Ontological Approach to Analysis of Big Data Metadata

Julia Rogushina
Institute of Software Systems
National Academy of Sciences of Ukraine
Kyiv, Ukraine
ladamandraka2010@gmail.com

Anatoly Gladun
International Research and Training Center of IT and Systems
National Academy of Sciences of Ukraine
Kyiv, Ukraine
glanat@yahoo.com

***Abstract*—Now data becomes the most valuable commodity that allows to make informed decisions in various areas of human life. In this article, we look at the features of Big Data generated by the Internet of Things (IoT) technology, and also present the methodology for Big Data processing with use of semantic modeling (ontologies) at all stages of the Big Data life cycle. Use of Big Data semantic model allows eliminating such contradictions in these technologies as the heterogeneity of devices and things that causes the heterogeneity of the data types produced by them.**

Machine learning is used as an instrument for analyzes of Big Data: it provides logical inference of the rules that can be applied to processing of information generated by smart home system. In this methodology, the authors propose the use deep machine learning, based on convolutional neural networks because this model of machine learning corresponds to processing of unstructured and complex nature of the IoT domain.

This approach increases the efficiency of IoT Big Data processing and differs from traditional processing systems by using NoSQL database, distributed architectures and semantic modeling. We propose the conceptual architecture of the Big Data processing system for IoT and describe it on example of the NoSQL database for the smart home. This architecture consists of five independent levels. A combined approach of semantic modeling and data mining methods can be used at each of these levels. Currently, this platform can be combined with a lot of open source components.

***Keywords*—Big Data, ontology, metadata**

I. Introduction

"Big Data" is a term that refers to a group of technologies oriented on obtaining of qualitatively new knowledge from large amounts of data that cannot be handled by traditional methods and serve. Exponential growth of data generated in electronic form and stored in data banks determines the actuality of such technologies.

We can consider some set of data as Big Data (and analyze it with Big Data technologies) if it has one or more of following features named 5V: Volume – great amounts of data that require specialized means of processing; Velocity – great speed of new data generation and transformation; Variety – different data formats and types that complicate data integration; Veracity – messiness or trustworthiness of the data that cannot be converted into information; Value – big parts of data may are not useful for any users.

The analysis of large data sets is an interdisciplinary task that combines mathematics, statistics, computer science and special knowledge of the domain. This direction of IT is closely allied with intelligent information systems (IIS) and applied aspects of artificial intelligence (II).

For effective practical use of Big Data we need to analyze them at the semantic level with use of domain knowledge. Today mankind generates more and more Big Data volumes. However, this information has no direct value, but is obtained only as a result of data analysis. Obtaining of knowledge from Big Data uses machine learning (ML) [2] that summarizes experience of some system stored electronically and tries to improve the behavior of this system.

ML results are not true but probabilistic and statistic, and their quality depends on how much the data processed are close to those used in practice. This fact defines that selection of pertinent sets of Big Data is very important step of it's analysis. Metadata of Big Data can be used as a main information source that characterized the semantics of data content. The major problems in Big Data technology [1] deal with semantics are: use for data integration; detection of Big Data sets pertinent to user task; and removal of data ambiguity (for example, homonymy). The solutions of these problems use metadata linked with Big Data are [2]. Although metadata management has been known for decades, but processing of Big Data requires development of new strategies and approaches.

II. Problem definition

Method of Big Data metadata analysis allows to select the task-pertinent data sets from heterogeneous sources and data warehouse on semantic level with the help of domain knowledge. The natural language (NL) part of such metadata (unstructured or semistructured annotations, descriptions, etc.) is ambiguous, and this fact causes the need in methods of ambiguity resolution (for example, for homonyms and polysemantic terms). We propose to match such metadata with user task description by methods of NL analysis enhanced with Big Data ontology contained domain-specific knowledge for semantic processing of Big Data metadata and their links with the domain concepts. Development a prototype of such ontology is also a part of this work.

