Semantic processing of metadata for Big Data: standards, ontologies and typical information objects

© Julia Rogushina[1][0000-0001-7958-2557], © Anatoly Gladun[2][0000-0002-4133-8169]

[1] Institute of software systems of National Academy of Sciences, Kyiv, Ukraine,
[2] International Research and Training Center of Information Technologies and Systems of National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine,
Kyiv, Ukraine,
ladamandraka2010@gmail.com, glanat@yahoo.com

Abstract. We consider the main aspects of semantic technologies applied to obtain information from Big Data and use of these technologies in representation and analysis of metadata that describe Big Data. existing metadata standards and their usage for Big Data are analyzed. Data Mining methods usage allows to acquire the necessary knowledge from unstructured elements of such metadata. Background knowledge about user task can be used for structuring of metadata elements and acquisition of information about information objects required by user. Ontological approach provides interoperability of such background knowledge. We propose to use semantic Wiki resources as a source of personified ontologies that represents current user needs in Big Data analysis.

Keywords: Big Data, ontology, metadata, semantic Wiki

1 Introduction

The main goal of Big Data metadata usage is to describe their context, content and structure, as well as managing methods and accumulation of their history. Metadata management has to provide data protection from loss, unauthorized deletion, saved or destroyed, and access to such management must be organized through the allocation of access rights and compliance with certain security rules. The Big Data semantics is reflected usually in unstructured natural language (NL) descriptions that are part of metadata, therefore the processing of such information requires much more effort than the processing of structured information. It causes an interest to research work directed at analysis of Big Data meta-descriptions structuring with use of existing metadata standards.

2 Metadata and their properties

Metadata in the broadest sense is data that provides information about other data. But such a definition is too simple and unconstructive. Metadata has various purposes. It helps users find relevant information and discover resources. It also helps organize electronic resources, provide digital identification, and archive and preserve resources. Wikipedia defines metadata as data from a formal top-level system that describes some data system as structured data that characterizes certain entities for their identification, retrieval, evaluation and management [1]. Metadata is a separate type of information resources (IR) that require specific means of representation, creation and processing (here IR is comprehended as any entity that is capable to transmit and store information or knowledge [2]).

Although creation of metadata was originally intended only to describe data, recently metadata is used to describe a wide variety of IRs and information objects (IOs) described by IRs – conceptual diagrams, ontologies, thesauri, cloud services, IoT devices, organizations, etc. Use of metadata allows to characterize the life cycle of information, actions and needs of different subjects of data processing. Metadata can be used to specify the semantics of IRs and IOs (e.g., to describe the model of domain at the semantic level with use of ontologies, taxonomies, thesauri, etc.) that can be processed by intelligent information systems (AAS). Conceptual modeling languages of different levels of formality (RDF, OWL, OWL2 ontology languages) are used for such specifications.

Metadata were used long before the advent of computer systems in various archives and library catalogs. Examples of such metadata are bibliographic descriptions of the literature sources used in publications and annotations of articles. It should be noted that such "pre-computer" metadata mainly characterize NL IRs but can be used for other IO (e.g., film annotations).

The development of information technologies causes a significant expansion of metadata functions and their diversity. Metadata representation and management tools created for various types of IOs and information systems (IS) with various functions depends both on the purpose of these ISs and their implementations. The semantics of metadata, their functions and means of their representation are determined by the information technologies used to create such IRs, and the specifics of the software and IRs processed by these IRs. The spread of electronic libraries [3] where IRs of different types are stored, data repositories and knowledge bases that implement Semantic Web technologies [4] has caused increasing interest in the semantization of metadata [5]. In addition, IR interoperability and reuse cause the need in metadata exchange between different ISs.

Although a huge number of publications have been devoted to metadata in recent years. Problems related to the creation of metadata systems for specific research domains and areas of activity, development of electronic libraries, various repositories of digital information objects, information systems for specific areas of application are most often discussed. but systematic analysis of metadata general properties and functions is not in focus of attention. Moreover, researchers use a large number of non-integrated metadata definitions that reflect different points of view on this term and on the scope of metadata use. Comparison of such definitions and specifics of their application is considered in [6].

Metadata is information that makes data useful [7]. This definition describes the scope of metadata, but is too general for practical use. E.g., for Big Data it defines the role of metadata, but does not specify the requirements for ways of their practical representation. Metadata is intended both for computer processing and for human interpretation of information about digital and non-digital objects [8]. In [9], metadata is defined as structured data that contains the characteristics of the entities for the purposes of their identification, retrieval, evaluation and management. Metadata can be considered as a structured data that describes IR characteristics.

