Use of Ontologies for Metadata Records Analysis in Big Data

Julia Rogushina[1], Anatoly Gladun[2], Serhii Pryima[3]

[1] Institute of Software Systems of National Academy of Sciences of Ukraine, Kyiv, Ukraine
[2] Institute of Special Communication and Information Protection of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
[3] Tavria State Agrotechnological University, Melitopol, Ukraine
ladamandraka2010@gmail.com, glanat@yahoo.com, pryima.serhii@gmail.com

Abstract. Big Data deals with the sets of information (structured, unstructured, or semi structured) so large that traditional ways and approaches (based on business intelligence decisions and database management systems) cannot be applied to them. Big Data is characterized by phenomenal acceleration of data accumulation and its complication. In different contexts Big Data often means both data of large volume and a set of tools and methods for their processing. Big Data sets are accompanied by metadata which contains a large amount of information about the data, including significant descriptive text information whose understanding by machines lead to better results of Big Data processing.

Methods of artificial intelligence and intelligent Web-technologies improve the efficiency of all stages of Big Data processing. Most often this integration concerns the use of machine learning that provides the knowledge acquisition from Big Data and ontological analysis that formalizes for domain knowledge for Big Data analysis.

In the paper, the authors present a method for analyzing the Big Data metadata which allows selecting those blocks of information among the heterogeneous sources and data repositories that are pertinent for solving the customer task. Much attention is paid to the matching of the text part of the metadata (metadata annotations) with the text describing the task. We suggest to use for these purposes the methods and instruments of natural language analysis and the Big Data ontology which contains knowledge about the specifics of this domain.

Keywords: Big Data, metadata, domain ontology, thesaurus, natural language text, homonymy, multimedia data, standard.

1 Introduction

The term "Big Data" refers to a group of technologies and methods oriented on analysis and processing of large amounts of data (both structured and unstructured) that cannot be processed by traditional methods; they serves to obtain qualitatively new knowledge. The actuality of this IT direction is determined by the exponential growth in the amount of data generated in electronic form and stored in data banks for future use. The analysis of large data sets is an interdisciplinary task that combines mathematics, statistics, computer science and special knowledge of the domain.

For effective practical use of Big Data we need to analyze them at the semantic level with use of domain knowledge.

2 Big Data Definition

A particular set of data should be considered as Big Data if it has one or more of following features named 5V:

- *Volume* refers to the vast amounts of data generated every second that require specialized processing facilities;
- *Velocity* refers to the speed at which new data is generated and the speed at which data moves around.
- *Variety* refers to the different formats and types of data which makes its integration difficult;
- *Veracity* refers to the messiness or trustworthiness of the data that cannot be converted into information and, therefore, have no value;
- *Value* – only part of data may be useful for users.

The main types of Big Data are: structured data (SQL databases); semi structured data (information security instructions, customer profile data, Web server logs, websites, emails, etc.) and unstructured data (audio files, video files, images, information cubes, etc.) that can be stored in NoSQL databases. Big Data provides the binding of geographically distributed data sets, taking into account operations such as replication and sharring (split into fragments).

Moreover, Big Data combines various unrelated data sets, processing large volumes of unstructured data (the part of such data in the total amount of Big Data is the largest).

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 processing and analysis. Due to the enormous volumes and velocity of information receiving, such processing can be performed only automatically. The knowledge obtained by processing may have a practical value of such type:

- rules built by means of machine learning;
- results of these rules application to the analysis of new data.

Examples of the first type knowledge are the decision tree for the task of medical diagnostics or the multilayer neural network that identifies people by their photographs from the social network. Examples of the second type knowledge are the

diagnosis for particular patient based on the decision tree and the recognition by the neural network of the person whose image was received by the social network user.

Obtaining of this knowledge from Big Data is based on statistical processing and machine learning (ML) [2]. Without examining in detail the methods and possibilities of ML, it should be noted that machine learning is a synthesis of some system experience stored electronically to further improve the behavior of this system that becomes more effective.

ML results are probabilistic and statistic, their quality depends on how much the data processed are close to those used in practice. Thus, the actual problem is finding exactly those arrays of Big Data, which are pertinent to a specific user's task (containing implicitly necessary knowledge), reliable, actual and qualitative. These Big Data parameters are not evaluated directly, but through an analysis of their metadata.

3 State of Art in Big Data

The major problems that exist today in Big Data technology and require solution are defined in [3] :

1. The problem of data integration that can be presented as a combined problem that requires: (1) determining the problem to be solved with the help of Big Data; (2) detection (search) of relevant parts of the data in Big Data repositories and sources; (3) executing ETL (extraction, transformation, loading) in appropriate formats and storing data for further processing; (4) removal of data ambiguity (for example, homonymy); (5) data processing for solving the problem.

2. The problem of heterogeneity overcoming between different sets of Big Data. Semantics can be considered as a means of creating a bridge between heterogeneous data.