III. Metadata Used for Big Data Description

Metadata is a special kind of information resources, their creation often requires considerable effort and substantial costs, but they significantly increase the value of the data and provide extended opportunities for their use. Metadata is defined as a structured, coded data that describes the characteristics of various (text, multimedia, etc.) objects that facilitates the identification, detection, evaluation and management of these objects. Metadata describes the meaning and properties of information in order to improve management, understanding, classification and exploitation of this information.

Metadata used for Big Data description is a data block physically joined to Big Data in its storage. This metadata provides information on the characteristics and structure of Big Data set: name; the origin of data, data source information; information about the author and date of the data creation; data size and format, control total; number of dataset records; image resolution; a brief description of the data etc. [3]. It is important to note that all changes of Big Data state initiate changes of metadata. The structure, functions and properties of metadata depend considerably on the particular technological realization and on the features of the described resources, as well as on the scope and specificity of applications. However, the interpretation of the term "metadata" is not defined unambiguously. Now specialists use a lot of different definitions of metadata. The most significant of them are: metadata is data about data [4]; metadata is information that makes the data useful [5]; metadata is machine-processed data that describes some resources, both digital and non-digital [6]; metadata is information that implies its computer processing and interpretation of digital and non-digital objects by people [7]; metadata is structured information that describes, explains, indicates location and, thus, facilitates the retrieval, use and management of information resources [8]; metadata in the Web is semistructured data, usually agreed with the corresponding models that provide operational interoperability in a heterogeneous environment [9].

IV. Integration of Artificial Intelligence with Big Data Analysis

The great number of publications in this sphere shows a high interest to use of methods of traditional artificial intelligence (AI) and intelligent Web technologies to acquisition of knowledge from Big Data. Most often, researchers work in direction of ML use and integration of ontological analysis for various phases of Big Data analysis to apply the domain knowledge. The Ontology Summit 2017 "AI, learning, reasoning and ontologies" [10] analyses the use AI methods of for ML, logical inference and ontological analysis focused on Big Data and integrates various research approaches in this area divided into some groups:

- Application of ML for extraction of knowledge and improving of domain ontologies – creation and improvement of sufficient domain knowledge (knowledge bases and ontologies) about the world for a truly intelligent agent, the use of automation and various ML approaches to knowledge extraction and ontological analysis;
- Usage of domain knowledge to improve results of ML – challenges and role of background knowledge and ontologies in improvement of ML results, the requirements for ontologies used in ML for various data sets (in particular, for Big Data);
- Integration of ontological analysis with logical inference – the reasoning techniques and mechanisms oriented on ontological knowledge representation in various forms.

Background knowledge in Big Data is processed by ontological analysis and logical inference by ML means to prepare data for training and testing (reduction of large, noisy data sets to managed ones) and eliminating the ambiguity of terms.

Before the learning phase of ML we have to define such input information:

- Description of solved task;
- Target function of ML that depends from objectives of system's behavior improvement (for example, increasing of the recognition accuracy, expansion the number or class of identified objects, acceleration of recognition);
- Data source that contains information required for analysis, its type, origin and structure (information received from the system experience of interaction with one user or with the entire community of users, information received from one or more external sources, etc.);
- Methods and means that provide integration of the obtained results with the existing knowledge of system.

Quite often Big Data for analysis is obtained from various external sources. Velocity of Big Data analysis depends on the amount of information being processed. So prior filtering of information decreases the time of it's analysis. For example in case of analysis of the television streams we can analyze not all of them but only the selected part of the programs pertinent to user's problem.

