Automated Systems & Technologies
25-26 May 2015 • St. Petersburg, Russia

VISUAL MODELS FOR BIG DATA ANALYSIS

Tatiana Gavrilova, Margarita Gladkova

Graduate School of Management
Saint-Petersburg State University
Volkhovskiy Per. 3
199004, St. Petersburg, Russia
gladkova@gsom.pu.ru

Abstract
The paper presents an approach for visual analysis and structuring of Big Data based on the principles of ontological engineering and cognitive psychology. Ontologies are to be used as a basis for big data volumes of information, we tried to follow the principle of good shape. The data structuring procedure is the key element of any model design and development. The suggested methodology proposes better vision and understanding of huge amounts of business information. Ontologies that describe the main concepts of exemplary domains are used both for deeper comprehension and better information sharing. The application of this methodology is considered for the business data on global entrepreneurship monitoring.

INTRODUCTION

Nowadays, companies are dealing with an exploding amount of data and information. The information comes from everywhere and not just generated inside the companies. Data gives both the new opportunities and new ultimate challenges for business. On the one hand, the ability to collect, manage and analyze data effectively allows to make better decisions and build competitive strategies. On the other hand, it is really difficult to extract the value from data that comes from everywhere.

It can be stated that organization became information driven. In 2012 The

Economist Intelligence Unit conducted a survey of 607 executives from across the globe, with 38% based in Europe, 28% in North America, 25% in Asia-Pacific and the remainder coming from Latin America and the Middle East and Africa. The sample was senior, C-level and board executives and other high-level managers. Respondents worked in a variety of different functions and hailed from over 20 industries such as finance, technology, manufacturing, healthcare and pharmaceuticals, consumer goods and retail. Nine in ten survey respondents agree that data is now an essential factor of production, alongside land, labour and capital. At the same time, the research showed that organizations are struggling with the enormous volumes of data and often with poor quality data [8].

And what companies face with is a Big Data phenomenon which they need to elaborate on. Big data is not a precise term; rather it's a characterization of the never-ending accumulation of all kinds of data, most of it unstructured. It describes data sets that are growing exponentially and that are too large, too raw or too unstructured for analysis using relational database techniques. Typically, big data can be characterized by the four "V's":

- Volume: the amount of data being created is vast compared to traditional data sources;
- Variety: data comes from different sources and is being created by machines as well as people;
- Velocity: data is being generated extremely fast — a process that never stops;
- Veracity: big data is sourced from many different places, as a result there is a need to test the veracity/quality of the data.

The Big Data may be gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies, software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks as well as by the people. In order to process Big Data one can use classification, visualization, data mining, crowdsourcing techniques, signal processing, simulations and other various tools.

During the last decade, visual data and knowledge representation has become one of the key considerations in data base design [14], e-business [20] and other applications [13]. But major of the works now are connected with the resulting visualization ([22], while this paper follow object-oriented approach (OOP) approach of the preliminary visual structuring. We will deal not with well known ERP-diagrams or UML notations but with method that is heavily associated with ontology development. The ideas of using ontologies and visual structuring in business applications were discussed in many works ([7], [11], [8]) and now are implemented in many sectors.

Ontological engineering can also be used as an effective research instrument to

study how the structure and patterns of the domain data and knowledge are related to other content pieces. Much of the research in this field so far has focused on a limited number of formal representations that are typically easy to be developed while cognitive and methodological issues are rather underestimated. Furthermore, categorization and laddering as the creative synthesizing activities also did not receive much attention in the literature. Regardless of how ontological engineering is used, in all cases it is necessary to analyze the design procedure.
This paper traces the cognitive foundations of the data model design using the methods of structured ontological engineering. The purpose of the described methodology is to provide data analysts with the distinct recommendations in ontology design and orchestrating for better knowledge transfer and sharing.