3. The problem of open data linking.

4. The problem of semantics use for data integration and for development of database management systems. Moreover, semantics can be used in the existing system to identify data inconsistencies, to generate new knowledge using a logical inference machine, or to link specific data that are not related to machine learning methods more precisely.

As the analysis of scientific publications [4] shows, today the problems deal with metadata used in Big Data are the most acute than ever. More and more organizations realize that their business efficiency depends from robust (workable) metadata of Big Data that obtains the necessary context and the origin of key data assets.

Although metadata management has been known for decades, today new strategies and approaches are being developed:

- support for the continuous development of data processing environment;
- search for more effective ways of business management with metadata.

Therefore, we need to review the strategies and techniques of working with the metadata available to modern organization, and to find out how to build successful strategies for the adoption and use of metadata.

Organizations and companies are interested in two types of Big Data. 1. data created by people, which are mainly distributed through the Web (social networks, cookies, emails, online television, online broadcasting, etc.). 2. heterogeneous data generated by various electronic devices. For example, such technologies as Internet of Human and Internet of Things generate mixed traffic of Big Data that is used co-operatively for predicative analysis to knowledge acquisition for understanding, planning and anticipation of actions for these systems. At the same time, the question of data quality is urgent. In fact, since Big Data is characterized by large volumes, it is "raw" by nature. Therefore, a solution of this problem is required.

4 Problem Definition

It is necessary to develop a method of Big Data metadata analysis which allows selecting those blocks of data among heterogeneous Big Data sources and data warehouse that are pertinent for solving the customer's problem. It should be born in mind that both the task definition and the annotations of Big Data are natural language (NL) unstructured or semistructured texts. Therefore their matching can be based on methods of NL analysis but with the Big Data ontology, which contains knowledge about the specifics of this domain and allows semantical processing of other elements of Big Data metadata (to match the parameters of the metadata structure with the domain concepts). Creating a prototype of such ontology is also a part of this work.

5 Directions of Integration of Intelligent Web Technologies with Big Data Processing

The analysis of publications proves a high interest in the application of artificial intelligence (AI) methods and intelligent Web technologies to the Big Data processing. Most often, such integration refers to the use of ML for knowledge acquisition from Big Data and ontological analysis – to apply the domain knowledge to Big Data analysis. Development of their models and methods, as well as assessment of their efficiency, is one of the priority directions of current scientific research.

The interest in this is evidenced by the materials of the Ontology Summit 2017 "AI, learning, reasoning and ontologies" [5] devoted to the issues of use AI methods of for ML, logical inference and ontological analysis focused on Big Data. Following areas were considered:

- ML usage to extract knowledge and improve 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 background knowledge to improve machine learning results – appraisal of challenges and role of background knowledge and ontologies in improvement of ML results, the requirements for ontologies used in ML.
- Usage of ontologies in logical reasoning and vice versa – the reasoning techniques and mechanisms oriented on ontological knowledge representation in various forms.

Ontological analysis and logical inference in Big Data processing by means of ML provides the use of background knowledge to prepare data for training and testing (reduction of large, noisy data sets to managed ones) and eliminating the ambiguity of terms.

Learning phase of ML needs in definition of:

- task that is solved by computer system;
- direction of system's behavior improvement (for example, increase the recognition accuracy, expand the number of identified persons, accelerate recognition);
- sources of data that contains the information required for analysis (from the experience of the interaction of this system with a specific user or with the entire community of users, from external sources, from similar systems, etc.);
- means of integration of the obtained results with the system knowledge.

If we need to use the external experience presented in Big Data then we have to find relevant Big Data sources. Such finding uses the metadata that accompanies Big Data and analyze its semantics. Part of the metadata that is generated automatically does not contain enough information about content. Possibility to obtain the necessary knowledge from Big Data is defined by semantic analysis of their annotations.

Such annotations that are created in process of Big Data storing in the corresponding repositories. They can be considered as unstructured or semistructured NL texts and we can apply to them standard tools of NL analysis similar to the Web search. Unfortunately, in the general case such problem is not solved effectively, and therefore it is advisable to apply a priori additional knowledge about Big Data domain.

Despite the high interest in Big Data and variety of technological means for their processing, there are no metadata standards specific to Big Data. The reason for this is the complexity and variety of Big Data.

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. Therefore, matching of annotations with task definition determines the pertinence of certain arrays of Big Data to this task.

If organization analyzes the Big Data that accumulates in the process of its own operation there is no need for such a comparison. But quite often Big Data for analysis is obtained from various external sources. Big Data analysis is based on ML methods which velocity depends on the amount of information being processed. Prior

filtering of content decreases the time of it's analysis. For example in case of analysis of the television streams it is better not to process all of them but first select the part of the programs pertinent to user's problem. The source of annotations of such Big Data is a TV guide.

