Проаналізовано проблеми, що виникають під час роботи з окремими джерелами, з використанням даних репозитаріїв та баз даних. Визначено поняття якості інформаційного продукту. Описуються основні види невизначеності. Побудовано метод оцінювання якості інформаційного продукту. Це дозволяє прогнозувати якість даних в каталозі Великих даних та отримати певні ефекти від впровадження. Зокрема, може бути підвищена ефективність пошуку залежностей у Великих даних

Ключові слова: великі дані, зменшення невизначеності, фактор ризику, Ф-залежність, корисність інформаційного продукту

Проанализированы проблемы, возникающие при работе с отдельными источниками, с использованием данных репозитариев и баз данных. Дано определение качествп информационного продукта. Описываются основные виды неопределенности. Построен метод оценки качества информационного продукта. Это позволяет прогнозировать качество данных в каталоге Больших данных и получить определенные эффекты от внедрения. В частности, может быть повышена эффективность поиска зависимостей в Больших данных

Ключевые слова: большие данные, уменьшение неопределенности, фактор риска, Ф-зависимость, полезность информационного продукта

UDC 004.052.2
DOI: 10.15587/1729-4061.2018.123064

UNCERTAINTY REDUCTION IN BIG DATA CATALOGUE FOR INFORMATION PRODUCT QUALITY EVALUATION

N. Shakhovska
Doctor of Technical Sciences, Professor,
Head of Department*
E-mail: nataliy.b.shakhovska@gmail.com
O. Vovk
PhD, Associate Professor*
E-mail: olenavovk@gmail.com
Yu. Kryvenchuk
PhD, Assistant*
E-mail: yurkokryvenchuk@gmail.com
*Department of Artificial intelligence
Lviv Polytechnic National University
S. Bandery str., 12, Lviv, Ukraine, 79013

1. Introduction

Big data information technology is the set of methods and means of processing different types of structured and unstructured dynamic large amounts of data for their analysis and use for decision support. There is an alternative to traditional database management systems and solutions class Business Intelligence. This class attribute of parallel data processing (NoSQL, algorithms MapReduce, Hadoop) [1].

Big Data features are:

– working with unstructured and structured information;

– orientation on the fast data processing;

– leads to the fact that traditional query language is ineffective while working with data.

Information objects describe a certain subject area, consolidated data and relationships between objects constitute the Big data catalogue. One of the problems that arise from the process of consolidation is the indeterminacy of data as the result of doubling, inexactitude, absence, contradictory data. Also, indeterminacy arises from the installation of wrong connections between objects. Therefore, there is a task of reduction of indeterminacy for upgrading of data.

Since the data comes from various sources, some set of data may be missing in the data source, and the other may overlap in various information products. Therefore, there is a problem of doubling, absence, imperfection, and vagueness of data.

Indeterminacy can arise at the level of attribute tuple and relation (indeterminacy in the circuit description).

The appearance of indeterminacy in the attribute and tuple due to multidimensionality display leads to the spread of uncertainty in all copies of a particular concept.

Since the Big data catalogue of millions of data items subject area, the traditional means of handling indeterminacy (interval maths, multivalent logic) become ineffective because of the large number of operands.

Thus, the specificity of Big data catalogue (the presence of a diverse set of sources, data doubling, ambiguity of describing data sources) leads to the fact that the indeterminacy in traditional relational databases is considered within a relationship and could occur at the level of attribute and tuple-level attitude in this case extends through the perception of the user information on the entire Big data catalogue. Therefore, for processing indeterminacy in the Big data catalogue, a different approach must be used, the use of which was unnecessary in relational databases and data warehouses.

The uncertainty reduction is the actual problem nowadays. First of all, we collect information from various sources and this information may be double, contradictory, etc. After that, we try to analyze this information (find dependencies, classification, clustering, etc.) Inexact information allows us to find inexact dependencies. That is why such information can't be used in decision support systems. As a result, the data availability is reduced.

That is why uncertainty reduction in Big data catalogue is an actual problem.

2. Literature review and problem statement

Classify types of indeterminacy by the nature of their manifestation in the Big data catalogue. One of the first works in this direction is [2].

In the [3], it is emphasized that indeterminacy, as the objective form of life surrounding of the real world, is conditioned, on the one hand, by the objective existence of randomness as forms of need, but on the other hand – the imperfection of each act of reflection real phenomenon in the human consciousness. Imperfection of reflection unstoppable through the universal connection of all objects of the real world and the infinity of their development. Indeterminacy is expressed in a variety of conversion possibilities in reality, the existence of the set (as a rule, endless number) of the states in which an object changes in dynamics, may be in future time.