These properties of metadata influence on ways of their representation, that is, metadata provides the basis for data structuring, but such structuring can be both formalized (computer-oriented) and informal (user-oriented). Thus, the use of metadata is directly related to the processing of unstructured and semistructured information and to determine the level of data structuring. It should be noted that the metadata used to describe Web IRs are usually semistructured, but they correspond to some standards and consistent models which ensure their operational interoperability in a heterogeneous environment [10].

In [11] metadata is defined as any descriptive information about other data sources that contributes to the organization, identification, presentation, location, interoperability, management and use of this data. Thus, metadata not only describes the composition of data, their structure and characteristics (place and time of storage, format, etc.), but also characterizes the information technology that supports this data, access methods and users.

Metadata definition proposed in [12] characterizes not IR as a whole, but a certain element of data related to this IR. This approach dealt with IOs is most in line with the specifics of Big Data storing in large repositories because it provides identifying of data subset with selected properties that is pertinent to a particular user task.

If metadata determines the semantics of information then it can be used to improve its search and retrieval, understanding and use. E.g., [13] considers the use of ontologies and thesauri for semantic annotation of IR and their elements oriented on machine learning and knowledge acquisition techniques. Depending on the purposes of annotation, ontologies of different complexity can be used (from controlled dictionaries and glossaries to ontologies with complex set of relations). Thesauri and controlled dictionaries can be used in document comments. These dictionaries are not completely formal (e.g., the semantics of relations between their terms are not defined formally), and such annotations usually indicate terms from a limited dictionary that can be used to improve search. Examples of such dictionaries are MeSH (Medical Subject Headings) (http://www.nlm.nih.gov/mesh/meshhome.html) and TGN, (http://www.getty.edu/ research / tools / vocabulary / tgn / index.html).

Dublin Core (http://www.dublincore.org/) is an example of a lightweight ontology that is used to specify the characteristics of electronic documents. Now this ontology is most widely applied for metadata semantization.

FOAF (Friend of a Friend) (http://www.foaf-project.org/) is an other lightweight ontology aimed at creating an annotated network of homepages for people, groups, companies, etc. It is implemented in RDF Schema and contains such basic classes as agent, person, organization, group, project, document, image, as well as some basic properties of instances of these classes.

The OntoWeb (http://www.ontoweb.org/) and KnowledgeWeb (http://knowledgeweb.semanticweb.org/) ontologies describe people, organizations, projects, publications, etc., and are commonly used for meta-descriptions of publications of international conferences of the EU.

The specific composition of metadata functions depends on the characteristics of IS that uses them, the nature of IR processed by this IS and of IOs described by this metadata, the basic information technology of the system, the needs of IS users and many other factors. But in general we can consider some fixed set of metadata properties and define their values for every concrete implementation.

Metadata properties:

1. *Relativity* of IR differentiation into data and metadata – metadata for one IR can be considered as data in another, and vice versa. E.g., the ontology used to annotate NL text is an element of metadata, and the same ontology in the ontology repository [14] is data.

2. *Multilevel description* of IR properties: any IR can be described in terms of some more abstract system of concepts that can form a hierarchy of levels, i.e. lower-level metadata is described with use of higher-level metadata. Such metadata hierarchy can include arbitrary number of levels. E.g., the Meta Object Facility (MOF) standard [15] has three levels and Dublin Core has two levels.

3. *Heterogeneity* of IRs, IOs and data that can be described by metadata: the properties that allow characterizing metadata depend on the specifics of the data and the scope of their use.

4. *Alienability* of metadata from IR: metadata can be stored independently and autonomously or be embedded in the IR that they characterize. Examples of autonomous metadata: database schemas, ontology repositories. Examples of embedded metadata: HTML-markup Web pages, Wiki-markup, annotation of the article in the document, glossary contained in the text of the standard specification. If unstructured metadata is embedded in NL document, it is necessary to extract this information from the text for computer processing.

5. The degree *of content dependence* is determined by the specifics and meaning of the metadata itself. E.g., the date of creation and storage of document, the type of file in which it is contained, page URLs do not depend on the content. Document annotation, statistical characteristics of text (frequency characteristics of word occurrences of the dictionary, text length etc.) are determined by the content.

6. The degree of *domain dependence* is determined by the objectives of metadata creation: meta-descriptions created by expressive means of standards that are not focused on any specific domain can be used extensively but such approach reduces the expressiveness of metadata. Specialized meta-descriptions (e.g., in various fields of scientific research, museum work and education) can be used only in certain domain but take into account the specifics of the data of this domain. Therefore domain-specific metadata standards are developed for domain where metadescriptions play a significant role in data processing.