Another example of Big Dataderived from various sources and annotated only by NL descriptions is the information resources on the availability of job vacancies offered by the European Employment Services (EURES) which brings together about 400 "euro-advisers" from national employment services, associations of employers, trade unions, local and regional authorities and higher education institutions; they are actively used by the ESCO (European Skills, Competences, Qualifications and Occupations), the multilingual classifier of European Skills, Competences, Qualifications and Professions. Their annotations can be filtered with the help of competence descriptions, semantically marked by the domain ontology concepts.

6 Metadata for Big Data

Metadata is a structured, coded data that describes the characteristics of media objects that facilitates the identification, detection, evaluation and management of these objects. Metadata is used to describe the meaning and properties of information in order to better understand, classify, manage and exploitation the data.

Metadata for Big Data is a data block physically joined to Big Data in it's storage. This metadata provides information on the characteristics and structure of Big Data set: name; the origin of data, data source information; of the source; XML tags indicating the author and date of the document creation; attributes indicating the size and format, control total; number of dataset records; resolution of image; a brief description of the data etc. [6, 7].

The properties of metadata, its composition and functions depend considerably on the technological realization, on the features of the resources they describe, as well as on the scope and specificity of applications.

The vast number of publications is devoted to the metadata, however, the interpretation of the term "metadata" has not been formed completely yet. 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.

Now a lot of metadata definitions are used by specialists. We have chosen the most significant ones: metadata is data about data [8]; metadata is information that makes the data useful [9]; metadata is machine-processed data that describes some resources, both digital and non-digital [10]; metadata is information that implies its computer processing and interpretation of digital and non-digital objects by people [11]; metadata is structured information that describes, explains, indicates location and, thus, facilitates the retrieval, use and management of information resources [12]; metadata in the Web it is semistructured data, usually agreed with the corresponding models that provide operational interoperability in a heterogeneous environment [13].

Metadata contributes to improving the quality of data, which is determined by the following characteristics: consistency (whether the submission is homogeneous, or whether there is duplicate data that is overlapping or conflicting); completeness (whether all data is available); accuracy (coincidence of saved and actual values); timeliness (whether the current saved value is relevant). Metadata also provides improved data analysis (OLAP, OLTP, Data Mining), where it is necessary to understand the domain of the data source in order to ensure adequate computation and interpretation of results. Metadata provides the use of general terminology and language of interaction within the organization, eliminating ambiguity and ensuring the consistency of conclusions within the company.

Big Data processing is closely linked to it's metadata, especially for semistructured and unstructured data. It is important to note that all changes of Big Data state initiate changes of information about the origin immediately recorded as metadata. The goal of obtaining the origin and the life cycle of the data is the possibility of argumentation of analytical results: similar to scientific research, if the results cannot be justified and repeated, they do not deserve trust.

Thus, effective processing of Big Data and acquisition of valuable knowledge demand a flexible framework for management of it's metadata-based processing. It allows to creating universal environment for interoperability of heterogeneous data blocks, to standardize the processing stages and to develop the processing platforms.

7 Metadata Standards Applicable to Big Data

Matching of Big Data annotation from *metadata* with the user's task description is carried out t 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.

In the standards of the ISO/IEC 11179 series, metadata is defined 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, without which data is not metadata. A set of circumstances, goals, or prospects for which some data is used as metadata is called a *context*. Thus, metadata is data in a given context.

Metadata can be stored in a database and be organized with the use of any model. The model describing metadata is called a *meta-model*. For example, the conceptual model presented in ISO/IEC 11179-3 is a meta-model in this content.

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.

A significant part of Big Data is multimedia information. Now a lot of 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, types of encodings, purpose etc.

The Moving Picture Experts Group for the Joint Standardization Committee propose a family of multi-media standard MPEG [14]. Some of them (MPEG-1 (ISO/IEC 11172) [15], MPEG-2 (ISO/IEC 13818) [16], MPEG-4 (ISO/IEC 14496)) deal only with compression of multimedia information [17]. Other ones describes the semantics of multimedia content.

MPEG-7 ("Multimedia Content Description Interface" ISO/IEC) [18] is a semantic multimedia-oriented standard. It assumes a different degree of attention to details in its descriptions. MPEG-7 contains description tools – DT (Description Tools); Definition Language DDL (Description Definition Language) and System Tools. It defines a standard set of *descriptors* for different types of information, standardizes the way of defining its descriptors and their interconnections. DT contains two components: descriptors that define the syntax and semantics of each property (metadata element), and the *description schemes* that set the structure and semantics of the relationships between their components, which can be either descriptors, or description schemes.

Since descriptive possibilities have to be unambiguously and completely interpreted in the application context, they can differs for different user domains and different applications, that is the same material can be described through different types of properties that are relevant to the scope of use and application possibilities. For example, a graphic image at the lowest level of abstraction can be described by form, size, texture, colour, palette, trajectory and position; and audio can be described through tone, tempo change, position in the soundtrack, while the upper level will contain semantic information "This is a scene with a green car going along the road on the left and a dog crossing the road on the right, accompanied by a background sound of rain". 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. In many cases multimedia resources are described by non-structured or semistructured NL text. However, the problem with dependent of these descriptions from particular NL. This is especially important for processing of names, titles, places, etc. The Description Tools of MPEG-7 allow creating content descriptions (that is, a set of description schemes DS and the corresponding descriptors D) containing information on the creation and use of content; reality displayed in the content; set of objects, etc.