If we plan to use the external experience presented in Big Data then we have to find relevant Big Data sources. We can do it with the help of the metadata that accompanies Big Data by analyzing of metadata semantics. Automatically generated part of the metadata does not contain enough information about content semantics. Available metadata is technical information that characterizes the time of the content creation, its volume, formats, etc., but does not relate to the information content of the data. This makes it impossible to provide a uniform description of the data semantics. But a big part of Big Data is accompanied by annotations or explanations, usually provided in natural language. Required information from Big Data can be retrieved by analysis of their annotations. Therefore, matching of annotations with task definition determines the pertinence of certain arrays of Big Data to this task. Big Data annotation from metadata is matched with the user's task description. Such matching is executed on the stage of data retrieval and selection, because direct comparison of Big Data content with this description inappropriate due to the extremely large volume and absence of structuring. Various annotations of Big Data are created in process of it's storing into the repositories. All types of annotations described Big Data on different levels can be considered as unstructured or semistructured NL texts. Therefore we apply to them standard tools of NL analysis similar to the Web search. Unfortunately, in the general case such retrieval problem is not solved effectively, and therefore it is advisable to apply a priori additional knowledge about Big Data domain. Analysis of scientific reports and corresponding standards shows that despite the high interest in Big Data and variety of technological means for their processing, today any metadata standards specific to Big Data is not developed. Such situation is caused by the complexity and variety of Big Data.

Metadata improves data analysis (OLAP, OLTP, Data Mining) by understanding of the data source domain in order to ensure adequate computation and interpretation of results. It provides the use of general terminology for interaction with user.

V. Metadata Standards Applicable to Big Data

Taking into account the lack of specific for Big Data standards for metadata, it is reasonable to analyze the existing metadata standards used for information that can have 5V properties and able to represent the content semantics. The standards of ISO/IEC 11179 series define metadata as data that defines and describes other data. This means that the metadata is data, and data becomes metadata when they are used in this way. This occurs in specific circumstances, for specific purposes, with defined prospects. A set of circumstances, goals, or prospects for which some data is used as metadata is called a context. Thus, metadata is data in some fixed context.

Metadata is stored in some database that is organized with the use of any formal meta-model that describes metadata. For example, the conceptual model defined in ISO/IEC 11179-3 is a meta-model in this content. A significant part of Big Data is multimedia information. We analyze some widely used standards used for describing of multimedia semantics. Now many various formats for multimedia representation are developed by different software and hardware manufacturers, but there is no unique standard common to everyone, because each manufacturer develops its own convenient approach that can subsequently be disseminated. Existing formats for saving multimedia in electronic form (GIF, TIFF, PIC, PCX, JPEG, PNG, etc.) differ in methods of information compression, encodings types, purpose of use etc.

Much of them are not oriented on describing of multimedia semantics and deal only with technical characteristics of multimedia, and only some of them represent the meaning and subject domain of data. The Moving Picture Experts Group for the Joint Standardization Committee propose a family of multi-media standard MPEG [11]. Some of them (MPEG-1 (ISO/IEC 11172), MPEG-2 (ISO/IEC 13818), MPEG-4 (ISO/IEC 14496)) deal only with compression of multimedia information. Other ones describe the semantics of multimedia content.

Standard MPEG-7 ("Multimedia Content Description Interface" ISO/IEC) [12] describes the semantic aspects of multimedia content with different degree of attention to details. MPEG-7 proposes the fixed set of descriptors for different types of information that formalize the defining of descriptors and their interconnections.

Multimedia descriptive differ for various domains and applications because the same content can be described at different abstraction levels through different properties relevant to the scope of use. For example, a graphic image at the lowest level of abstraction can be described by size, number of color, forms and positions of objects, while the upper level will contain semantic information connected graphical elements with domain concepts| Usually high-level descriptions of multimedia are represented by non-structured or semistructured NL text. For example, "Red dog named Lada de Mandraka stands near the black car". There may also be intermediate levels of abstraction. The level of abstraction is related to the way of information obtaining: many low-level properties can be extracted automatically, while high-level properties require human participation.

MPEG-21 [13] is a "Multimedia Framework" is oriented on for semantic search. It is developed for representation of content management infrastructure in a distributed environment. This standard defines the basic syntax and semantics of multimedia elements, dependencies between them and the operations that they support. It is serving to establish interoperability between multimedia information resources.