7. The degree of *structuring* is determined by focus of metadata processing. Metadata can consist of structured and unstructured elements. Metadata oriented on human users can be both structured and non-structured, e.g., they can be represented by NL texts (annotations, abstracts, etc.). Computer processing causes the use of structured metadata that can be analyzed automatically in different ISs.

8. The level of *granularity* of the IR description determines what elements of IR are described by metadata. These elements can be collections of IRs as a whole, individual IRs, documents and individual elements of documents – IOs contained in these documents. E.g., metadata can describe portal based on Wiki technology, individual Wiki pages and IOs from these pages represented by Wiki templates.

9. The degree of *dynamism* is determined by the conditions under which and how often metadata can be changed. E.g., the database schema is relatively unchanged and static, while the catalogs of the electronic library change when new IOs are included into them. Changes of metadata can be caused by changes of existing data (additions, deletions, edits) or by new records added while preserving the previous ones.

10. The degree of *formalization* is determined by the means used to represent metadata. Metadata representation can be based on expressive means with varying degrees formalized – natural languages, semi-formal languages (e.g., NL with a limited vocabulary), a set of metadata elements from some ontology (e.g., from the Dublin Core of FOAF), formal languages (e.g., OWL language [16]).

There are many other properties of metadata that can be considered in various studies (e.g., presentation tools, storage methods, and explicit representation), but they are not fundamental for describing Big Data and therefore we do not considered them in this paper.

Such set of properties provides the basis for comparison of various solutions and their matching with user aims. We can compare existing metadata systems from the viewpoint of user task that has been solved on base of information described by this metadata. It can help in identifying which metadata schemas should be applied in order to meet the needs of the information creators and users.

E.g., if we try to select some subset of Big Data for analysis then the granularity is defined by type of IOs that we want to process.

Unfortunately, present metadata systems have some common disadvantages [17] that complicate data processing:

- low efficiency of information recovery in metadata systems;
- inconsistent input of metadata changes, which causes to inconsistencies and duplication of information;
- insufficient automation of the metadata management;
- focusing on work with one fixed type of information (IRs or IOs with fixed set of properties);
- lack of a unified metadata model for all types of IOs;
- lack of common understanding of the base unit of metadata description – an instance of metadata, which is described by a set of parameters that do not intersect with other sets described by other metadata;
- incompleteness of metadata elements set which usually do not contain information about the means of data processing and storage.

3 Structured and unstructured data

Unstructured data (USD) is information that does not have a predefined data model with formally fixed elements and relations [18]. Such uncertainty leads to problems related to its storage (traditional databases are not designed for such uncertainty) and analysis. Today USD is potentially the greatest source of new knowledge, and the volume of USD influences the accuracy of the results. The properties of USD and the means of their processing are analyzed in more detail in [19]. The most usable part of USD consists of NL texts.

NL information is a sequence of words of natural language (NL) of arbitrary length, combined according to poorly formalized linguistic rules and represented in electronic form. Such data can be analyzed as USD due to the fact that structural elements contained into NL data are not explicitly represented, and therefore their acquisition requires a lot of time and efforts.

If some elements of metadata do not have a formalized structure, then acquisition of the necessary information from this element needs in use of methods that focus on the analysis of USD. The main part of Big Data semantics is represented by NL annotations and descriptions, and therefore analysis of Big Data meaning and score of use is defined by methods NL-oriented processing of USD.

USD analysis can use semantic markup to associate IR or IO elements with ontology elements (e.g., a fragment of NL text is associated with a class or instance of an ontology class, and another element is associated with the value of its property). But direct application of ontologies for IR semantics is not easy because the main part of users is not specialists in ontological analysis. Therefore, it is more useful to use simpler markup instruments, such as semantic Wiki markup. Such semantic wikification can be performed by both domain experts and technical staff. In this case, the problem of IR semantization is divided into the set of simpler subtasks:

1. selection or construction of the domain ontology of O, used as a structure of the IR database (this subtask is provided by knowledge engineers in collaboration with domain experts);

2. generation of basic elements of semantic markup according to selected ontology and rules of their application database (this subtask is executed by knowledge engineers);

3. creation of semantically marked IRs (this subtask is executed at first by domain experts, and further can be delegated to technical staff);

4. construction of more universal rules for marking arbitrary IRs in accordance with the semantics of the selected domain, in the most difficult case - establishing links with the elements of the ontology to replenish and improve it (this subtask is provided by knowledge engineers in collaboration with domain experts).