MPEG-21 [19] is a "Multimedia Framework" standard designed to create a content management infrastructure in a distributed environment for semantic search. It 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.

Such metadata are available for representation semantics of multimedia Big Data. NL annotations of the semantic content of the material should be included in these meta-descriptions.

RDF (Resource Description Framework) is another promising approach to creating semantic metadata for various types of information created within the Semantic Web project. RDF is intended to standardize the definition and use of Web metadata resources, but it is also applicable to the description of Big Data. RDF 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.

An important feature of the RDF standard is extensibility: RDF gives a possibility to specify the structure of the source description by using and extending the built-in concepts of RDF schemes, such as classes, properties, types, collections. The RDF model scheme includes inheritance of classes and properties.

Certain patterns and standards for describing typical resources are provided to users to simplify and unify the creation of resource meta-descriptions. The most thoroughly developed set of elements for metadata creation is "Dublin Core Metadata Elements" [20].

8 Lifecycle of Big Data Analysis

Big Data analysis differs from traditional data analysis primarily because of the characteristics of the processed data, such as volume, velocity and diversity. In order to meet the various requirements for implementing Big Data analysis, a step-by-step methodology is required for organizing activities and tasks related to the acquisition, processing, analysis and reuse of data. The traditional life cycle of the Big Data Analysis can be divided into the following stages, as shown in Fig. 1: 1) Assessment of the task that solution requires results of Big Data analysis; 2) Identification of data (internal, external, location); 3) Data collection and cleaning; 4) Data extracting (receiving, forwarding, entry into the data bank); 5) Data inspection and clearing; 6) Data aggregation and submission for analysis; 7) Data analysis; 8) Data visualization; 9) Use of the analysis results.

We changed some stages of the traditional life cycle of Big Data analysis by adding Semantic Web elements, in particular deal with ontological modeling of domain knowledge, into certain stages.

At the stage of *data identification* the data sets necessary for carrying out analytical projects (tasks) and their sources are defined. Domain ontology helps in appropriated data sources – for example, NoSQL data warehouses with some relevant to task characters (geographic, temporal, domain-specific) and data types. Detecting a wider range of relevant data sources may increase the likelihood of detecting hidden regularities and correlations in Big Data.

At the stage of *data collection and cleaning*, the final formation of Big Data packages for the purposes of the task is accomplished using semantic analysis of metadata text annotations and the selection of relevant data sets for solving the problem. Semantic approach is used for selection of Big Data sets that are relevant to the user's task.

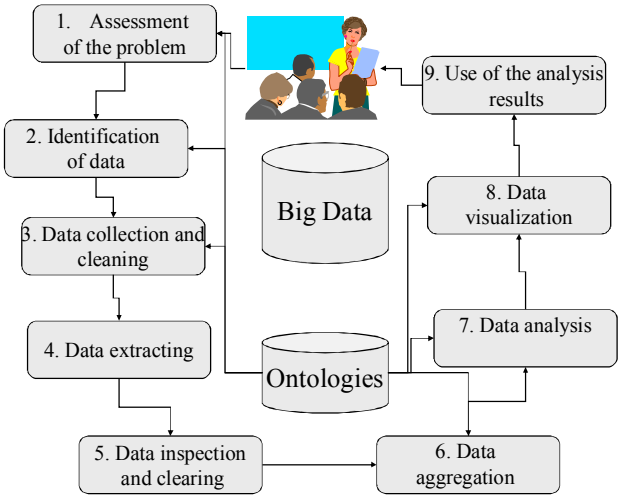

Fig.1. Life Cycle of Big Data Analysis

Some data identified as input data for analysis may come in formats that are incompatible with the Big Data application. This is especially true for data from external sources. The stage of the lifecycle of *data extraction* is designed to extract incomparable data and convert it into format that the base Big Data software can use to analyze the data.

The stage of *data inspection and clearing* is designed to create complex rules for checking and deleting any known inaccurate data (duplicate data, data omissions, excess data, etc.). For package analysis, data inspection and clearing can be performed using the offline ETL operation. For real-time analysis, a more complex internal memory system is needed to validate and clear data as they flow from the source.

The stage of *data aggregation and presentation serves* for consolidation of data sets that can be distributed across multiple data sets through common fields, such as by date or identifier (ID). In other cases, the same data fields can be displayed in multiple data sets. In any case, you need a data convolution method or you need to define a data set representing the correct value. Completion of this stage can be complicated due to differences in: data structure – although the data formats may be the same, the data structure model may differ; semantics – the value marked differently in two different sets of data may mean the same thing, for example, "surname" and "last name". At this stage domain ontology can be used for matching of various names of the same concepts, for determining of relations between them (hierarchy, synonymy, semantic closeness, etc.).

