A conceptual modelling-based approach to generate data value through the end-user interactions: A case study in the genomics domain

Carlos Iñiguez-Jarrín

Escuela Politécnica Nacional, Departamento de Informática y Ciencias de la Computación,
Facultad de Ingeniería de Sistemas
Ladrón de Guevara, E11-253, Quito, Ecuador
P.O. Box 17-01-2759
carlos.iniguez@epn.edu.ec

Universitat Politècnica de València, Valencia, Spain
ciniguez@pros.upv.es

Abstract. In the current Big data ecosystem, identifying the data with the real value to an organization, or in other words the “data value", is a key issue for the decision making process. Understanding data implies a challenging cognitive process, which involves the know-how of domain experts. We propose an approach based on conceptual modelling to discover the data value through the interactions made by users when exploring the data. Our main ideas are: 1) To create a base of domain knowledge represented by interactions; 2) To formalize the interactions of the users with the data ecosystem. Our goal is to express high-level interactions between end-users and the data scenario, which represents the cognitive process followed to enact value from data. Such interactions together a subjacent conceptual model will be the mechanisms to recommend the next data exploring steps. In this paper we provide a solution design to generate value from the huge amount of data.

Keywords. Data value, interaction design, visual data exploration

1 Introduction

"Big data" is a umbrella term for expressing the huge ecosystem of structured and unstructured data, which has become a issue of great interest for industry due to it represents a potential resource to get an in-depth understanding of data. Because the

heterogeneous characteristic of data, "Big data" requires data analysis mechanisms more powerful than the traditional ones. [1]. “Decision makers of all kinds, (…), would like to base their decisions and actions on this data” [2]. Turning data into meaningful information becomes a challenge [3][4]. Understanding the data is a complex task that involves to locate relevant information from the large amount of existent data and the cognitive process involved to obtain the “data value”.

But, what is the meaning of "data value"? We define the "data value" like the meaningful information obtained from a sense-making process that adds value to the accomplishment of user goal. Data become information when they are ascribed value [3]. In a sense-making process, domain experts create, modify, and evaluate schemas of relations between items [5] [6] . Such schemas are representations of conceptual models consisting of concepts and their relationships that allow people to abstract a problem. In the Software Engineering and Data Base areas, conceptual models are usually used to organize and shape the body of knowledge, as the skeleton fulfils its function in the human body. However, understanding such conceptual models in order to explore and generate value from data is not an easy task even by skilled technical people. Therefore, the need for an intuitive mechanism for manipulating the conceptual model as support for sense-making process by non-skilled users becomes a challenge.

This is where the prediction based on user-interactions and the interactive visualization come into play. In the data sense-making process scenario, predictive interactions could be the way to facilitate the conceptual model understanding. Capturing and formalizing the resulting domain expert interactions become source of tracks about their preferences, behaviour and decisions. Thus, inferring over such interactions in order to predict the next steps to take by users together a suitable visual guide can help them to decrease the complexity on understanding the conceptual model.

The aim of this research is obtaining the data value through a sense-making process to provide meaning from data. To achieve such a goal, in this paper we describe the design of an approach based on underlying conceptual model and end-user interactions in order to generate data value from a huge data set. The domain expert will work on visualizing and exploring data and the set of registered interactions together an underlying conceptual model will be the source of recommendation to generate meaningful information.

In order to apply our research, we selected the genetic diagnosis of diseases as a case of study, where from a huge amount of genetic data, domain experts (clinicians, physicians, scientists, etc.) carry out a sense-making process to detect whether a person has an illness or estimate the risk of developing some disease. From a genetic sample taken from one o many patients, the practitioners explore and analyse manually the genetic data to find relationships between sample genetic mutations and relevant information of genetic diseases, which is available on public genetic databases. This is like finding a needle in a haystack. Finally, the relevant findings that contribute to the diagnosis are consolidated y reported.

Our proposal is related to the analysis phase, where the coalescence between appropriate interactive mechanisms and the genome conceptual model [7] may be a suitable mechanism to generate value of the large amount of genetic data.

The paper is structured as follows: the next section summarizes the related work then, in Section 3, we present the Design Science research methodology [8] applied to this research. In Section 4, we describe the problem statement and identify the research questions we plan to answer in the proposed work. Finally, section 5 presents an overview of our solution design.

2 Related Works

In this section, we briefly review relevant works from Human-computer Interaction, Information Retrieval, and Data Visualization and Machine Learning that are related to improving the way in which users get meaning from the data.

Le et al. [9] discuss the use of analytic trails technology as part of Smarter Decisions to support the users when conducting visual data analysis. Although the use of the analytic trails is focused on the aspects of analytic provenance, asynchronous collaboration, and reuse of analyses, it is important to mention that the captured interactions become the source of "trails" of analysis tasks representing the analytic steps taken by the user during visual data exploration.