It should be noted that this approach, despite its simplicity, has a significant drawback - semantic Wiki-markup of IR built for some domain can not be applied automatically for another domain. Therefore, analysis of Big Data metadata from viewpoint of one user task cannot be use for other task of other user without additional processing.

Problem can be solved by use of high-level ontologies and taxonomies as a base of integration of semantic markup concepts from different domains. Pertinent knowledge for such ontologies can be acquired from various online encyclopedias based on semantic Wiki technologies (e.g., the portal version of the Great Ukrainian Encyclopedia e-VUE [20]). .

Semantic markup also allows to analyze the semantic similarity between the concepts of the selected ontology and use it further for the analysis of USD [21].

4 Metadata for Big Data

The properties of metadata, their composition and functions significantly depend on the IS implementation technologies and characteristics of IRs described by them, as well as on the domain and application specifics.

Some set of data can be considered as Big Data if it has one or more of the so-called "5V" characteristics: volume; speed; diversity; certainty; value [22]. Metadata that characterizes Big Data may contain information about the source of the data; about the author and the date of the document creation; data size and format; the number of records in the data set; description of this data, etc. In Big Data processing, metadata analysis is crucial because metadata contains information not only about the origin of the data [23, 24], but also about their content. Big Data during the lifecycle needs in for metadata management to ensure effective use of information .

Metadata for Big Data [25] is structured or semi-structured information that allows to create, manage and use Big Data at different times and in different areas of activity, as well as select such Big Data sets that are relevant to the user task [26]. Big Data can contain structured data (e.g. SQL database); semi-structured data (e.g. customer profile data, Web server logs, Web sites) and unstructured data (e.g.) (NL documents, audio files, e-files, images , information cubes, etc.). This information is stored in NoSQL databases. The analysis of publications shows a high interest in the application of artificial intelligence methods and intelligent Web-technologies to Big Data processing. Most often, such integration concerns the use of machine learning (ML) to acquire knowledge from Big Data and ontological analysis - to apply domain knowledge to make learning procedures more efficient. Direct processing of Big Data by ML is not possible because of time restrictions (big volume of data requires very long work of ML algorithms) but it can be improve by background knowledge about domain rules and relations.

Various natural and artificial languages are used to describe metadata. NLs are the richest and most expressive means of metadata representation (e.g., annotations of publications, various information about IRs and their authors) but they are oriented on human for users and are not intended for computer processing because they do not provide unambiguous and rigorous interpretation of metadata.

Artificial languages used to describe metadata are formal and available for automated processing. In the broadest sense, any formalized languages can be used for this purpose which allow to explicitly fix the structure of metadata. Examples of such languages: languages of DBMS data description , conceptual modeling, description of ontologies, business processes, workflows; ODL language of object description; W3C consortium languages: OWL, OWL2, RDF, RDFS; markup languages (e.g. XML, HTML, XHTML); XML schema language.

5 Metadata standardization

Metadata standardization is the basis for interoperability and reuse of both the metadata itself and the IRs that characterize this metadata. Therefore, international standardization organizations pay close attention to the development of metadata formats that are designed for formal descriptions of various IR and IO types. Such standards include a set of fields (attributes, properties, metadata elements) that allow to characterize particular element of data. Such standards can be used (with different efficiency) to describe Big Data. Therefore, it is advisable to consider in more detail the currently developed standards for metadata, which should be applied to Big Data, and monitor their improvement in order to use these standards to make at least a partial USD structuring.

In Ukraine today three international standards dealt with metadata (ISO 15489-1: 2016 [27], ISO 15836-1: 2017 [28], ISO 15836-2: 2019 [29]) are accepted as national standards by the confirmation method [30, 31].

Standard ISO “15489-1: 2016 Information and documentation - Records management - Part 1: Concepts and principles” defines the basic concepts and principles of document and information management. It describes the information fields that are part of the metadata structure. For Big Data, these fields allow to display the following information: Big Data content description (form, format, relations between Big Data blocks, etc.); environment of Big Data creation; relations with other Big Data blocks (sharing, replication) and metadata; identifiers and information required for data extraction and representation; actions and events related to these Big Data (date, time of action, change of metadata, etc.).

Standard ISO “*15836-1: 2017 Information and documentation - The Dublin Core metadata element set - Part 1: Core elements”* describes Dublin Core elements that use for IR descriptions. In this standard, a resource refers to any object that can be identified (e.g., in the field of computer science, resources are represented by individual documents, texts, audio and video files, Web pages, databases, etc.). Big Data and their metadata also meet this definition and can be considered as resources.

