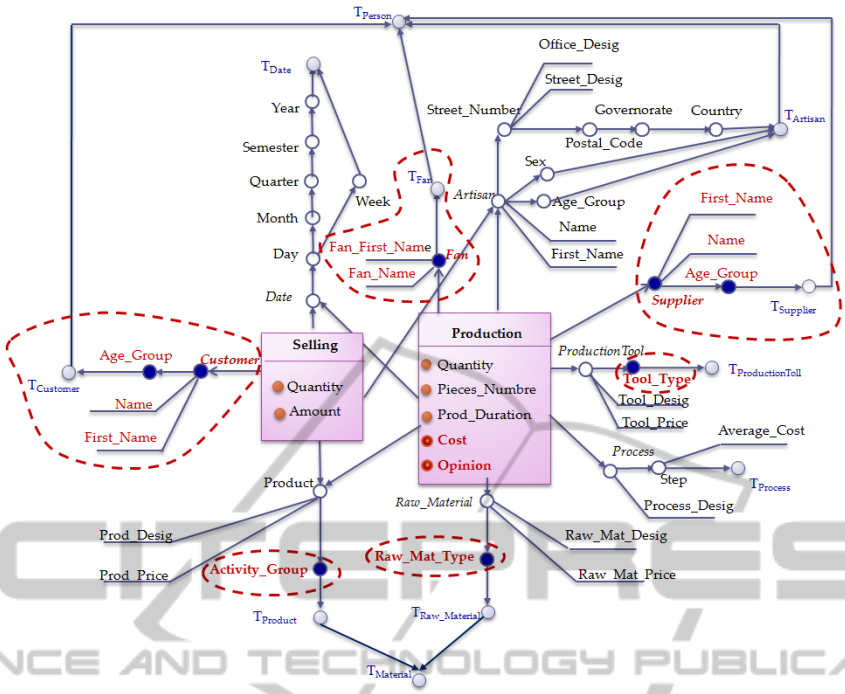

Figure 8: DW schema enriched by discovered MC.

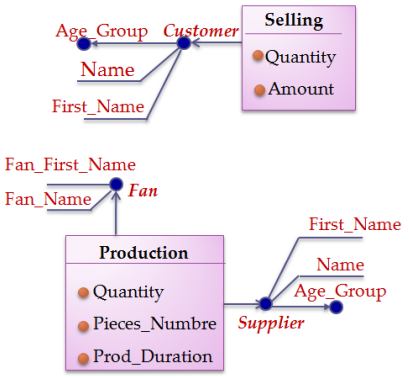

Figure 7: Adding the ***Customer***, ***Supplier*** and ***Fan*** dimensions.

Each group represents a CM, so we have three new MC: ***Customer***, ***Supplier*** and ***FAN***. These MC are considered as three new dimensions. ***Customer*** is added as dimensions to the ***Selling*** fact. ***Supplier*** and ***FAN*** are added as dimensions to the ***Production*** fact as depicted in figure 7.

Figure 8 shows the DW schema enrichment for storing various cumulative data about handicraft women derived from SN in the the extended Dimensional Fact Model (xDFM) (Mansmann, 2008). Dimensions are modeled as aggregation paths. All paths of a dimension converge in an abstract Tnode, which corresponds to the aggregated value all. A level node in a dimension consists of at least one key attribute, but may include further attributes represented as underlined terminal nodes. As shown in Figure 8, three new dimensions are added to the initial DW schemas which are ***Customer***, ***Supplier*** and ***Fan***. These dimensions are discovered by performing the hierarchical clustering based on a set of features. ***Customer***, ***Supplier*** and ***Fan*** reflect the communities to which handicraft women are connected at SN. However, an ***Artisan*** can be also a ***Supplier***, a ***Customer*** or a ***Fan***. That is why a generalization link is defined ***Tperson***. Similarly, ***Raw_Mat_Type***, ***Activity_Group*** and ***Tool_Type*** proprieties are added as levels successively in the ***Raw_Material***, ***Product*** and ***Fan*** dimension hierarchies. These properties are deducted using the hierarchical clustering based on the appropriate features. However, a ***Product*** can also be a ***Raw_Material***. Hence, a generalization link is added ***Tmaterial***.