In [4], such types of indeterminacies are defined, the nature of which is:

– value is unknown (missing);

– incompleteness of the information;

– illegibility (usage of distribution for installation of the variety of knowledge);

– the inaccuracy (concerns numerical data);

– non-determination of conclusion procedures of the solutions;

– unreliability of the data;

– multivalence of interpretations;

– linguistic indefinability.

Let us consider the indicated types of equivocations in more detail and find out places of their occurrence in relation.

Uncertainty of types 3–8 are categorized in [5] as wobble of the data and predominantly occur at a level of a tuple or subset of values of attributes.

The zero information is most often met at a level of attribute value.

The incompleteness is a condition of a tuple, in which there are missing values. It is possible to attribute an illegibility, inaccuracy and contingency to physical uncertainty, one of the sources of which is limitation exactly of numeric data types or loss of accuracy in a run time of mathematical operations (here attribute uncertainty arises owing to activity with intervals).

The unreliability and multivalence of interpretations arise in connection with inexact analysis or ambiguous mapping of objects in relation. In relation this type of uncertainty is modeled using padding attribute. The values of this attribute mean the confidence of a tuple or subset of attributes values in a tuple.

The multivalence of interpretation is one of the sources of inconsistencies.

The linguistic uncertainty is connected with usage of natural language for knowledge submission, which has a qualitative nature, and there can be related to misunderstanding of a word or misunderstanding of the contents of the proposal.

Such type of uncertainty is met in systems of text information processing (machine translation system, self-conditioning system, etc.).

The reviewed types of equivocations can be superimposed against each other or to be a source of one another.

Nowadays, the methods of elimination are missing, inexact and indistinct data [1–3] are designed. Therefore, it is necessary to elaborate methods, which can work with all types of uncertainty [6].

Uncertainty of these types may be in database, data warehouse and Big data catalogue (Fig. 2) [7].

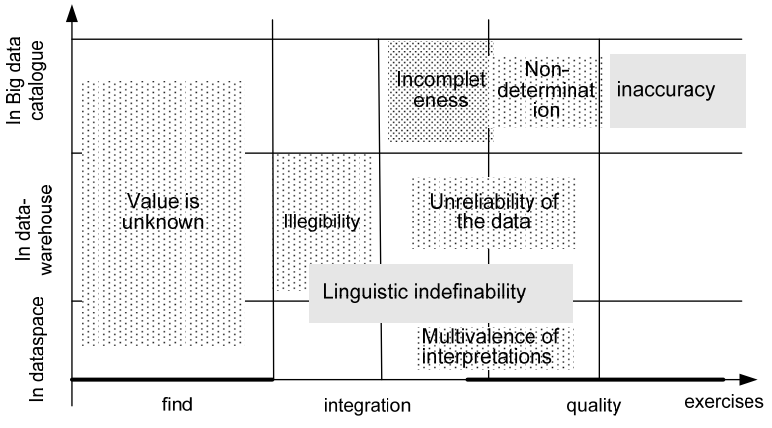

Fig. 1. Types of indeterminacy in the consolidated data in Big data catalogue and levels iof their withdrawal

Incompleteness in the level of the data warehouse arises from attacks – block data source, hiding of information as well. Indeterminacy in the level of dictionary and catalogue of data arises primarily from software failures, and because of attacks at the data sources.

Incompleteness in the data warehouse is a source of several problems: NULL-values in data cube; sparse cube; high complexity of computing. Uncertainty in Big data catalogue causes impossibility of data integration.

Let us consider more detail types of indeterminacies and show their appearance in the data warehouse and simple data. In the [6], it is analysed that indeterminacy results from the consolidation of data into a single source (local or virtual), and, therefore, one will have to deal with structured data. As a single source we will use a relational model.

Missing of data occurs due to the lack of description of the required properties in the catalog of data and dictionary. Absence can occur either because the required characteristics are not found in the Big data catalogue information products, or they not included in the catalog or dictionary through the lack of confidence. For the removal of this type of indeterminacy, the repeated use of the agent, maybe with the diminished level of trust to data is necessary.

The inaccuracy of data occurs in the level of characteristics (attributes in the relational databases) and means that the object has value of characteristic, but this value is unknown:

$$s = \{A,\ unk\}, \tag{1}$$

where s is the object that describes the characteristics of processing of consolidated data, unk is the lack of importance, A is the subset of remaining attributes from the tuple of consolidated data.

$$unk \cup \mathrm{A} = s, unk \cap \mathrm{A} = \varnothing. \tag{2}$$

Presenting this type of indeterminacy is identical to the data warehouse. The indeterminacy in the data directory is

a source of noise in all the information obtained from the source data with an unknown attribute.