RDF (Resource Description Framework) [14] is a part of the Semantic Web project designed for creating semantic metadata for various types of information. RDF is intended to standardize the definition and use of Web metadata resources, but it is also applicable to the description of Big Data. It uses the base data model "object – attribute – value". RDF Schema gives a possibility to define a specific dictionary for RDF data and specify the types of objects to which these attributes can be applied, that is, mechanism of RDF Schema provides a basic system of types for RDF models. RDF standard is extensible and can specify the structure of the source description by using and extending the built-in concepts of RDF schemes (classes, properties, types, etc.).

Standards for describing typical resources help to simplify and unify the creation of meta-descriptions. The most well-known set of elements for metadata creation "Dublin Core Metadata Elements" [15] can be used for description of the Big Data sets.

VI. Big Data Ontology

Ontologies in knowledge engineering are used for formal and declarative description of some domain [16]. A wide range of ontologies available through the Web confirms the popularity of this approach among various groups of developers and users of Web applications, including Big Data. Such ontologies differ by the volume, expressive means, purpose, degree of knowledge formalization, etc. [17]. Domain ontology is the part of the domain knowledge limited the meaning of terms that do not depend on changing part of domain knowledge. It determines the agreements about domain terms [18].

Big Data ontology contains classes for selection of typical for Big Data information objects (video, audio, streaming video, semistructured data from sensors) with sets of relevant semantic properties. Examples of classes are Big Data formats and source types; examples of properties are geographical location, time of creation, size, annotation. Big Data ontology allows to represent the semantics of links between individual Big Data fragments (temporal, geographic, communicational (for example, information about communications by smartphones), by device identifiers, by subject, by purpose, etc.). It fixes also the quality parameters of Big Data such as noise, accuracy, degree of trust to the source, signal quality, completeness, etc.

The individuals of Big Data ontology can be matched with the individuals of user task ontology to search the pertinent sources for analysis.

To use ontological knowledge for comparing such information objects as annotations – unstructured NL texts – it is necessary to provide mechanisms for linking elements of their content with ontology terms. Such mechanism can use the task thesaurus (dictionary of the basic concepts of language linked with separate words or phrases with certain semantic connections between them [19]) based on the domain ontology [20]. Task thesaurus is a set of concepts necessary to describe and solve a problem for which the user is trying to find some information by analysis of some Big Data set. Thesaurus concepts can be imported from domain ontology. Thesauri are used in semantic markup of NL texts [21]. The similarity of Big Data annotation and user task is estimated by the semantic proximity between their thesauri.

VII. Conclusions

The analyzing the existing means of Big Data description shows the lack of generally accepted standards for metadata representation. Therefore, we propose to match the natural language annotations of Big Data with user task with the help of ontological representation of knowledge about Big Data and task domain. Prototype of Big Data ontology formalizes information about Big Data structure, sources, data sets, etc. and provides filtering of data pertinent to particular user task. This ontology includes elements of standards for description of various information types that can be used for Big Data representation.

References

[1] Marz, N., Warren, J. (). Big Data: Principles and best practices of scalable real-time data systems. New York; Manning Publications Co.2015

[2] Bizer C., Boncz P., Brodie M.L, Erling O. The meaningful use of Big Data: four perspectives – four challenges *SIGMOD Rec. 40 (4)* 2012. pp.56–60.

[3] Smith K., Seligman L., Rosenthal A., Kurcz Ch., Greer M., Macheret C., Sexton M., Eckstein A. Big Metadata: The Need for Principled Metadata Management in Big Data Ecosystems *Proceedings of the Company DanaC@SIGMOD*, Snowbird, UT, USA, 2014. – P. 46-55.

[4] Jeusfeld M.A. Metadata Encyclopedia of Database Systems, Springer, 2009, 3, p.1723-1724. http://www.springerlink.com/content/h241167167r35055/.