The stage of *data analysis* use Data Mining and ML techniques for generation of new knowledge by Bid Data processing. Ontologies are used on this stage for integration of these new rules with concepts of user task domain.

It is necessary to note the importance of the first two stages of this life cycle – the task setting, for which the Big Data analysis is carried out, and the selection of the Big Data set, which is pertinent to this task. Supposing these steps are unsuccessful, then, despite the complexity and effectiveness of the data analysis methods, the results will not meet the user's needs.

9 Ontologies and Big Data

In knowledge engineering, ontology is understood as a detailed description of some problem area, which is used for the formal and declarative definition of its conceptualization [21]. Often ontology is called the knowledge base of a special type, which can be divided, alienated and used independently in the framework of the considered domain [22]. Now use of ontologies as adequate means for describing different domains is a generally accepted fact, and 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 applications with Big Data.

Such ontologies are describes in various languages and associated with a wide variety of domains. They differ by the volume, expressive means, purpose, degree of knowledge formalization, etc. Classifications of ontologies differ in the parameters the classification an in general, can be divided into two groups – semantic and pragmatic. Semantic classifications group ontologies according to the parameters connected with the content of information: domain; the degree of formality of the knowledge presented; the level of expressiveness and the level of information detailed description [23]. Pragmatic classifications group ontologies according to the purposes of their development and sphere of use.

Domain ontology is the part of the domain knowledge that limits the meaning of its terms which do not depend on another (changing) part of knowledge of this domain. Such domain ontology can be considered as a set of agreements on the domain, and the rest of domain knowledge is a set of empirical and other laws of this area. Thus, the ontology determines the degree of agreement of terms by the specialists in this domain [24].

Different sources offer different formal models for ontology representation. However, every of them contains set of terms (notions, concepts) that can be divided into set of classes and set of instances; set of relations between concepts where some relation groups (relations "class-subclass", hierarchical and taxonomic relations and the synonymy relations) can be clearly distinguished, as well as functions – a special case of relations for which the n-th element of the relations is uniquely determined by n-1-th previous elements; axioms and functions of interpreting concepts and relations.

To build an ontological Big Data model, it is necessary to separate the set of classes from the set of class instances. It is also advisable to separate object relations between instances of different classes from data relations, namely the relations between instances of attributes and their values. To describe the Big Data ontologies we use the following formal model:

$$O = \langle X, R, F, T, M \rangle \quad (1),$$

that contains the following elements:

- $X = X_{cl} \cup X_{ind}$ – the set of ontology concepts, where X_{cl} is the set of classes, X_{ind} is the set of class instances, such where $\forall a \in X_{ind} \exists A \in X_{cl}, a \in A$;
- $R = r_{ier\_cl} \cup \{r_i\} \cup r_{ier\_prop} \cup \{p_j\} \cup p_{ier\_prop}$ is the set of relations between elements of ontology, where
- r_{ier_cl} is the hierarchical relation between the classes of ontology – they are structures of partial ordering with the upper element Thing that can be established between classes of ontology and are characterized by such properties as antisymmetry and transitivity, $r_{ier\_cl} : X_{cl} \to X_{cl}$;
- $\{r_i\}$ is the set of object properties that establish the relationship between instances of classes: $r_i(a, a \in X_{ind}) = b, b \in X_{ind}$; $r_i : X_{ind} \to X_{ind}$;
- r_{ier_prop} is hierarchical relations between the object properties of the ontology classes;
- $\{p_j\}$ is a set of data properties that establish the relations between instances of classes and values with T: $p_i(a, a \in X_{ind}) = t, t \in T$, $p_i : X_{ind} \to T$;
- p_{ier_prop} is hierarchical relations between the properties of these instances of ontology classes;
- $F = \{F_{cl} \cup F_{prop}\}$ is a set of characteristics that can be used for logical inference above the ontology;
- T is a set of data types (for example, a line, a whole) for values of data properties of ontology classes;
- M is the set of non-logical rules of SA.

Such an ontology Big Data 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:

- different formats of devices generating Big Data;
- the purpose of these devices;
- geographical location;
- time characteristics;
- reliability of the source;
- conditions for access;
- volumes and speed of the update.

Big Data can be both created by human and generated by electronic devices, they can come from different sources and be presented in different formats or types. Therefore, Big Data ontology displays typical sources of Big Data – from the activities of people (both individuals and organizations) through information and communication equipment (from social networks, smart phones, computers, cash registers, ATMs, etc.) and from automated devices (sensors, sensor networks, camcorders, GPS, Internet of Things devices, automated productions, drones).

Ontology can also fix the quality parameters of Big Data – noise, accuracy, degree of trust to the source, signal quality, completeness, etc.