Imperfection is a condition of the object, which is a subset of missing values characteristics. If this subset is empty and we talked about the relational view of data, we get the traditional tuple. Lack of information is also a partial case of incomplete information when the number of unknown tuple attribute values equal to 1. Imperfection may appear as in the case in which data are integrated and in the data dictionary as a result of failures of intelligent agent determining the structure of the source:

$$s = \{A,\ \{unk\}\},\ \ |unk| < |A|. \tag{3}$$

This type of uncertainty is modeled as well as in the data warehouse, but, unlike data warehouse arises in the relation (catalogue of data).

Indeterminacies of types 3–8 classified as the ambiguity of data mostly occur at the facility or a subset of the values of the characteristics, which form a procession. They arise as a result of attacks on the data sources (information products).

Lack of precision occurs due to incomplete studying or ambiguous displaying of characteristics values. It can be formed using the additional attribute (attributes) in relation scheme. The values contain the level of confidence in the validity of a subset of the values of non-key attributes.

$$s = \{A,\ unk_1,\ unk_2,\ldots,\ unk_n\},\quad A \in K, A',\ 1 \le n \le |A'|,$$

$$unk_{attr} = P^{attr}(i,j),\ \ |A| \ge \{unk_1, unk_2, ..., unk_n\}, \tag{4}$$

where K is the set of importance keys, A' is the subset of the values of non-key attributes.

The level of confidence can be marked using a numerical scale, linguistic assessments, fuzzy values.

The inexactitude is a result of mathematical operations and interval values processing. This type of uncertaity is modeled by an additional attribute and can occur due to the lack of precision in data dictionary.

Unlike data warehouse, this type of uncertainty occurs in Big data catalogue quite often in connection with the processing of data stored on different platforms used to solve different classes of problems.

$$s = \{A, \{unk\}\},\ \ \{unk\} \subset A,\ \ Design(A) \in \{unk\}. \tag{5}$$

Non-determination of conclusion procedures occurs when we should save intermediate or final results of the decision support procedure. Also, non-determination occurs in the facts table and in the aggregated attributes. It is modeled by extending of the data scheme and occured exclusively in the consolidated database:

$$s = s \cup \{unk\},\ \ \{unk\} \notin A,\ \ Design(s) \in \{unk\}. \tag{6}$$

Unreliability is a type of indeterminacy, which is considered one of the characteristics of the object. Although the nature of this feature is uncertain, we use traditional numerical values as domen of this attribute. Unreliability can be applied to traditional values of mathematical operations. It arises as a result of the trust definition in the data source. Unreliability is modeled by additional attribute to the data directory scheme. The value of this attribute is changed as a result of the Big data catalogue. It appears as a characteristic of the inverse value of trust in the data source.

$$s = s \cup \left[unk_j\right],\ \ unk_j \notin A,\ \ unk_j = \frac{1}{P(j)}. \tag{7}$$

The multivalence interpretations are a source of irreconcilability. This type of indeterminacy arises most often in the data directory by obtaining information from various sources and the inability to determine the validity of the data. For displaying this type of indeterminacy, we add additional attribute to relation scheme. It contains a degree of confidence in the validity of the data procession. The multivalence interpretations occur only in relation.

Linguistic indeterminacy is connected with the use of a natural language in information resources (in text files and web resources), which have a qualitative character. It can be owing to misunderstanding (lack of knowledge) of a word meaning or misunderstanding of the sense of the offer. Such type of indeterminacy is met in systems of formulating of textual information (the machine translation system, system for self-training, etc.). In the context of Big data catalogues linguistic indeterminacy arises owing to processing semi-structured information (texts, web pages, etc.).

Types of indeterminacies can be imposed or be considered by a source of appearance of each other. For a task of diminution of indeterminacy, the method which is used for indeterminacy reduction in storages of data of regular type – indeterminacy elimination on the basis of a method of extracting of knowledge is improved.

Unknown value of the attribute is considered as a class mark, and the problem of elimination of indeterminacy is transformed into a problem of reference to a class. Use of this method allows eliminating the indeterminacy like "unknown" and "imperfection" at the level of value of the attribute and a subset of attributes. However, unlike Big data, it is necessary to consider still the trust level to the data source, that is work with indeterminacy at the level of the relation.

One of the methods of modeling of inexact, lack of precision and partial data is the insertion of the additional attribute in the catalog sources which value specifies the trust degree to indeterminate data.

In [8, 9], the method of decision tree was used for uncertainty reduction in Big data. However, this method works well only with structured sources.

In [10], Fuzzy Self-Organizing Map and algorithm using fuzzy c-mean (FCM) were used to model uncertainties based on a centralized-batch processing framework. They integrated a fuzzy self-organizing map algorithm with MapReduce framework in order to execute a parallel computing on Big data. However, we can use this method only in data processing, but not in data preprocessing. Particularly, we can't find the importance of the data source in case of duplicated data.