The resource life cycle is a sequence of events that involve the resource creation and use. The 15-element "core" specified in this standard is part of a larger set of metadata dictionaries and technical specifications supported by the Dublin Core Metadata Initiative (DCMI) [32]. The basic elements can be used in combination with metadata terms from other compatible dictionaries in the context of application profiles, as indicated in the abstract DCMI [DCAM] model.

Standard ISO “*15836-2: 2019 Information and documentation - The Dublin Core metadata element set - Part 2: DCMI Properties and classes”* expands and supplements the first part of this standard. It provides programmers with a common universal language for metadata creating and analyzing. This universal language provides an extended description of metadata elements, using their updated properties and classes. The ISO 15836-2 standard increases the initial set from 15 basic properties to 40 properties and 20 classes to increase the accuracy and expressiveness of the descriptions in the Dublin Core. Standard focuses on describing the general properties of metadata elements required for basic interoperability between different programming languages and domains of their use. The properties and classes described in this standard are intended to be used in conjunction with metadata terms from other compatible dictionaries in the context of application profiles.

6 Metadata and typical information objects

Analysis of modern metadata systems shows that they allow to describe not only IRs as a whole, but also IOs typical for a certain domain described in these IRs. Typical information objects (TIOs) are characterized by the set of semantic properties described in the metadata of each instance of such TIOs.

TIOs can describe both IRs and IOs (documents, database elements, multimedia information) and real-world objects (personalities, organizations, geographical objects, etc.) [33]. TIO is created according to domain and task specifics: if IS processes a certain number of elements with a similar set of properties and characteristics, then it is advisable to allocate a separate TIO for them.

Use of TIOs allows to classify metadata information with a complex structure at the semantic level: e.g., the values of some elements of Dublin Core metadata can be attributed to certain TIOs (Table 1) that determine the rules of their analysis and processing. Structure of such TIOs can be defined by classes and instances of classes of appropriate domain ontology instead of NL-descriptions.

Table 1. TIC of Dublin Core metadata elements

Name	TIO
title	Domain concept
creator	Personality, Organization, Service TIOs
subject	Domain concept
description	NL USD
publisher	Personality, Organization, Service TIOs
contributor	Personality, Organization, Service TIOs
Date	Structured data Date type
type	TIO (concept from ontology “Resources”)
format	TIO (concept from ontology “Data types”)
identifier	Reference
source	Reference
coverage	TIO (concept from ontology “Geographical objects”)
language	TIO (concept from ontology “Languages”)
relation	Reference
rights	NL USD

The structure and relationship between TIOs can be represented in different ways. E.g., if we use domain ontologies as a source of background knowledge then TIOs correspond to classes of ontologies, and their characteristics correspond to the properties of class instances. If we use semantic Wiki-resources then structure of TIOs is imported from templates with the sets of semantic properties.

7 Use of Data Mining for Big Data metadata analysis

Now a lot of methods of knowledge acquisition are developed and realized for various types of IRs - structured, partially structured and unstructured [34]. Analysis of such methods shows that the addition of structural elements significantly reduces the decision space and processing time. Quite often such information about structure of elements is based on background knowledge about TIO. This is especially important for Big Data metadata analysis, because TIO properties are used as parameters of the data sample which values are analyzed by the methods of Data Mining [35], and therefore the correct selection of TIO determines the result of Big Data processing as a whole.

Data Mining is a process aimed at identifying new significant correlations, patterns and trends as a result of analyzing a large amount of stored data with use of pattern recognition techniques and statistical methods. Data Mining methods have become especially effective with the development and accumulation of Big Data. We can say that Data Mining is a process of automated acquisition of new knowledge from IRs. Such knowledge is implicitly presents in the processed information, but we need in semantic analysis of Big Data metadata to select such sets of data that contain pertinent knowledge.

The results of Data Mining largely depend on the data they process: on their completeness, relevance, relevance of the task and quality, and on the knowledge on the basis of which this data is selected. Therefore, if the construction of a data set is based on the analysis of their metadata, it is the composition and quality of metadata significantly determine the quality of knowledge that can be obtained from IR.

The most common uses of Data Mining are related to the tasks of classification, clustering and forecasting. Data Mining tools allow to find automatically new rules and regularities in the data, to build hypotheses about the relations between data elements. Since the formulation of the hypothesis about dependencies is the most difficult task, the advantage of Data Mining over other methods of analysis is obvious. But for their effective use, these results must be related to the appropriate conceptual apparatus, which is formalized by means of knowledge representation, e.g., with the help of ontologies [36]. In many cases, such connection is established through semantic metadata - those elements of metadata that are associated with a particular representation of knowledge, e.g., with elements of the ontology of the corresponding domain. The knowledge acquired in this way from the data, in turn, allows to improve the domain ontology of which will be further used to create metadata. Thus, the creation of metadata and their use to improve ontologies is a cyclical process that supports more efficient storage and use of data.