METHODOLOGICAL BASIS

The idea of using visual structuring of information to improve the quality of databases is not new. For more than twenty years concept mapping ([21], [6], [15]) has been used for providing structures and mental models that support the process of information understanding. Knowledge and data analysts are making visible the skeleton of the studied discipline and showing the domain's conceptual structure. Often this structure is called "ontology" ([18]).
However, ontology-based approach to information representation is a relatively new development. Ontology is a set of distinctions we make in understanding and viewing the world. There are numerous definitions of this milestone term ([12], [11]), that together clarify the ontological approach to knowledge structuring while giving enough freedom to open-ended, creative thinking. Ontologies now supposed to be one of the most universal and sharable forms of conceptual modelling.
Frequently, it is impossible to express all the information as a single ontology. Accordingly, subject knowledge storage provides for a set of related ontologies. Some problems may occur when moving from one ontological space to another, but constructing meta-ontologies may help to resolve these problems. Meta-ontology provides more general description dealing with higher level abstractions. Figure 1 illustrates different ontology classifications in the form of the mind map. Mind-mapping ([4]) and concept mapping ([17]) are now widely used for visualizing of the ontologies at the design stage.

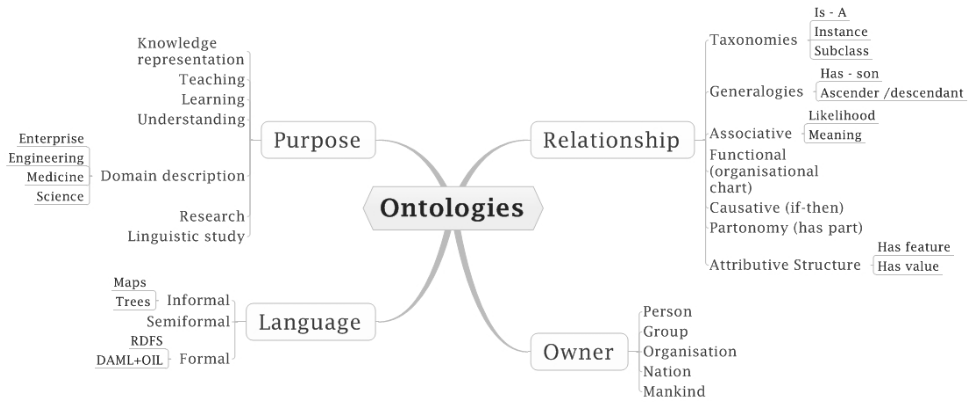

Fig. 1. Summarizing the ontology classifications in a mind-map

ONTOLOGICAL ENGINEERING

Ontological engineering as presented at Figure 2 covers all the issues of ontology development and applications. But it is an unfortunate tradition that technological aspects are much more explored than the methodological ones. Ontology development still faces the knowledge acquisition bottleneck problem, as it was described in the work ([12]). The ontology developer comes up against the additional problem of not having any sufficiently tested and generalized methodologies recommending what activities to perform and at what stage of the ontology development process these activities should be performed. Even the last decade when some effective tutorials on ontology development were presented ([18], [16]), the absence of structured guidelines and methods hindered the development of shared and consensual ontologies within and between teams; the extension of a given ontology by others; and its reuse in other ontologies and final applications.

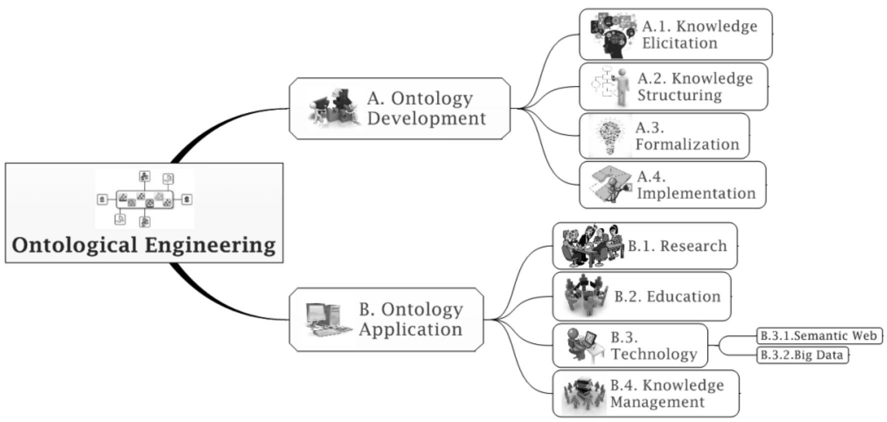

Fig. 2. Ontological Engineering

Simple Recipe for Ontology Design

While in major works the emphasis is put on ontology specification (or coding), we would like to elucidate again the essentials of ontology capture in the simplest form as a recipe for "dummies":