[5] Grotschel M., Lugger J. Scientific Information System and Metadata. Konrad-Zuse-Zentrum fur Informationstechnik, Berlin. – http://www.zib.de/ groetschel/pubnew/paper/groetschelluegger 1999.pdf

[6] Halshofer B., Klas W. A Survey of Techniques for Achieving Metadata Interoperability *ACM Computing Surveys*, Vol. 42, No. 2, 2010.

[7] Metadata Standards and Applications. Introduction: Background, Goals, and Course Outline. ALCTS. – http://www.loc.gov/catworkshop/courses/ metadatastandards/pdf/MSA Instructor Manual.pdf.

[8] Uniform Resource Identifier (URI): Generic Syntax. – http://tools.ietf.org/html/rfc3986 .

[9] Lagose C. Metadata for the Web. Cornell University. CS 431 – March 2, 2005.

[10] Baclawski K., Bennett M., Berg-Cross G., Fritzsche D., Schneider T., Sharma R., Westerninen A. Ontology Summit 2017 communiqué–AI, learning, reasoning and ontologies. Applied Ontology, 2018, P.1-16. – http://www.ccs.neu.edu/home/kenb/ pub/2017/09/public.pdf .

[11] MPEG-21 Multimedia Framework, Introduction, ISO/IEC, http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm.

[12] MPEG-7 Overview, ISO/IEC, 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm

[13] MPEG-21 Overview v.4, 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm.

[14] Lassila O., Swick R. Resource Description Framework (RDF) Model and Syntax Specification *W3C Recommendation*. – http://www.w3.org/TR/REC-rdf-syntax.

[15] Dublin Core Metadata Elements http://www.faqs.org/rfcs/rfc2413.html.

[16] Gruber T., What is an Ontology? – http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.

[17] Guarino N. Formal Ontology in Information Systems *Formal Ontology in Information Systems. Proc. of FOIS'98*, 1998. – P. 3-15.

[18] Gladun A., Rogushina J., Schreurs J. Domain Ontology, an Instrument of Semantic Web Knowledge Management in e-Learning *International Journal of Advanced Corporate Learning (iJAC)*, V. 5, Issue 4 (2012). – P.21-31. – http://online-journals.org/i-jac/article/view/2288.

[19] ISO 25964-1:2011, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval / Geneva: International Organization for Standards, 2011.

[20] Gladun, A., Rogushina, J. (2012). Use of semantic web technologies and multilinguistic thesauri for knowledge-based access to biomedical resources *International Journal of Intelligent Systems and Applications*, 4(1), 11.

[21] Gladun, A., Rogushina, J., Valencia-García, R., Béjar, R. M. (2013). Semantics-driven modelling of user preferences for information retrieval in the biomedical domain *Informatics for health and social care*, 38(2), 150-170.

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ДЛЯ АНАЛИЗА МЕТАДАННЫХ BIG DATA

Рогушина Ю. В., Гладун А. Я.

Работа посвящена разработке онтологических средств анализа Big Data, к которым невозможно применить традиционные аналитические подходы, основанные на решениях бизнес-аналитики и системах управления базами данных.

Авторы представляют метод анализа метаданных, описывающих Big Data, который позволяет выбирать те блоки информации среди разнородных источников и хранилищ данных, которые пертинентны задаче пользователя. Большое внимание уделяется сопоставлению аннотаций (естественно-языковой части метаданных) с текстом, описывающим задачу. Предлагается использовать для этого средства анализа естественного языка и онтологию Big Data, содержащую знания о специфике этой предметной области.

Использование искусственного интеллекта и интеллектуальные веб-технологии повышают эффективность всех этапов обработки Big Data. Распознавание текстовой части метаданных выполняется на основе онтологии Big Data, которая содержит знания о их специфике. Разработан прототип такой онтологии, представлена архитектура интеллектуальной системи сопоставления аннотаций Big Data с использованием тезаурусов.

Received 21.12.18