

Visualization and descriptive analytics of wellness data through Big Data

Shujaat Hussain, Sungyoung Lee
 Ubiquitous Computing Laboratory
 Dept. of Computer Engineering,
 Kyung Hee University
 Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do,
 446-701, Korea

Abstract—Visualization and analytics are now very important in today's data deluge. The data generated today is very complex and unstructured and it is very difficult to comprehend the data. Due to these factors, visualization and analytics plays a very important role in human understanding. In this paper a big data platform is proposed which takes data from Hadoop distributed file system and the relational database for health analytics and visualization. The data is structured and displayed in terms of analytics and graphs. This platform is capable of transforming the data into a predefined structure so that additional insights can be driven from the data. Visualization and analytics on big data enhances the decision making and knowledge discovery. The framework is capable of handling unstructured data as well as structured data. A case study is given where variety of data is shown in form of analytics and visualization for supporting decision making and additional analysis.

I. INTRODUCTION

Big data is creating many opportunities for different and diverse fields to achieve deeper and faster insights that can enhance the decision making. Visualization can play an important part in this decision making. Techniques such as reporting, dashboards, and data visualization is being widely used now with emergence of big data. The goal of visualization is to provide the end users and experts with an understanding of the information contents. There are two modes of visualization in the system i.e. for the end user and an expert mode. Visualization brings context to the data and removes misunderstandings and improve forecasts There is a need to effectively visualize data to highlight the anomalies, trends and different attributes related to the user. Multiple layer visualization with different data views is very necessary as the visualization will aid users, domain experts and different stakeholders of the system. The rest of the paper is organized as follows: Section II presents the related work and proposed framework in section III. Case study are illustrated in section IV. Finally, we conclude our work in section V.

II. RELATED WORK

A novel network traffic visualization scheme is proposed based on exploiting frequent item-set mining with the aim to visualize traffic patterns extracted from communication logs using hypergraphs [1]. A Intra-user and inter-user social media from a big data framework is proposed. It analyzes social

media and visualization from the server and display to a user. It gets the signal fusion from heterogeneous sensors [2]. Kandogan ET al, created a feature ranking and annotation method. They do annotation interaction to help support understanding of the structure of data [3]. The new emerging directions are highlighted to create clear, meaningful and integrated visualizations that give biological insight, without being overwhelmed by the intrinsic complexity of the data [4]. A novel Massive Model Visualization via a rendering approach called MMDr using spatial hierarchies, with the goal of a better understanding, to achieve interactive frame rates on extremely large data sets, and help to accomplish more advanced research into Massive Model Visualization [5]. A tuning parameter taxonomy is proposed which consists of behavioral, Numerical statistically and temporal classifications and are further sub divided. This taxonomy help analyze different attributes and dimensions of the data being visualized [6]. An overview of current real-time massive model visualization technology, with the goal of providing readers with a high level understanding of the domain [7]. They have included simulations and different rasterization algorithms for rendering. A novel approach to automatic image annotation based on two key components: (a) an adaptive visual feature representation of image contents based on matching pursuit algorithms; and (b) an adaptive two-level segmentation method [8]. For parameter tuning McAdams et al uses a subtle, but significant, change in the design: the addition of a tuning parameter in place of an increase in component precision. Statistical models are used to develop a framework for the tuning parameter design method in [9].

III. PROPOSED TECHNIQUE

The visualization architecture is shown in figure 1. The data comes from the Hadoop distributed file system and relational database. The parameter tuning takes unstructured Data from big data storage as input and output is filtered data from Parameter tuning component.

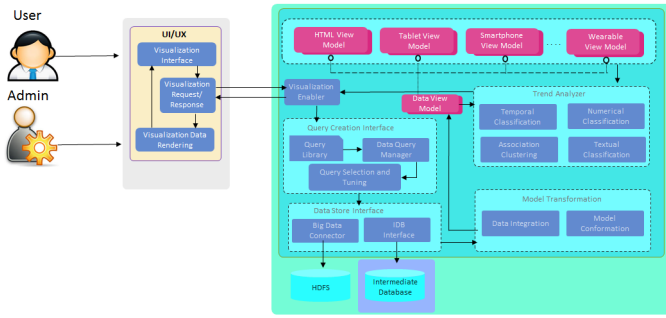


Fig. 1. Architecture of descriptive analytics

A. Visualization Enabler

The visualization enabler is request response interface as well as the graph template holder. One of its function is to send the user request to the query creation interface. The other function is to plot the filtered data from the trend analyzer for the graphical view.

B. Query Creation Interface

The query creation interface consists of query library, data query manager and query selection and tuning. The query library contains the set of queries and the query manager maps the user request to those queries. These requests are broken into the three parts. i.e. conditional attributes, temporal attributes and attributes that needs to be returned to the user in the form of visualization and analytics. The query selection and tuning component fine tunes the query for data store interface as sometimes the query cannot be mapped exactly.

C. Data Store Interface

The data store interface contains a big data connector. The big data connector decides whether the data has to be queried from the Hadoop Distributed File System (HDFS) or the relational database known in the architecture as intermediate database (IDB). If the temporal attribute is above the threshold i.e. if the user query is for more than a month or so, the data will be retrieved from the HDFS.

If the threshold is not met, then the faster option is taken and the IDB interface queries data from the intermediate database.

D. Model Transformation

The data is sent to the model transformation module. It consists of model integration and model conformance. The main purpose of this model is to transform the data extracted from the big data repository. The data is unstructured and needs to be transformed so that further processing can be done. The model integration transforms the data in a format specified like JSON, XML and a tabular format. The model conformance removes redundancy and validate the model that data is transformed into to speed up the analysis.