In [11], the types of Big data uncertainty are described. However, the author analyzed only Unscalable computation ability, Ubiquitous uncertainty and weak relations. That is why all types of uncertainty should be processed in Big data catalogue.

In [12], one aspect of uncertainty is addressed by developing a new methodology to establish the reliability of user-generated data based upon causal links with recurring patterns. The authors associate a large data set of geo-tagged

Twitter messages in San Francisco with points of interest, such as bars, restaurants, or museums, within the city. This model is validated by causal relationships between a point of interest and the number of messages in its vicinity. But we can't use this model for multiple data sources analysis.

3. The aim and objectives of the study

The aim of the study was to create the method for each type of uncertainty reduction for increasing the quality of Big data analysis. Also, the definition of information product (InP) quality was given. The model of consolidated data creation allows us to find the probability of exact data source. This allows evaluating the usage of the information product for the Big data analysis process.

To achieve this aim, the following objectives had to be solved:

1. Development of a new model of consolidated data for Big data catalogue creation.
2. Improvement of the method of reducing the indeterminacy of consolidated data.
3. Development of the method for determining the viability of an InP based on the method of indeterminacy reduction.

4. Development of the model of consolidated data

The model of consolidated data is a final set of attributes $\{A_1, A_2,..., A_n\}$, set of attributes $\{A\_unk_1, A\_unk_2, A\_unk_p\}$ with indistinct or non-determinate definitions and set of attributes $\{Unk_1, Unk_2,..., Unk_m\}$, which domains are the numerical data, probabilistic data, value of function of accessory of indistinct sets, degree of the validity of multiple-valued logic, percentage, coefficients, various scales or linguistic estimates. Also, the scheme of consolidated data consists of the scheme of the synonyms dictionary *Dic* and model of the Big data catalog *Cg* [13]:

$$Cg' =< \{C1, C2,..., Cn\}, \{C\_unk1, C\_unk2, C\_unkp\}, \{Unk1, Unk2, \ldots, Unkm\}, Dic, Cg >. \quad (8)$$

The tuple of the consolidated data *dc* is the information description of the object *t* of the data source *S* presented in the form of a set (procession), importance of characteristics (attributes). The subset of attributes contains data on the object, data source and synonymic names of the object, and these data can be incomplete, indistinct or non-deterministic. The object, presented in this tuple, exists, but the part of the information on it is absent, imperfect, fuzzy, non-deterministic, etc.

The values of the consolidated data attributes are divided into groups.

1. Exact (known) – the importance of the primary key, external key (may be absent). Mark them through C.
2. Absence – no information physically. We use $\perp$ for this group.
3. Indeterminacy – set of attributes Unk used for subsets of attributes; Unk indicates a truth degree of these attributes. The default value of the attribute Unk is assigned the value, which means the highest degree of truth.

Let's notice that, in case of absolute trust to each value of a tuple, we receive a traditional relational tuple and we apply traditional operations over it.

The procession of the consolidated data *dc* is a set of values object substance:

$$dc =< C, C\_unk, Unk, \{dic\}, \{cg\} >, \quad (9)$$

where *C* is the subset of attribute values with distinct values, *C_unk* is the subset of attribute values with fuzzy and non-deterministic values, *Unk* is the subset of attribute values with truth degrees of attributes *C_unk*, {*dic*} is the set of values of the data dictionary, {*cg*} is the set of values from the directory data.

Datawarehouse of consolidated data is the set of relationships with the scheme Cg' and tuples set of consolidated data *dc*.

The model of consolidated data contains data from all types of sources of Big data catalogue.

5. Development of operations on the model of consolidated data

Because the data warehouse of the consolidated data is expansion-of the data warehouse constructed on the of relational model, we will improve operations.

For processing and analysis of indeterminacies using in query the relational operators, we should use the selection operator by the values of a set attributes *Unk*. In the data warehouse, there is a similar cut operation. Let *r* and *s* be related to the scheme *R*, r' and s' be related to the scheme $R \cup Unk \cup Dic \cup Cg$. Then $r \cap s$, $r \cup s$ and $r - s$ is the relation with scheme *R*, $r' \cap s'$, $r' \cup s'$ and $r' - s'$ is the relation with scheme $R \cup Unk \cup Dic \cup Cg$.

Considering the probability of attacks (indeterminacy like "multivalence"), we choose those data sources, the level of faith of which is higher than similar:

$$r' = r \cup \sigma_{\max(P(\pi(Cg)))}(Dic) \cup Cg. \quad (10)$$

Expansion to the relation works correctly in case of assignment of the *Unk* attribute of the lowest degree of trust to all values (a priori it is considered that this information which is brought in the relation is truthful and full, and nothing is known about the rest information). Selection of such method of representing the degree of validity is by default carried out, proceeding from the principle of isolation.