Ontology allows to represent the semantics of links between individual Big Data fragments (temporal, geographic, communicational (for example, information from smart phones that were in the conversation mode), by device identifiers, by subject, by purpose, etc.). Below (Fig. 2), examples of Big Data ontology elements corresponded to various elements of its ontological model (1) are demonstrated:

$$X_{cl} = \{"Big\_Data\_resource","standard","type","format","metadata\_format",...\}$$

$$X_{ind} = \{XXX101,...,MPEG7,...,JPG,...\};$$

$$"metadata\_fornat" r_{ier\_cl} "format" ;$$

$$\{r_i\} = \{"has\_type","has\_resource","based\_on",...\};$$

$$\{p_j\} = \{"annotation","size","date",...\}.$$

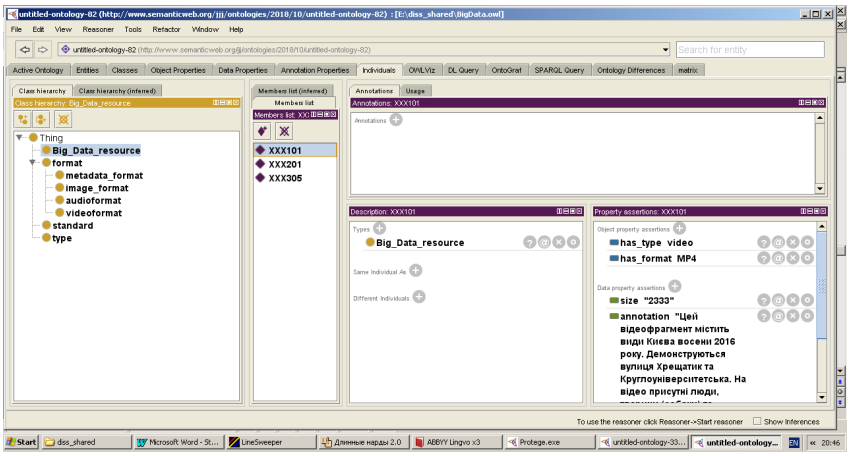

Fig.2. Elements of Big Data ontological model.

The tools of ontology visualization make it easier to analyze the relationship between its elements (Fig. 3).

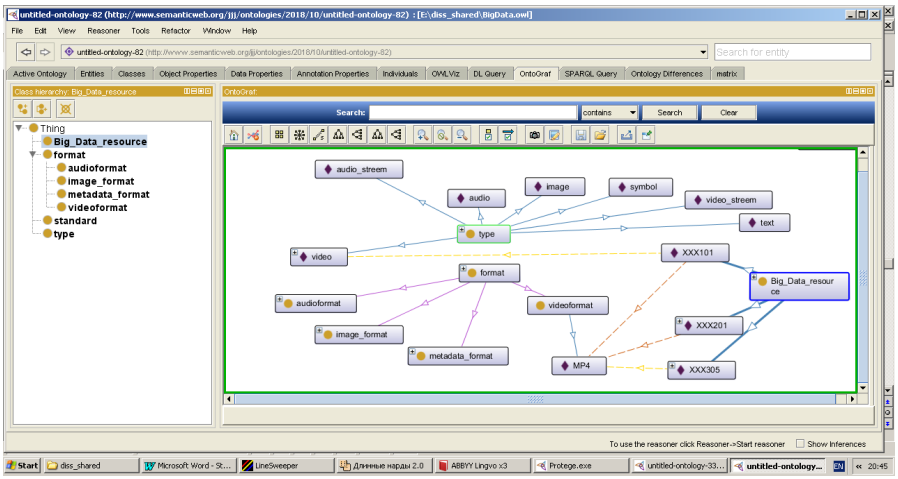

Fig.3. Visualization of Big Data ontology.

The elements of this ontology are matched with the ontology of the user's task to search the pertinent sources of Big Data.

10 Comparison of Natural Language Texts

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 a mechanism can be based the task thesaurus which represents the user needs of base of the domain ontology.

In general, a thesaurus is a dictionary of the basic concepts of language linked with separate words or phrases with certain semantic connections between them [25]. Thesaurus can be considered as a special case of ontology [26]. *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. The weight of each of thesaurus concepts characterizes the importance and pertinence of this concept for the current user task. Thesaurus concepts can be imported from domain ontology. Thesauri are used in semantic markup of NL texts [27]. The similarity of two NL texts is estimated by the semantic proximity function between their thesauri.

For search of the pertinent Big Data sets the user's task thesaurus Th_{task}, $Th_{task} = \{< t_m, w_m >, m = \overline{1,q}\}$ is compared with the thesauri of the Big Data annotations from the sets I, $I = \{annot(Big\_Data\_resource_j)\}, j = \overline{1,n}$, and the coefficients of their proximity $K_j, m = \overline{1,q}$, is calculated:

$$K_j = \sum_{m=1}^{q} f(t_m) * w_m, m = \overline{1,q}, \text{ where}$$

$$f(t_m) = \begin{cases} 0, & t_m \notin \text{annot}(\text{Big\_Data\_resource}_j) \\ 1, & t_m \in \text{annot}(\text{Big\_Data\_resource}_j) \end{cases}$$

There is assumed that $t_m \in \text{annot}(\text{Big\_Data\_resource}_j)$ if the annotation of the resource $\text{annot}(\text{Big\_Data\_resource}_j)$ has fragment of the text that in accordance with the lexical knowledge base correlates with the term of thesaurus t_m. Processed resources are ordered in dependence on the values K_j, and user receives for further analysis such Big Data sets where the value of the semantic proximity function is higher than the given value of estimation K.