6 CONCLUSION

In this paper, we proposed four stage methodology to define a data warehouse schema from social networks. Starting in a first stage, by the design of the initial data warehouse schema based on the existing approaches. In a second stage, we have generated a NOSQL Data warehouse schema by applying a set of transformation rules. Then, based on users require-

ment, clustering of social networks profiling data is performed which allows the dynamic discovery of multidimensional concepts. Finally, the enrichment of the NoSQL data warehouse schema by the discovered MC is realized to ensure the DW schema evolution.

We have especially detailed the dynamic content-based discovery of dimensions, hierarchies and measures using hierarchical clustering. This latter, is performed using profiling ontology with adequate similarity measures. The detailed stages are experimented on the real case study of the BWEC project.

We are currently studying the NOSQL data bases, we intend to define transformation rules from a conceptual data warehouse schema to NOSQL database and the evolution rules. Moreover, we think it would be interesting to formally specifying transformation rules to allow the automatic schema generation.

ACKNOWLEDGEMENT

We are very thankful to the Algerian Tunisian Project dealing with the improvement of handicraft women business in emerging countries through affordable technologies and social networks. This project is financed by the Tunisian Ministry of Higher Education, Scientific Research and Information and Communication Technologies Higher Education and Scientific Research sector.

REFERENCES

A. Thusoo, Z. Shao, S. A. D. B. N. J. J. S. S. R. M. and Liu, H. (2010). Data warehousing and analytics infrastructure at facebook. In *International conference on Management of data SIGMOD'10*, pages 1013–1020.

E. Gallinucci, M. Golfarelli, A. W. and Rizzi, S. (2013). Meta-stars: Multidimensional modeling for social business intelligence. In *International Workshop On Data Warehousing and OLAP DOLAP13*, pages 11–18.

L. Sautot, B. Faivre, L. J. and Molin, P. (2014). The hierarchical agglomerative clustering with gower index: A methodology for automatic design of olap cube in ecological data processing context. In *Ecological Informatics*, pages 1–14.

M. Ceci, A. and Malerba, D. (2011). Olap over continuous domains via densitybased hierarchical clustering. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems KES'11*, pages 559–570.

Mansmann, S. (2008). Extending the olap technology to handle nonconventional and complex data. In *Ph.D. dissertation, University Konstanz, Department of Computer and Information Science Germany*.

Moalla, I. and Nabli, A. (2014). Towards data mart building from social network for opinion analysis. In *International Conference on Intelligent Data Engineering and Automated Learning IDEAL14*, pages 295–302.

N. U. Rehman, S. Mansmann, A. W. and Scholl, M. H. (2012). Building a data warehouse for twitter stream exploration. In *International Conference on Advances in Social Networks Analysis and Mining ASONAM'12*, pages 1341–1348.

Nabli, A. (2013). *Approche d'aide la conception automatise d'entrept de donnes: Guide de modlisation*. Presses Acadmiques Francophones.

P. Kazienko, K. Musial, E. K. T. K. and Brdka, P. (2011). Multidimensional social network: Model and analysis. In *International Conference on Computational Collective Intelligence Technologies and Applications ICCCI'11*, pages 378–387.

R. Yangui, A. N. and Gargouri, F. (2014a). Shicaro: Semi-supervised hierarchical clustering based on ranking features using ontology. In *International Conference on Management and Technology in Knowledge, Service, Tourism & Hospitality SERVE'14*, pages 233–238.

R. Yangui, A. N. and Gargouri, F. (2014b). Soim: Similarity measures on ontology instances based on mixed features. In *4th International Conference on Model & Data Engineering MEDI2014*, pages 169–176.

Usman, M. and Pears, R. (2011). Multi level mining of warehouse schema. In *NDT, volume 136 of Communications in Computer and Information Science*, pages 395–408.