The Web access to data processed by Data Mining adds many specific requirements for analysis methods that causes distinguishing of separate direction. Web Mining systems [37] allow to find patterns in the Web IRs and transform Data Mining technology to analyze big volumes of unstructured, heterogeneous and distributed data. They can be used for Big Data processing but need in selection of appropriate data sets. Web Mining requires the following steps:

- input stage - receiving "raw" data from sources (server logs, texts of electronic documents or Big Data storage);
- preprocessing stage - data are represented in the form necessary for the successful construction of a model (with use of background knowledge);
- pattern discovery stage;
- model analysis stage - interpretation of the obtained results.

Data Mining provides processing of NL texts that usually contains the most useful information. The analysis of such data is also separated into a special direction of Data Mining – Text Mining [38]. This is usually done by identifying patterns and trends using statistical and linguistic methods.

The use of background knowledge about domain allows to significantly increasing the efficiency of Data Mining in all directions. But various directions differ by external knowledge that they can use. E.g., Text Mining needs iv eternal linguistic knowledge bases and thesauri, and Web Mining operates with metadata ontologies and standards, information about semantic markup etc.

One of the current areas of application of background knowledge in Data Mining is the analysis of Big Data and their metadata. This is caused by the extremely large volumes of the data themselves and their dynamics, which leads to the dynamism of the metadata that describes them. Therefore, important requirements for the methods of their analysis are the speed and availability of heuristics, which can significantly reduce the analysis time. E.g., knowledge of the "class-subclass" relation between metadata parameters allows for improved learning sampling.

Gaining of such background knowledge consists of the following subtasks:

1. search for IR that pertinent to user tasks;

2. obtaining the necessary background knowledge from these IRs;

3. use of the acquired knowledge for data analysis.

In the case of Big Data analysis, these tasks are specified as follows:

1.1. selection of the Big Data repository,

1.2. search or creation of domain ontology that contains background knowledge about the user task;

1.3. analysis of Big Data metadata in order to select a set of data that are pertinent to tasks of the user, using background knowledge of the selected domain ontology;

1.4. generation of the required data set (subsets of Big Data under certain conditions) using ontology knowledge;

2. obtaining from the domain ontology the terms and relations that are needed for a more effective analysis of a large amount of information (e.g., to reduce the number

of data parameters or to reduce the number of records under more precise conditions of the problem);

3. use of the received knowledge for analysis of the received data set and for interpretation of the received result.

Thus, ontologies allow both to semantically analyze metadata describing Big Data (e.g., to replace terms in the description of the problem with synonyms or semantically similar concepts, to narrow or expand the query) and to analyze the data itself (e.g., using restrictions on possible parameter values or influence others from some data).

8 Semantic Wiki resources as a source of background knowledge for Big Data metadata analysis

Research of methods of obtaining background knowledge, which characterizes the Big Data, is an important area of research aimed at processing such data because, as a rule, pertinent ontologies are not offered for Big Data sets by their creators.

The high time complexity, which is influenced by the large size of the feature space in Big Data, causes problems in the use of traditional methods of artificial intelligence to such information. For their optimization it is expedient to apply the available knowledge about domain which includes both Big Data and the user tasks that require information from these Big Data. Use of existing background knowledge allows not to re-acquire this knowledge and use it to logical inference derive and establish relations between the elements of Big Data metadata.

The effectiveness of this approach depends of the pertinence of the selected knowledge base to user task and of the knowledge representation means. Today, the most common solution for reusable representation distributed knowledge in terms of sharing and reuse is ontologies. But the construction and search for ontologies that are pertinent to a specific problem is a difficult problem. Search of pertinent ontology cannot be fully automated, although comparing Big Data metadata with metadata descriptions of ontologies in the repository allows pre-selection. The problem is complicated by the fact that many professionals working with Big Data and their metadata do not have sufficient experience in working with ontologies. Therefore, it is advisable to use as a source of background knowledge such IR that satisfy the following conditions:

1. simple enough for understanding of their content and properties by many users;
2. the Web accessibility;
3. use of open formats;
4. possibility of automatic generation of ontologies with a fixed set of concepts.

Semantically marked Wiki-resources satisfy all these requirements. The expressive capabilities of Semantic MediaWiki [39] – a semantic extension of MediaWiki [40] – allows to explicitly capture the content of relationships between Wiki pages that correspond to ontology classes. It is much easier for majority of users to generate appropriate ontological structures from semantic Wiki resources.