11 Solution of Homonymy in Big Data Annotations

The NL ambiguity causes different interpretations of words. One of commonly encountered problems deals with homonyms. Examples of homonyms: "hyperbole" as a stylistic figure in which an attribute is exaggerated and "hyperbole" as a flat curve, "bow" of a ship – "bow" and arrow, "row" (line) – "row" (quarrel).

Recognition is also applied to various types lexical homonyms: homophones (words pronounced alike but different in meaning: too – two, here – hear, meat – meet, see – sea); homoforms (words having the same sound composition only in a certain grammatical form), homographs (words spelled alike but different in meaning).

If a word in NL text has several variants of semantic meaning, then it is necessary to choose the proper variant by the context (with the help of knowledge from domain ontology). All recognized examples are gathered into the set of precedents that can be processed by one of ML algorithm for generation of some rules for homonymy solution (e.g. the decision tree). These rules usually depend on particular NL. Information for recognition can be acquired from semantic or non-semantic Wiki-resources, as well as from dictionaries of homonyms of various natural languages. If we use Wiki then the content of the word is defined as the text of the relevant Wiki article, from which links to other articles are obtained [28]. We can use Wiki for recognition of new terminology or for new meanings of existing terms because Wiki resources are much more dynamic and up-to-date versus traditional vocabularies.

The procedure for homonym recognizing consists of the following steps:

1. The text for recognition comes to the input.
2. Pre-processing of the text (fragmentation of text on sentences and words).
3. Normalization of the text (conversion of words into the infinitive, change of endings, etc.)|
4. Comparison of every word with homonym database.
5. If word match with homonym from database then the algorithm of homonym recognition is executed.

The algorithm of homonym recognition is based on the decision tree. The decision tree (also called the tree of classifications) is used in Data Mining for predicted models. The structure of this tree contains the following elements: "leaves" and "branches". The branches correspond with values of attributes – parameters that defines value of target function, the leaves correspond with values of result (target function). In order to recognize (classify) a new case, we go down from the root of this tree according to values of case attributes to the leaf.

The process going from top to bottom is an example of an absorbing "greedy" algorithm, and today it is one of widespread strategies.

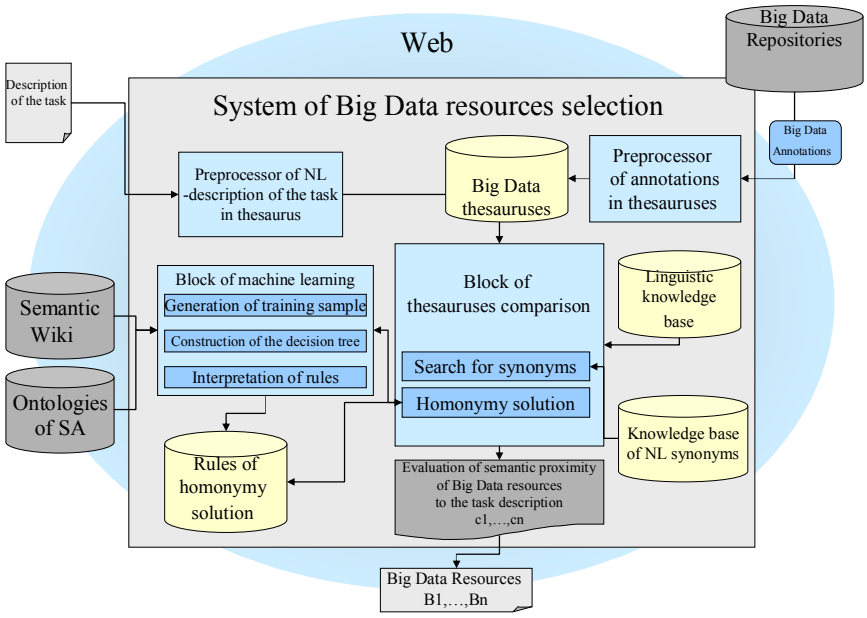

Fig. 4. Generalized architecture of Big Data annotation comparison system

The general recursive scheme of the decision tree constructing by the learning sample:

Step 1. If learning sample has examples with different results then go to 2. Else link the last branch with leave corresponded with this result and stop.

Step 2. Select one of m attributes of learning sample (various algorithms differ one from another by criteria of this selection) and link the last branch with node corresponded with this attribute A_i that has q_{A_i} values.

Step 3. Divide learning sample on q_{A_i} subsets with m-1 attributes, link the attribute node with branches corresponded with all these values and execute step 1 for all these subsets of learning sample with m-1 attributes.