Athukorala [4] analyses how user interaction modelling can be applied to provide better support in exploratory information–seeking. He proposes to model the user behavior to allow information retrieval systems to infer the state of exploration from observable aspects of user interactions**.**

JIT interactive analytics [10] is a proposal to join computational and visual techniques. JIT analytics is performed in real-time on data that users are interacting with to guide visual-analytic exploration where enriching visualizations with annotations suggests to users possible insights to examine further.

In [11], RockQuery is proposed as a Visual Query System (VQS)/Data Query Tool that combines ontology views with interaction community techniques to reduce the overload of information by presenting visually to the user only the information that is required by the data exploration task at hand.

In [12], a knowledge navigation infrastructure is presented to explore literature related to dengue disease. A content acquisition engine drives the delivery of dengue-specific literature from public repositories by means of previously identified keywords.

In [13], Optique is presented as a tool to enable the access to Big data from a ontology based approach to formulate visual queries.

In [14], focused on exploratory search, is presented an approach where users can directly manipulate document features (keywords) through a graphic user interface to indicate their interests and reinforcement learning is used to model the user by allowing the system to trade off between exploration and exploitation.

All the works cited describe mechanisms aimed to allow users to get insight of large dataset. Although they use annotations and keywords as source of automatic learning approaches, they do not focus on the interactions set as a source of valuable information. We propose an approach where the process to get value from data is supported by both the inference over the user-interactions as guide to discover related data

and the conceptual model as guide to allow users to visually navigate on obtaining value from data.

3 Research Methodology

This research will follow the guidelines of the Design Science Methodology [8]. Design Science is oriented to information systems and software engineering research, an appropriate approach to the nature of our research. The main object of study of this methodology is an artifact in a problem context. In our case, the artifact is:

A conceptual modelling approach to generate data value based on user interactions

And the problem context consist of:

Data analysis.

Our research is considered as a utility-driven exploratory research project. Exploratory research since our stakeholders (SENESCYT and Escuela Politécnica Nacional) are willing to sponsor an exploratory research and utility-driven since the research results must accomplish with its budgets and goals. Such goals can be expressed in a hierarchical structure where the achievement of high-level goals is the result of the achievement of every low-level goal. Our goals are defined and ordered from highest (G1) to lowest (G4) goals level, as follow:

- G1: Develop a conceptual modelling-based approach to generate data value through the end-user interactions.
- G2: Determine the conceptual modelling-based approach to generate data value through the interactions.
- G3: Predict the effects caused by the approach in the context of use.
- G4: Make a literature review to determine the existent approaches to address the generation of data value from both the conceptual model and interaction perspectives.

The defined goals derive problems classified on *design problems* and *knowledge questions*. Solving design problems imply changing the real world to suit human purposes, in contrast, solving knowledge questions imply acquiring knowledge about the world without necessarily changing it [15]. In order to deal with the mentioned problems, the methodology provides two rational nested problem-solving cycles, which consist of several tasks. Both cycles work in a nested way ensuring the design and evaluation of artifacts. The Fig. 1 shows the two nested cycles applied to our research project. The Design Cycle contains three tasks (from T1 to T3) to address design problems, whereas the Empirical Cycle contains five tasks (from T4 to T8) to address the knowledge questions. It is important to mention that, the design cycle is a subcycle from Engineering Cycle, for this reason not all tasks from engineering cycle are regarded into the design cycle. Design science projects are always restricted to the first three tasks of the engineering cycle [8]. The *treatment implementation* (T9) is out of scope of this research project.