The operator of cut involves analysis of illegible value set for attribute values *Unk*.

$$slice: \sigma^{cons}_{\left(Unk \Theta unk) \cup (C\_unk \Theta c\_unk) \cup \atop \cup \sigma_C(Dic) \cup \sigma_C(Cg)\right)}(cg') = \left\{ t \in dc \middle| t(Unk) \Theta unk, t(C\_unk) \Theta c\_unk, meta_{Unk, C\_unk} = 1, \atop \sigma_C(Dic)\ Is\ Not\ NULL, \sigma_C(Cg)\ Is\ Not\ NULL, unk = P(cg') \right\}, \quad (11)$$

where Θ is the set of binary relations symbols (marks) on pairs of values domains. For each attribute C_unk we used comparison operations. As a rule, we use only $=, \neq, <, \leq, \geq, >$.

Advanced slice operator is distributive relatively to binary Boolean operations:

$$\sigma^{cons}{}_{A=a}(r' \gamma s') = \sigma^{cons}{}_{A=a}(r') \gamma \sigma_{A=a}(s'), \quad (12)$$

where $\gamma = \cap, \cup$ or $-$, r' i s' is the relation over the same scheme.

The data warehouse drill-down operation is analogue to projection operation in the relational model. For the projection realization in consolidated data we should find connection between subset of attributes Unk and subset of attributes *C_unk* and check synonyms in the dictionary Dic for the attribute name *C_unk*. Therefore, the improved drill-down operation is presented as follows:

$$\text{drill-down}:\pi_X^{cons}(cg') = \\ = \text{IIF}\begin{pmatrix} \neg\ \text{ISNULL}\left(\sigma_{Cg=R\cup C\_unk=X}(c\_unk)\right); \\ \pi_{X\cup\pi_{Unk}(\sigma_{Cg=\ meta(C\_unk,Unk)=1}(c\_unk))}(dc); \\ \text{IIF}\left(\sigma_{C\cup C\_Unk=X}(Dic);\pi_{\sigma_{C\cup C\_Unk=X}(Dic)}(r);\pi_X(dc)\right) \end{pmatrix}, \quad (13)$$

where IIF (condition; operation1, operation2) is the operation introduced in the standard SQL 92. If the condition is performed condition 1, otherwise condition 2; ISNULL(r) – logical operator that results in true if the relation *r* operand does not contain tuples and defect – in that case. Also, we need the search of synonym attribute in the dictionary of synonyms *Dic* ($\sigma_{C\cup C\_Unk=X}(Dic)$) and replacement ($\pi_{\sigma_{C\cup C\_Unk=X}(Dic)}(r)$).

The connection operator is used to link related facts and relation of measurements in consolidated data, since it is based on the relational model.

Traditional connection operator can not be used for Big data catalogue and data warehouse with consolidated data, because for statistical analysis it is necessary to connect related facts relational dimensions. If subsets of attributes *Unk* is non-empty for the facts and dimensions, such connection is incorrect. Also, operator connections are affected by the fact that there is a need not only to connect with those attributes specified as input parameters, but also to check for synonyms in a dictionary of synonyms *Dic*. For improving service connection, one should consider cases where the relationship is completely connecting or not connecting fully. For full connecting relations of input attributes set *Unk* does not affect the operation of the connection. If the set of attributes *Unk* contains indeterminacy as a foreign key relationship, which is a connection, then this measure of indeterminacy is transferred to all the rest of the attribute values of this ratio. In the case of incomplete connections of attribute Unk with tuples from subordinate tables that do not occur in the relation, the value will be equal to the highest degree of confidence.

$$across: r \underset{cons}{\triangleright\!\triangleleft} cg' = \\ = \text{IIF}\begin{pmatrix} \sigma_{C\cup C\_Unk=X}(Dic);\pi_{\sigma_{C\cup C\_Unk=X}(Dic)}(r \underset{\bullet}{\triangleright\!\triangleleft} cg'); \\ \pi_{(R,B,NVL(Unk,\min))}(r \underset{\bullet}{\triangleright\!\triangleleft} cg') \end{pmatrix}, \quad (14)$$

where *r* is the traditional relation, cg' is the relation with the consolidated data, *R* is the set of relation attributes *r*, S is the set of relation attributes cg', not including a subset of attributes Unk ($Cg' = Cg \cup Unk$), B is the set of attributes with *S*, which are not covered in relation *r* ($B \subset Cg$, $B \not\subset Cg \cap R$), min is the importance, which means the lowest level of faith, $NVL(Unk,\min)$ is the operation that assigns min for all values Unk for connecting related processions cg', $\underset{\bullet}{\triangleright\!\triangleleft}$ is the left connection. It is necessary to check connections of synonyms ($\sigma_{C\cup C\_Unk=X}(Dic)$). If not, the operations of the left connection for relations with schemes S 'and R and the projection of the attributes-synonymous are processed.