In this case, attributes are linked with presence or absence of various terms (their forms, synonyms etc.) in context of description of various meanings of term-homonim in learning sample. Such learning sample is generated by Wikipedia, other Wiki resources, vocabularies and definitions. Size and pertinence of learning sample defines the quality of recognition.

Decision rule based on the decision tree: "if the word {c1, c2 ... cn} is combined with the word D in the text, then the word D is selected". The intelligence system of annotation comparisons (Fig. 4) has to perform these actions.

12 Conclusions

The results of analyzing the existing means of Big Data description show the lack of generally accepted standards for such metadata representation. Therefore, the proposed methods for the analysis of natural language annotations of Big Data are by far the most adequate means of comparing the semantics of Big Data sets with particular user tasks to select the pertinent information for analysis.

References

1. T. Erl, W. Khattak, & P. Buhler (2016). Big Data Fundamentals. Prentice Hall: Upper Saddle River, NJ, USA.
2. N. Marz & J. Warren (2015). Big Data: Principles and best practices of scalable real-time data systems. New York; Manning Publications Co.
3. C. Bizer, P. Boncz, M.L. Brodie, O. Erling, The meaningful use of Big Data: four perspectives – four challenges, SIGMOD Rec. 40 (4) (2012) 56–60.
4. H. Abbes, F. Gargouri: M2Onto: an approach and a tool to learn OWL ontology from MongoDB database // Madureira, A.M., Abraham, A., Gamboa, D., Novais, P. (eds.) ISDA 2016. AISC, vol. 557, 2017. – Pp. 612–621. doi:10.1007/978-3-319-53480-0_60.
5. K. Baclawski, M. Bennett, G. Berg-Cross, D. Fritzsche, T. Schneider, R. Sharma, A. Westerninen 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 .
6. K. Smith, L. Seligman, A. Rosenthal, Ch. Kurcz, M. Greer, C. Macheret, M. Sexton, A. Eckstein "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.
7. A. Dey, G. Chinchwadkar, A. Fekete, Ramachandran K. Metadata-as-a-Service //in Proceedings of the 31st IEEE International Conference on Data Engineering Workshops (ICDEW), 2015. – P.6-9.
8. M.A. Jeusfeld Metadata // Encyclopedia of Database Systems, Springer, 2009. – 3. 1723-1724. – http://www.springerlink.com/content/h241167167r35055/
9. M. Grotschel, J. Lugger Scientific Information System and Metadata. Konrad-Zuse-Zentrum fur Informationstechnik, Berlin. – http://www.zib.de/groetschel/pubnew/paper/groetschelluegger 1999.pdf
10. B. Halshofer, W. Klas A Survey of Techniques for Achieving Metadata Interoperability // ACM Computing Surveys, Vol. 42, No. 2, 2010.

11. Metadata Standards and Applications. Introduction: Background, Goals, and Course Outline. ALCTS. – http://www.loc.gov/catworkshop/courses/ metadatastandards/pdf/MSA Instructor Manual.pdf
12. Uniform Resource Identifier (URI): Generic Syntax. – http://tools.ietf.org/html/rfc3986 .
13. Lagose C. Metadata for the Web. Cornell University. CS 431 – March 2, 2005.
14. MPEG-21 Multimedia Framework, Introduction, ISO/IEC, http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm .
15. MPEG-1, ISO/IEC, 1996. – http://mpeg.telecomitalialab.com/standards/mpeg-1/mpeg-1.htm
16. MPEG-2, ISO/IEC, 2000. – http://mpeg.telecomitalialab.com/standards/mpeg-2/mpeg-2.htm
17. Overview of the MPEG-4 Standard, ISO/IEC, 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-4/mpeg-4.htm
18. MPEG-7 Overview, ISO/IEC, 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
19. MPEG-21 Overview v.4, 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm.
20. Dublin Core Metadata Elements http://www.faqs.org/rfcs/rfc2413.html .
21. T. Gruber What is an Ontology? – http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.
22. N. Guarino Formal Ontology in Information Systems // Formal Ontology in Information Systems. Proc. of FOIS'98, 1998. – P. 3-15.
23. L. Obrst, W. Ceusters, I. Mani, S. Ray, B. Smith The evaluation of ontologies // Semantic Web, Springer US, 2007. – P.139-158. – http://philpapers.org/archive/OBRTEO-6.pdf.
24. A.Y. Gladun, J.V. Rogushina Semantic technologies: principles and practics. – K.: ADEF-Ukraine, 2016. – 308 c. [in Ukrainian]
25. ISO 25964-1:2011, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval / Geneva: International Organization for Standards, 2011.
26. A. Gladun, & J. Rogushina (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.
27. A. Gladun, J. Rogushina, 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.
28. J. Rogushina Semantic Wiki resources and their use for the construction of personalized ontologies // CEUR Workshop Proceedings 1631 , 2016. – P.188-195.