- *Goals, strategy, and boundary identification:* The first step in ontology development should be to identify the purpose of the ontology and the needs for the do-main knowledge acquisition. It is important to be clear about what type of the ontology (see Figure 1) is being built and what level of granularity the concept has.
- *Glossary development or meta-concept identification*: This time consuming step is devoted to gathering all the information relevant to the learned domain.
- *Laddering, including categorization and specification:* The high level hierarchies among the concepts should be revealed and the hierarchy should be represented visually on the defined levels. This could be done via a top-down strategy or bottom-up structuring strategy.
- *Orchestration*: This term means the harmonious organization ([19]). The final step is devoted to updating the visual ontology structure by excluding any excessive-ness, synonymy, and contradictions. The main goal of this final step is to create a beautiful or harmonious ontology.

Beauty is a characteristic of an object, or idea that provides a perceptual experience of pleasure, meaning, or satisfaction. The experience of "beauty" often involves the interpretation of some entity as being in balance and harmony with nature, which may lead to feelings of attraction and emotional well-being. Because this is a subjective experience, it is often said that "beauty is in the eye of the beholder" [15].

Visual Ontology Orchestrating

Bearing in mind that ontologies are to be used as a basis for big data volumes of information, we tried to follow the principle of good shape. The most substantial impulse to the formal definition of this concept was given by the German psychologist Max Wertheimer. His criteria of good Gestalt (image or pattern) ([19]) we partially transferred to ontological engineering:

- Law of Pragnanz (the law of good shape) – the organization of any structure in the nature or cognition will be as good as the prevailing conditions allow. 'Good' here means regular, complete, balanced, and/or symmetrical.
- Law of Parsimony – the simplest example is the best (the Ockham's razor principle): entities should not be multiplied unnecessarily.

In the case of building ontological hierarchies, we have to keep in mind that a well-balanced hierarchy corresponds to a strong and comprehensible representation of the domain knowledge. We enlist below some tips that we consider useful in formulating the idea of "harmony" ([12]):

- Concepts of one level should be linked to their parent concept by one type

of relationships, for example, “is-a”, “has part”, etc. This means that concepts of one layer have similar nature and level of granularity.
- The ontology tree should be balanced, that is, the depth of the paths in the ontological tree should be more or less equal (±2 nodes).
- Cross-links should be avoided as much as possible.