Otherwise, the operation of the left connection by the common attributes is realized, and then over the relation received from the previous operation of projection. The result of this operation is connection with the empty value of a subset of the Unk attributes and min value is saved in Unk.

It should be noted that when the dictionary of synonyms is empty ($Dic = \varnothing$) and the probability of appeal to data sources as a whole and their characteristics is equal to 1 ($Unk = 1$), we will receive a traditional relational connection.

6. Reduction of indeterminacy of consolidated data

The analysis of large amounts of data requires identification of groups of attributes that form the functional dependence. However, in the real world data sets are much more common in which important dependencies are defined only on a subset of the values of key attributes, call the following dependencies partial functional dependencies. That is, a partial functional dependency is an FD defined in some fixed ratio selection.

$$F_p:\ K = \{a_i\}, a_i \in A, D = \{a_j\}, a_j \in A, R' \subset R: K \to D\big|R'. \quad (15)$$

Many relations are not clearly determined, call them probabilistic dependencies of production.

Probabilistic productive relationship is the production rule in the selection of the basic relation that holds a significant number of objects for this selection. The threshold of significance should be determined by experts; or based on calculations of the probability of false selection of this relationship.

$$F_I:\ K = \{a_i\}, a_i \in A, D = \{a_j\}, a_j \in A,: P(k \in K \to d \in D) = p, \quad (16)$$

here *k* and *d* are the tuples of values of certain groups of attributes K and D, respectively.

The main indicator of the reliability of such dependence is the ratio of objects number with the probabilistic productive relationship to objects number in the selection:

$$P(F_I) = \frac{\left|\sigma_{k\in K\wedge d\in D}(R)\right|}{\left|\sigma_{k\in K}(R)\right|}. \quad (17)$$

Classification rule is called probabilistic productive relationship between subsets of attributes X and Y in the data warehouse with consolidated data cg', which occurs in the test set cg' with a degree of conformity (faith) *s*, where $(X = x) \to (Y = y)$.

The classification rule is constructed based on training data set cg', where the tag class value (value of attributes subset *Y*) is known. The classification rule generally built for the scheme cg', and therefore will not be affected by the new tuples arriving in the relation of the consolidated data repository (independence of the test set).

Mark of class is linguistic variable or traditional object characteristic that is the value of a subset of attributes *Y* and marks objects with similar (similar with degree s) values of a subset of attributes *X*. Domains attributes that belong to a subset of Y, $y \in dom(Y) = \pi_Y(Cg')$, must contain a finite and pre-known set of values.

Marks of a class are selected from a predefined set of values (they are known in test dataset), and reference to a

class of objects information about which just arrived in the data warehouse with the consolidated data, is carried out on the basis of classification rules. The marks will be added automatically, since the new data flow into the data space is also dynamic.

Calculation of the reliability performance of such a relationship is based on the possibility of such a schedule depending on the components of the probabilistic productive relationship:

$$P(s \in S \to t \in T) = \\ = \sum_{t_i \in T} P(s \in S \to t = t_i) = \sum_{t_i \in T} \frac{\sum_j \left| s = s_j \wedge t = t_i \right|}{\sum_j \left| s = s_j \right|}. \quad (18)$$

As in the case with *F*-dependencies (functional dependencies), a set of classification rules, which take place in a given relation can be represented by some subset of them, which by inference rules can get all the classification rules of the relationship. Since the classification rules are an extension with F-dependencies, you should consider transforming of functional dependencies axioms for classification rules.

Reflexive property. $P(s \in S \to s \in S) = 1$ for any relation $r(R)$.

Proof:

$$P(s \in S \to s \in S) = \frac{\left| \sigma_{s \in S \wedge s \in S} \right|}{\left| \sigma_{s \in S} \right|} = \frac{\left| \sigma_{s \in S} \right|}{\left| \sigma_{s \in S} \right|} = 1. \quad (19)$$

Replenishment: If

$$P(s \in S \to t \in T) = p, \ \ P(s \in S \wedge w \in D(W) \to t \in T) = p.$$

Proof:

$$P(s \in S \wedge w \in D(W) \to t \in T) = \frac{\left| \sigma_{s \in S \wedge w \in D(W) \wedge t \in T}(R) \right|}{\left| \sigma_{s \in S \wedge w \in D(W)}(R) \right|} = \\ = \left| \forall x \in r : q = \pi_{W=w}(x) \in D(W) \Rightarrow w \in D(W) \right| = \\ = \frac{\left| \sigma_{s \in S \wedge t \in T}(R) \right|}{\left| \sigma_{s \in S}(R) \right|} = P(s \in S \to t \in T) = p. \quad (20)$$