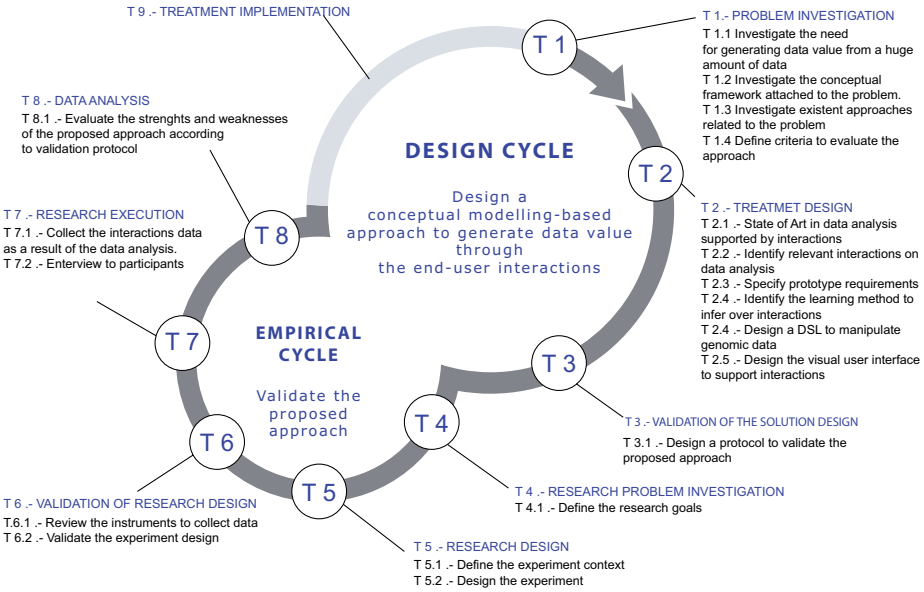

Fig. 1. Design Cycle and Empiric Cycle applied to the research

4 Research Questions

Our research aims to provide an conceptual modelling approach to generate data value throug user-interactions. To achieve this goal, We have defined the research questions showed in the Table 1.

Code	Research Question	Goal	Problem Type
RQ 1	What approaches exist for data value generation supported by conceptual models and end-user interactions?	G2, G4	KQ
RQ 2	How to develop a conceptual modelling approach to generate data value through the end-user interactions?	G1	DP
RQ 2.1	Identify the user-interactions that can be suitable to support generating data value	G1, G4	DP
RQ 2.2	Define the user interface characteristics that allow user to generate value from data.	G1, G4	DP
RQ3	What effects does the approach applied to the context of use cause?	G3	KQ

Table 1. Research Questions

Every research question is related to the target goal defined in the Section 3, and the research problem type: knowledge questions (KQ) and design problem (DP).

5 Solution Design Proposal

Our proposed approach is oriented to provide an environment where a domain user can be guided by predictive interactions when exploring a large dataset. The process is depicted in the Figure 2. In the first step, a graphical user interface allows user interact with conceptual model elements expressed in meta-information in order to order to explore and make sense of data showed. In the second step, the interactions performed are stored in a knowledge database. Every action performed when the user explores the data, represents an interaction consisting of a) the data attributes to specialize the exploration and b) the track of conceptual model nodes visited until while. In the step 3, an inference engine consisting of predictive algorithms takes the stored interactions in order to suggest possible steps in the search process. The outcomes are transformed to a suitable input for a DSL depicted in the step 4. The DSL use the set of data attributes to request from database, the related data that accomplish the search constraints, whereas the track of conceptual model selected is used to reasoning over the conceptual model in order to find out the connected concepts related to the last visited element. Those two sorts of information: related data and connected concepts are communicated to the graphical user interface as depicted in the step 5. The showed data represent the actual state of search whereas the connected concepts become on the possible next steps that systems suggest to the user in order to guide the process of get meaning of data.

Unlike data-driven traditional approaches, where the users requires knowing the conceptual model in order to create tailored queries to their needs, we aim to provide an interactive and predictive mechanism that combines both the data-driven and model-driven perspectives, that allows the user to obtain answers to their questions while discovering the knowledge guided by the conceptual model of the problem. Hence, end-users do not need to know the whole conceptual model which represents a complex task.