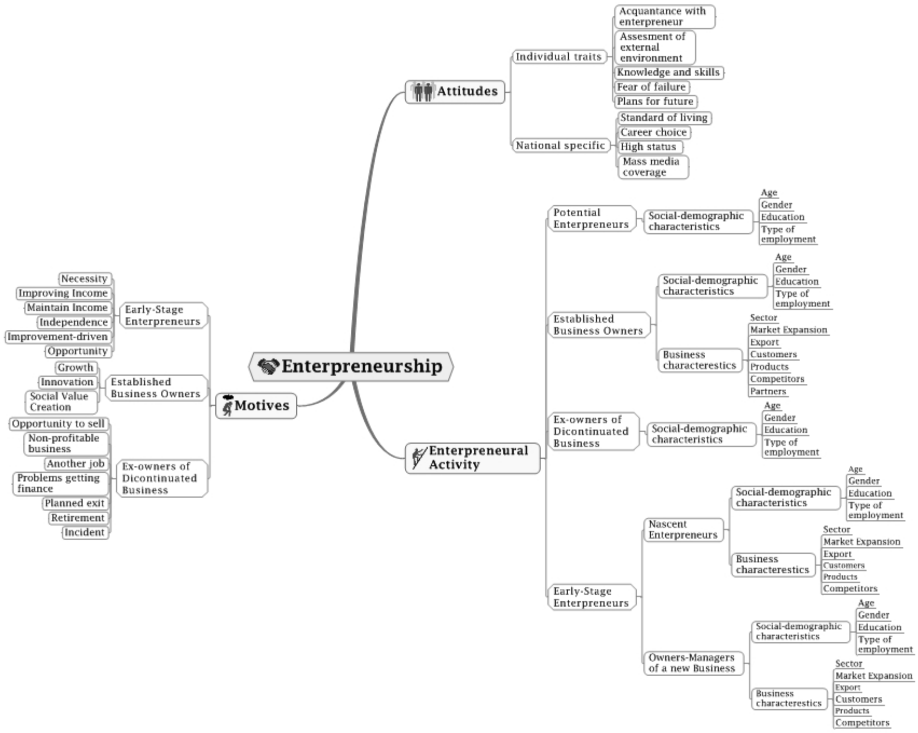

Fig. 3. BigData processing with the help of ontologies: Global Entrepreneurship Monitor

Let’s consider the possible ways of dealing with Big Data and how ontologies may structure them. The research group of the Graduate School of Management at St. Petersburg University of Research Center of Entrepreneurship every year work on the Global Entrepreneurship Monitor report [24]. This project is a joint project of the world’s leading business schools that conducts a series of cross-national research projects on entrepreneurial development and that facilitates the exchange of information on entrepreneurial activity in different countries. Since 2006, the Graduate School of Management, St. Petersburg State University and the National Research University—Higher School of Economics, Moscow represent GEM consortium. Data set is collected using special questionnaire revealing respondents’ attitudes to conditions of entrepreneurial activity and their involvement in the entrepreneurial process. The minimal representative sample in each country is 2000 adults. GEM methodology for the survey used a multistage, stratified, probabilistic sample of 197 thousands of respondents from 70 countries to represent the world adult population between the ages 18 and 64 year. While the set of 220 variables for all countries were analyzed, the structure of data may be simplified and "beautified" with the help of ontologies. On the basis of GEM conceptual model the structure of Entrepreneurship presented on Figure 3 was obtained.

The Figure 3 illustrates the components of Entrepreneurship such as attitudes to entrepreneurship, entrepreneurial activity, and entrepreneurial aspirations as well as the factors which define them. The research studies all stages of a business' life cycle: from conception of an idea to early stages (potential entrepreneurs) (nascent entrepreneurs), when a company is in the maturation phase; and from new companies (owners of new created companies), when a company already operates in the market, to established businesses and the potential discontinuation of business. Therefore it may be logical to build another ontological model describing the Entrepreneurship that is presented on Figure 4.

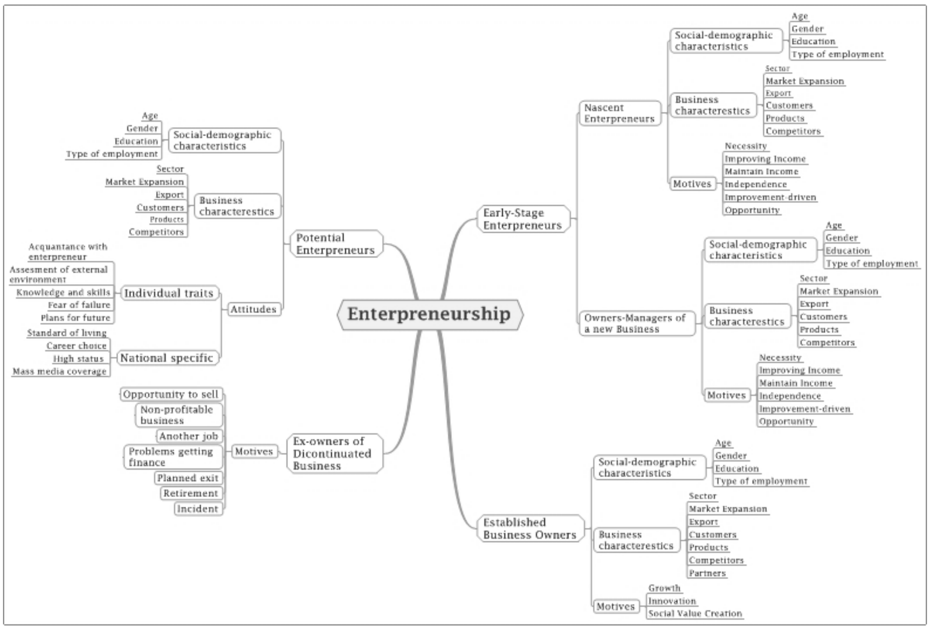

Fig. 4. Another look on GEM BigData processing with the help of ontologies

Note that with the help of ontology we united two main aspects of GEM's approach to Entrepreneurship. Assigning particular variables to each ontology segment will allow to simplify the understanding of data and its processing.