Additivity: If

$$P(s \in S \to t \in T) = p$$

and

$$P(s \in S \to w \in W) = 1,$$

then

$$P(s \in S \to t \in T \wedge w \in W) = p.$$

Proof:

$$P(s \in S \to t \in T \wedge w \in W) = \frac{\left| \sigma_{s \in S \wedge t \in T \wedge w \in W} \right|}{\left| \sigma_{s \in S} \right|} = \\ = \left| s \in s \to w \in W \right| = \frac{\left| \sigma_{s \in S \wedge t \in T} \right|}{\left| \sigma_{s \in S} \right|} = P(s \in S \to t \in T) = p. \quad (21)$$

Eliminating the uncertainties that occur among the values of the attribute Y in the relation r, is classification using a modified *chase* algorithm.

The point of the method:

1) search for tuples with the same values in the set of attributes X;

2) search for tuples with the same values in the set of synonyms attributes X;

3) calculation of the level of confidence in the source of tuple obtained in steps 1) and 2);

4) calculation of confidence to attribute sources of tuple obtained in steps 1) and 2);

5) determining the tuples with the highest level of confidence.

If we are able to classify the objects, it's necessary to build classification functions. Generally, in the space of data information about several types of classes can be stored, and each class type has its own subset of features. One and the same function can be used to specify multiple types of classes.

Classification functions are called the modified functional relationships that are performed for a specific subset of tuples in consolidated data repository.

The classification algorithm:

1) If $\sigma(cg') = \{\text{dc}_1(X_1) \downarrow, ..., dc_1(X_n) \downarrow\}$ **i** $\{dc_2(X_1) \downarrow, ..., dc_2(X_n) \downarrow,\}$

And $\{dc_1(X_1) \downarrow, ..., dc_1(X_n) \downarrow = dc_2(X_1) \downarrow, ..., dc_2(X_n) \downarrow\}$

And $\{dc_1(Y) \downarrow\}$ i $\{dc_2(Y) = \perp\}$ and If $\sigma_{X_1}(Dic) = \varnothing$

Then replace $\perp$ by $\text{dc}_1(Y)$ **i**

$$dc_1(P) = dc_1(P) / (\sum_i m_{1i} / n).$$

2) If $\{\text{dc}_1(X_1) \downarrow, ..., \text{dc}_1(X_n) \downarrow\}$

And {in dc_2 *m* with *n* importance of attributes – $\downarrow$, *n* – *m* importance of attributes – $\perp$, $m \le n$}

And $\{P \ge 1 - m/n\}$ and {on certain importances $\text{dc}_1(X^m) \downarrow = \text{dc}_2(X^m) \downarrow$}

And $\{\text{dc}_1(Y) \downarrow\}$ and $_2(Y) = \perp\}$,

Then change $\perp$ in *r* $\text{dc}_1(Y)$ **i**

$$dc_2(P) = dc_2(P) / \left(\sum_i m_{2i} / n \right).$$

3) If {in dc_i m_i with *n* importance of attributes – $\downarrow$, $m_i \le n$}

And {in dc_j m_j with *n* importance of attributes – $\downarrow$, $m_j \le n$}

And {on certain importances $\text{dc}_i(X^m) \downarrow = \text{dc}_2(X^m) \downarrow$}

And {on certain importances $\text{dc}_i(X^m) \downarrow = \text{dc}_2(X^m) \downarrow$}

And $\left\{ m_i / n \le m_j / n \right\}$ and $\left\{ P \ge 1 - m_i / n \right\}$

And $\{\text{dc}_i(Y) \downarrow\}$ and $\{\text{dc}_j(Y) \downarrow\}$ and $\{\text{dc}_2(Y) = \perp\}$,

Then change $\perp$ in $\text{dc}_j(Y)$ and

$$dc_2(P) = dc_2(P) / \left(\sum_i m_{2i} / n \right).$$

The method for determining the viability of an InP

Viability is the measure with which an InP is used in a specific subject area to achieve a specific goal with appropriate efficiency, productivity, and satisfaction of needs at intervals of terminal time.

InP is a function of the time of creation, the metadata (number of InPs, technical solutions, etc.).