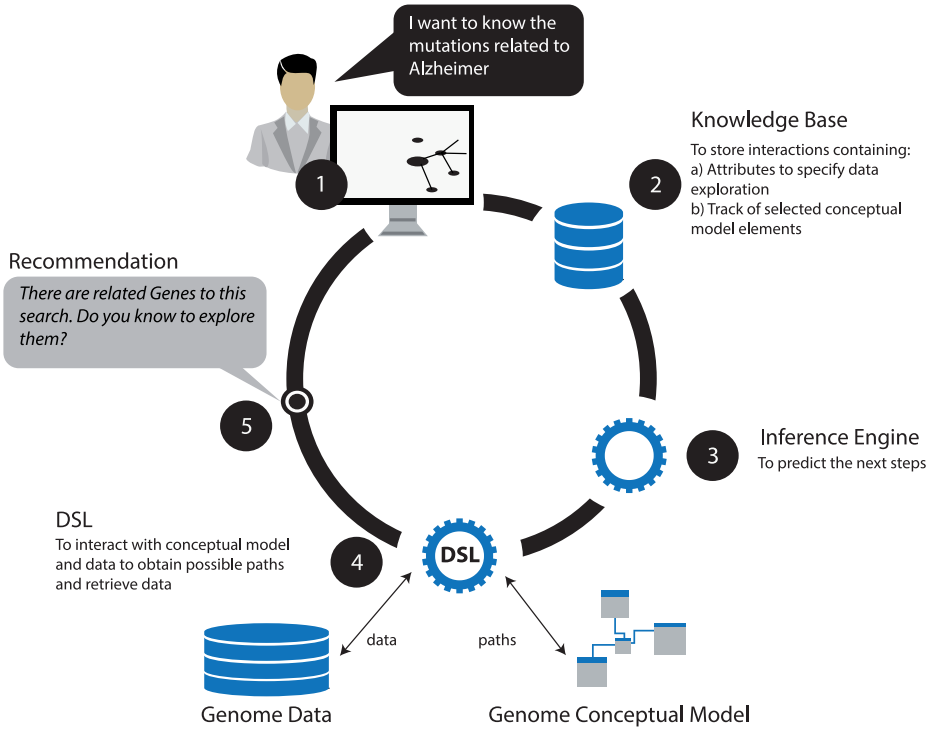

Fig. 2. Solution Design

The main contributions to be expected from this work are:

1. Exploratory study of interactions of genetists when diagnosing genetic diseases.
2. Design a prototype to generate data value through user interactions, applied to the genetic diseases diagnosis
3. Formalize the user-interactions involved on diagnosing genetic diseases.
4. Evaluate the user-interactions based approach on a real scenario.

6 Acknowledgements

The author gratefully acknowledge the financial support provided by the Escuela Politécnica Nacional, Secretaría Nacional de Educación, Ciencia y Tecnología (SENESCYT) and IDEO project for the development of this research project. Thanks to supervisor Óscar Pastor for his invaluable support and advice.

7 References

[1] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” in *Mobile Networks and*

Applications, 2014, vol. 19, no. 2, pp. 171–209.

[2] F. Danyel, D. Rob, C. Mary, and S. Drucker, "Interactions with Big Data Analytics population by running controlled," *Interactions*, vol. May-June, pp. 50–59, 2012.

[3] M. R. Lissack, "Of chaos and complexity: managerial insights from a new science," *Manag. Decis.*, vol. 35, no. 3, pp. 205–218, Apr. 1997.

[4] K. M. Athukorala, "Enhancing Exploratory Information-Seeking through Interaction Modeling," in *USER MODELING, ADAPTATION, AND PERSONALIZATION, UMAP 2014*, 2014, vol. 8538, pp. 478–483.

[5] D. M. Russell, M. J. Stefii, P. Pirolli, and S. K. Card, "of Sensemaking," pp. 24–29, 1993.

[6] D. H. Chau, A. Kittur, J. I. Hong, and C. Faloutsos, "Apolo," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 2011, vol. 46, p. 167.

[7] O. Pastor, J. C. Casamayor, M. Celma, L. Mota, M. Á. Pastor, and A. M. Levin, "Conceptual modeling of human genome: Integration challenges," *Lecture Notes in Computer Science including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, vol. 7260 LNCS. pp. 231–250, 2012.

[8] R. Wieringa, *Design science methodology*. 2014.

[9] J. Lu, Z. Wen, S. Pan, and J. Lai, "Analytic trails: Supporting provenance, collaboration, and reuse for visual data analysis by business users," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6949 LNCS, no. PART 4, pp. 256–273.

[10] E. Kandogan, "Just-in-time interactive analytics: Guiding visual exploration of data," *IBM J. Res. Dev.*, vol. 59, no. 2/3, p. 12:1-12:10, Mar. 2015.

[11] J. Lozano, J. Carbonera, M. Pimenta, and M. Abel, "RockQuery An Ontology-based Data Querying Tool," vol. 3, pp. 25–33, 2015.

[12] M. Rajapakse, R. Kanagasabai, W. T. Ang, A. Veeramani, M. J. Schreiber, and C. J. O. Baker, "Ontology-centric integration and navigation of the dengue literature," *J. Biomed. Inform.*, vol. 41, no. 5, pp. 806–815, Oct. 2008.

[13] A. Soylu, M. Giese, E. Jimenez-Ruiz, E. Kharlamov, D. Zheleznyakov, and I. Horrocks, "OptiqueVQS: towards an ontology-based visual query system for big data," *Proc. Fifth Int. Conf. Manag. Emergent Digit. Ecosyst. - MEDES '13*, pp. 119–126, 2013.

[14] D. Glowacka, T. Ruotsalo, K. Konuyshkova, K. Athukorala, S. Kaski, and G. Jacucci, "Directing exploratory search: Reinforcement Learning from User Interactions with Keywords," *Proc. 2013 Int. Conf. Intell. user interfaces - IUI '13*, pp. 117–128, Mar. 2013.

[15] R. Wieringa, "Design Science as nested problem solving," *4th Int. Conf. Des. Sci. Res. Inf. Syst. Technol.*, pp. 1–12, 2009.