E. Data View Model

Data view model depends on the device that user is requesting from. The view decides how much data needs to displayed depending on the size of the screen. The screen can be desktop view, a tablet view or a smartphone.

F. Trend Analyzer

The trend analyzer classifies the data depending on the metadata and clusters the data for visualization. The numerical and temporal classification separated the attribute nature and it contributes to the visualization in term of graphs and scales. The textual classification triggers the descriptive analytics. The association cluster module clusters the module through the clustering and summarization technique for similar attributes.

IV. CASE STUDY

After the query is posed to the platform, the query is broke down into different attributes. For example the query was **"Find the food consumption grouped by user age group 25-30 over last month"**. The query is broke in three attributes as shown in table 1.

TABLE I
QUERY BREAKDOWN IN QUERY CREATION INTERFACE

Return Attribute	Conditional Attribute	Temporal Attribute
Food	Age >25	Days=30
Age	Age >30	
Facts		

Table 1 show the return attribute which are food, age and facts about these foods. These are the parameters which will make up the visualization and descriptive analytics. Conditional attribute breaks downs the condition of the query for posing simple query to big data if needed. This table help map the query to the query library. When the match is made, then the query is tuned to append the temporal condition which is usually not attached to queries saved in the library.

The query is sent to the data store interface and posed to the big data connector. The big data connector poses it to the HDFS as the threshold of 30 days is above the data handled by the relational database.

The data is transformed into a tabular format. The transformation is a challenge as it is done in mapreduce and is necessary for classification of the data. The data stored in the HDFS is in flat structure and needs to be semantically structured.

The trend analyzer component classifies the data into temporal, numerical and textual classification.

TABLE II
TEXTUAL CLASSIFICATION

Food	Facts
Chocolate	Vitamin A & D, Magnesium
Spinach	Iron, Calcium, Vitamin A & C
Steak	Sodium, Magnesium

TABLE III
TEMPORAL CLASSIFICATION

Days	Dates
Wednesday	12 June
Thursday	13 June
Sunday	16 June

TABLE IV
NUMERICAL CLASSIFICATION

ID	Calories	Age
1	600	25
3	100	36
6	250	28

The user is shown the graph in figure 2 from the classified data through summarization and association technique.

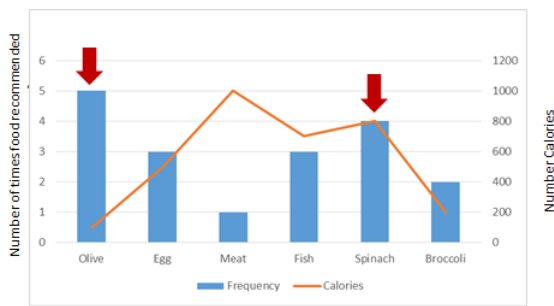


Fig. 2. Visualization for the query

The analytics through facts are associated with different symptoms and diseases and shown in figure 3.

Steak contains cholesterol and sodium content as well as saturated fat. It increases the risk of heart problems
Spinach contains protein, calcium, iron, magnesium, potassium, Vitamin A and B.



Fig. 3. Descriptive analytics for the query

V. CONCLUSION

Big data analytics and visualization go hand in hand as we need an effective way to display the data intuitively for the users and the developers. In this paper a platform is proposed which queries and transforms the data for analytics and visualization. It extracts the data both from big data repository and relational databases and classify it for faster processing in Hadoop. The graphs and visual summaries are provided to the user after data reduction.

For future work, we will try to extend the techniques in the trend analyzer, as now only association and summarization technique is used. Another drawback of expert intervention in query tuning will be addressed. Complex queries and multi-dimensional data structure will also be addressed in the analytics platform.

ACKNOWLEDGMENT

This work was supported by the Industrial Core Technology Development Program (10049079 , Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea)" and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) NRF-2014R1A2A2A01003914.

REFERENCES

- [1] E. Glatz, S. Mavromatidis, B. Ager, and X. Dimitropoulos, "Visualizing big network traffic data using frequent pattern mining and hypergraphs," *Computing*, vol. 96, no. 1, pp. 27–38, 2014.
- [2] Y.-T. Chang and S.-W. Sun, "A realtime interactive visualization system for knowledge transfer from social media in a big data," in *Information, Communications and Signal Processing (ICICSP) 2013 9th International Conference on*. IEEE, 2013, pp. 1–5.
- [3] E. Kandogan, "Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations," in *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 2012, pp. 73–82.
- [4] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum *et al.*, "Visualization of omics data for systems biology," *Nature methods*, vol. 7, pp. S56–S68, 2010.
- [5] J. S. Bennett, "Massive model visualization: An investigation into spatial partitioning," 2009.
- [6] <http://www.ibm.com/developerworks/library/ac-canada/>, [Online; accessed 17-Aug-2015].
- [7] A. Dietrich, E. Gobbetti, and S.-E. Yoon, "Massive-model rendering techniques: a tutorial," *IEEE Computer Graphics and Applications*, no. 6, pp. 20–34, 2007.
- [8] R. Shi, H. Feng, T.-S. Chua, and C.-H. Lee, "An adaptive image content representation and segmentation approach to automatic image annotation," in *Image and Video Retrieval*. Springer, 2004, pp. 545–554.
- [9] D. A. McAdams and K. L. Wood, "Tuning parameter tolerance design: Foundations, methods, and measures," *Research in Engineering Design*, vol. 12, no. 3, pp. 152–162, 2000.