It is also worth to mention that visual models provide the opportunity to ease search of necessary valuable information and makes it possible to segment the data and analyze and compare different groups of data.

CONCLUSION

Our research stresses the role of visual ontology orchestrating for developing big data storages quickly, efficiently and effectively. We follow David Johnassen's idea of "using maps as a mind tool". The use of visual paradigm to represent and support the design process helps a professional analyst to concentrate on the domain rather than on details. The development of beautiful knowledge structures in the form of ontologies provides comprehensive support and scaffolds the users in understanding of semantics hidden in big data.

REFERENCES

[1] Big data. Changing the way businesses compete and operate. EY. 2014.

[2] Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. 2011.

[3] Big Success with Big Data. Accenture Big Success with Big Data Survey. 2014.

[4] Buzan, T. Mind Map handbook. Thorsons, 2005.

[5] Capgemini and the Economist Intelligence Unit. The Deciding Factor: Big Data and Decision-making. 2012

[6] Conlon, T. Towards Diversity: Advancing Knowledge-based Modelling with Knowledge Acquisition. In Proceedings of 8th International PEG Conference, Sozopol, Bulgaria 1997. pp. 379-386.

[7] Davies, J., van Harmelen, F., Fensel, D. (Eds.). Towards the Semantic Web: Ontology-Driven Knowledge Management. New York: John Wiley & Sons, Inc. 2002.

[8] Gavrilova, T.A., Gorovoy,V. A. Ontological Engineering for Corporate Knowledge Portal Design // In "Processes and Foundations for Virtual Organizations", Eds. L. Camarinha -Matos and H. Afsarmanesh, Kluwer Academic Publishers, 2003. - p.289-296.

[9] Gavrilova, T., Laird, D. Practical Design of Business Enterprise Ontologie. In M. Bramer and V. Terzyan (Eds.), Industrial Applications of Semantic Web, New York: Springer. 2005. pp. 61-81.

[10] Gavrilova T., Andreeva T. Knowledge elicitation techniques in a knowledge management context. Journal of Knowledge Management. 2012. Vol. 16, Iss. 4. pp. 523 – 537.

[11] Gómez-Pérez, A., Fernández-López, M., Corcho, O. Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer. 2004.

[12] Guarino, N., Giaretta, P. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing. 1998. pp. 25- 32.

[13] Iliinsky N., Steele J. Designing. Data Visualizations. O'Reilly Media. 2011.

[14] Jacobson, Ivar; Grady Booch, James Rumbaugh. The Unified Software Development Process. Addison Wesley Longman. 1998.

[15] Jonassen, D.H. Designing constructivist learning environments. In C.M. Reigeluth (Ed.), Instructional design models and strategies (2nd ed.). Mahwah, NJ: Lawrence Erlbaum. 1998.

[16] Mizoguchi, R., Bourdeau, J. Using Ontological Engineering to Overcome Common AI-ED Problems. International Journal of Artificial Intelligence in Education. 2006. 11(2). pp. 107-121.

[27] Novak, J. D., Cañas, A. J. The Theory Underlying Concept Maps and How to Construct Them. Technical Report IHMC CmapTools, Florida Institute for Human and Machine Cognition, 2006.

[18] Noy, N., McGuinness, D. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

[19] Orchestration. Merriam-Webster Online Dictionary.

[20] Soukup T., Davidson I. Visual Data Mining: Techniques and Tools for Data Visualization and Mining. John Wiley and Sons. 2002.

[21] Sowa, J. F. Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley. 1984.

[22] Tufte E. Envisioning Information. Graphics Press. 1990.

[34] Verkhovskaia O.R., Dorokhina M.V. 2014. Global Entrepreneurship Monitor. Russia 2013.