Wiki-ontology is a special case of domain ontology [41] with limited expressive possibilities that are defined by the expressiveness of Wiki markup and its semantic extension and do not involve the application of characteristics for object properties and data properties. Such ontologies can be created automatically on base of the set of Wiki pages selected by user. In addition, this approach allows to filter the information needed to problem solving and therefore such Wiki-ontologies are much more compact. Reduction of ontology classes, individuals and relations significantly decreases the time for its analysis.

An important feature of Wiki-resources that are based on Semantic MediaWiki is the ability to generate a Wiki-ontology not for the whole set of pages, but only for a specific subset selected by the user explicitly by a list of pages or by user-specified semantic query. The parameters of such a query are the categories and conditions for the values of the semantic properties of the pages.

9 Conclusions

Semantic processing of Big Data metadata allows to obtain implicit knowledge about the data. standards help in structuring of metadata elements and their processing on semantic level. To date, there are no generally accepted, universal Big Data metadata standards, and the most commonly used is the universal Dublin Core metadata description standard.

We use semantic technologies and ontologies to be able to integrate data from internal and external sources and improve Big Data management, evaluation, and interpretation to perform artificial intelligence applications. Metadata is the main source of information about Big Data throughout its life cycle. In order to properly select data sets from Big Data, it is necessary to learn to automatically extract knowledge from their metadata using semantic technologies. It is expedient to use for this purpose such external sources of background knowledge as ontologies and thesauri.

We proposed to use Wiki technologies and their semantic extension as a source of background knowledge about the user task that has to be solved with the help of information acquired from Big Data.

References

1. Metadata. – https://uk.wikipedia.org/wiki/Метадані (in Ukrainian)
2. Dublin Core Metadata Initiative. DCMI TYPE Vocabulary.– http://dublincore.org/documents/demitype-vocabulary.
3. Reznichenko, V.A., Zakharova, O.V., Zakharova, E.G. (2005). Electronic libraries: information resources and services. Problems in programming, № 4, pp.60-72. (in Ukrainian)
4. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), P.34-43.
5. Dunsire, G., Willer, M. (2011). Standard library metadata models and structures for the Semantic Web. Library hi tech news.
6. Kogalovsky, M.R. (2012). Metadata, their properties, functions, classification and presentation means. Proc. of the 14th All-Russian Scientific Conference "Digital Libraries: Promising Methods and Technologies, Electronic Collections" —RCDL-2012, 2012. http:ceur-ws.org/Vol-934/paper3.pdf (in Russian)

7. 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
8. Halshofer B. Klas W. (2010) A Survey of Techniques for Achieving Metadata Interoperability. ACM Computing Surveys, Vol. 42, No. 2, Article 7.
9. Taylor C. An Introduction to Metadata. The University of Queensland, Australia. http://www.libraty.uq.edu.au/papers/ctmeta4.html
10. Lagose C. (2005) Metadata for the Web. Cornell University. CS 431.
11. Feng L., Brussee R., Blanken H., Veenstra M. (2007) Languages for Metadata. In: Multimedia Retrieval. Data-Centric Systems and Applications, Springer, P.23-51. http://www.springerlink.com/ content/m276p88003533q86/..
12. Jeusfeld M.A. (2009) Metadata. Encyclopedia of Database Systems, Springer, P. 1723-1724. http ://www. springerlink.com/content/ h241167167r35055/.
13. Corcho O. (2006) Ontology based document annotation: trends and open research problems /Intern. Journal of Metadata, Semantics and Ontologies. - Volume 1, Issue 1, January 2006. http://www.dia.fi.upm.es/~ocorcho/documents/IJMSO2006_Corcho.pdf .
14. Gladun A., Rogushina J. (2013) Repositories of ontologies as a means of knowledge reuse for recognition of information objects. Ontology of design, No. 1 (7), P.35-50. (in Russian)
15. Overbeek, J. F. (2006). Meta Object Facility (MOF): investigation of the state of the art. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.4092&rep=rep1&type=pdf.
16. OWL Web Ontology Language. Overview. W3C Recommendation: W3C, 2009. – http://www.w3.org/TR/owl-features/.
17. Kobelev, A.E., Vyazilov, E.D. (2010). Modern approaches to metadata creating. Modern problems of remote sensing of the Earth from space, 7 (4), P.194-203.http://d33.infospace.ru/d33_conf/sb2010t4/194-203.pdf.
18. Unstructured_data. – https://en.wikipedia.org/ wiki/Unstructured_data.
19. Rogushina J. (2019) Means and methods of unstructured data analysis. // Problems in programming, № 1, P.57-77. http://pp.isofts.kiev.ua/ojs1/article/view/348/346.
20. Andon P., Rogushina J., Grishanova I., Reznichenko V., Kyrydon A., Aristova A., Tyschenko A. (2020) Experience of the semantic technologies use for intelligent Web encyclopedia creation (on example of the Great Ukrainian Encyclopedia portal). Problems in programming, № 2-3. P.246-258. (in Ukrainian)
21. Rogushina J. (2019) Use of Semantic Similarity Estimates for Unstructured Data Analysis CEUR Vol-2577, Selected Papers of the XIX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2019), P.246-258. – http://ceur-ws.org/Vol-2577/paper20.pdf.
22. Demchenko, Y., De Laat, C., Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. In 2014 International Conference on Collaboration Technologies and Systems (CTS), P. 104-112.
23. Smith, K., Seligman, L., Rosenthal, A., Kurcz, C., Greer, M., Macheret, C., Eckstein, A. (2014). "Big Metadata" The Need for Principled Metadata Management in Big Data Ecosystems. Proc. of Workshop on Data analytics in the Cloud, P.1-4. https://www.researchgate.net/profile/Arnon_Rosenthal/publication/264382894_Big_Metadata_The_Need_for_Principled_Metadata_Management_in_Big_Data_Ecosystems/links/53da96030cf2e38c6338161f/Big-Metadata-The-Need-for-Principled-Metadata-Management-in-Big-Data-Ecosystems.pdf.
24. Dey, A., Chinchwadkar, G., Fekete, A., & Ramachandran, K. (2015) Metadata-as-a-service. 31st IEEE International Conference on Data Engineering Workshops, P. 6-9.

25. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile networks and applications, 19(2), 171-209.
26. Rogushina J., Gladun A., Pryima S. Use of Ontologies for Metadata Records Analysis in Big Data // Selected Papers of the XVIII International Scientific and Practical Conference "Information Technologies and Security" (ITS 2018). CEUR Vol-2318. – http://ceur-ws.org/Vol-2318/paper5.pdf.
27. ISO 15489-1:2016 Information and documentation — Records management — Part 1: Concepts and principles.
28. ISO 15836-1:2017 Information and documentation — The Dublin Core metadata element set — Part 1: Core elements.
29. ISO 15836-2:2019 Information and documentation — The Dublin Core metadata element set — Part 2: DCMI Properties and classes.
30. DSTU ISO 15489-1: 2018 Information and documentation. Records management. Part 1. Concepts and principles (ISO 15489-1: 2016, IDT). (in Ukrainian)
31. DSTU ISO 15836-1: 2018 Information and documentation. Dublin Core Metadata Element Set. Part 1. Basic elements (ISO 15836-1: 2017, IDT). (in Ukrainian)
32. Weibel, S. L., & Koch, T. (2000). The Dublin core metadata initiative. D-lib magazine, 6(12), 1082-9873.
33. Rogushina J. (2019) The use of thesauri to search for complex Web information objects based on ontologies. Problems of programming, № 4, P.11-27. (in Ukrainian)
34. Gladun A., Rogushina J. (2016) Semantic technologies: principles and practices. – Kyiv, ADEF-Ukraine, 308 p. (in Ukrainian)
35. Gladun A., Rogushina J. (2016) Data Mining: search for knowledge in data. - Kyiv, ADEF-Ukraine, 452 p. (in Ukrainian)
36. Nigro H.O. ed. (2007) Data Mining with Ontologies: Implementations, Findings, and Frameworks: Implementations, Findings, and Frameworks. IGI Global, 89 p.
37. Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. ACM Sigkdd Explorations Newsletter, 2(1), 1-15. https://arxiv.org/pdf/cs/0011033.pdf
38. Berry, M. W. Castellanos M. (2007). Survey of text mining. Survey of Text Mining:Clustering, Classification, and Retrieval. Computing Reviews, 45(9), 548.
39. Krötzsch M., Vrandečić D., Völkel M. Semantic MediaWiki// International Semantic Web Conference. 2006. Pp.. 935-942. URL: https://link.springer.com/content/pdf/10.1007/11926078_68.pdf.
40. MediaWiki. URL: https://www.mediawiki.org/wiki/MediaWiki.
41. Rogushina J. Analysis of Automated Matching of the Semantic Wiki Resources with Elements of Domain Ontologies // International Journal of Mathematical Sciences and Computing (IJMSC). 2017. Vol. 3. No. 3. Pp. 50-58. URL: http://www.mecs-press.org/ijmsc/ijmsc-v3-n3/IJMSC-V3-N3-5.pdf.