The method for determining the viability of the InP consists of the following steps:

1. Calculation of InP characteristics.

2. Expert definition of the weight characteristics.

3. Calculation of viability [14].

Step 1. A set of significant values for the site as an information product (dimensionless), obtained on the basis of theoretical and experimental studies, is given as:

$$Y=\{V,K,A,Km,O,Ac,N,Ms,C,Pr\},\ \ Y\to[0...1]. \qquad (22)$$

Importance of information V (y_1) is the parameter that has a dynamic character and exists only at the moment of interaction of data and methods in the information process for a particular social group (ς_i – type i social group): $V_{\varsigma_i}\to[0...1]$:

$$V=\sum_{i=1}^{n}V_{\varsigma_i},\ \ V_{\varsigma_i}=\frac{N_{\varsigma_i}(t)}{\sum N_{\varsigma_i}},\ \ t_1<t_i<t_2, \qquad (23)$$

where V_{ς_i} is the importance of information for the type of social group ς_i; t is the time of its using; $N_{\varsigma_i}(t)$ is the number of information messages for a social group over a period of time t; $\sum N_{\varsigma_i}$ is the total number of information messages in this social group.

Usefulness of information messages K (y_2) is the parameter that characterizes compliance with the needs of the user, that is, assessing the relevance of information messages in *Ip.IR*:

$$K(Ip_j)=(\sum_k P(Ip_j.IR\,|\,t_i)V(D_i\,|\,Ip_j.IR), \qquad (24)$$

where $P(Ip_j.IR)$ is the probability of receiving information messages from $Ip_j.IR$ in the moment of time t_i, $t_1<t_i<t_2$, D_i is the current value decision *Ip.IR*, *Ip.IR* is the information resource IP.

Adaptability A (y_3) (compliance with user requirements) is formed on the basis of an assessment of the ratio of information and intellectual resources in relation to K which makes it possible to determine the number of components (modules) in the InP, that is:

$$A=\frac{K(|\sigma_{t_i}(IR)|+|\sigma_{t_i}(HR)|)}{|IR|+|HR|},\ \ t_1<t_i<t_2, \qquad (25)$$

where $|\sigma_{t_i}(IR)|$ is the amount of information resource per time t_i, $|\sigma_{t_i}(HR)|$ is the amount of intellectual resource per time t_i, $|IR|$, $|HR|$ is the total amount of information and intellectual resources, respectively.

Convenience of communication with users Km (y_4) is the parameter describing the appearance or ease of use according to the expert's assessment, in which the InP is available for the maximum number of users. Their weight is estimated by the hierarchy analysis method (the parameter is determined according to the evaluation of the expert – Q):

$$Km=\frac{\sum_i Km_i\times Q}{\sum Km_i},$$

$$Km_i=1,\ \ Q=\overline{0...1}. \qquad (26)$$

InP service O (y_5) is the depth of linking (the number of transitions from the main link to the required one), etc. This parameter depends on the degree to which the IP meets modern requirements:

$O=\frac{1}{m}$, m is the depth of linking.

InP availability Ac (y_6) determines how freely users can use the InP (which was evaluated by the expert, the values are given in Table 1).

Table 1

Scale of ranking of "Availability"

Interpretation of influence	Value range
Complex access	0.0–0.2
Average access	0.3–0.6
Easy access	0.7–1.0

Prevalence of IP N (y_7) is the parameter that determines the number of IPs of this type:

$$N=\frac{COUNT(\sigma_{name\wedge Ip\_type}(M))}{COUNT(\sigma_{Ip\_type}(M))}, \qquad (27)$$

where $\sigma_{name\wedge Ip\_type}(M)$ is the operation of sampling from metadata by the type of information product *type* and its name *name*, $COUNT$ is the quantity determination function, $\sigma_{Ip\_type}(M)$ is the sampling operator of metadata by type.

Attendance IP Ms (y_8) is the parameter that determines the number of users:

$$Ms=\frac{COUNT(\sigma_{user}(Ip))}{COUNT(\sigma(Ip))}, \qquad (28)$$

where $COUNT(\sigma_{user}(Ip))$ is the number of information product users *Ip*.

Social affiliation C (y_9) is the parameter that defines the circle of users of the given IP (fuzzy ratio of PI positioning):

$$C=\max\left(\frac{COUNT(\sigma_{user=\varsigma_i}(Ip))}{COUNT(\sigma_{user}(Ip))}\right),\ \ i=1...n, \qquad (29)$$

where ς_i is the type i social group.

Value (Pr) (y_{10}) is the cost of operating the InP. The ranking scale of the "Cost" characteristic is shown in Table 2.

Table 2

Scale of ranking of the "Cost" characteristic

Interpretation of influence	Value range
Lack of value/low price	0.7–1.0
Average cost	0.4–0.6
High price	0.0–0.3

Step 2. The weight of InP characteristics is determined on the basis of expert evaluation. To do this, we used $\boldsymbol{X}=\{x_1,x_2,...,x_n\}$ – a set of experts, $\boldsymbol{Y}=\{y_1,y_2,...,y_p\}$ – a set of characteristics and $\boldsymbol{Ip}=\{Ip_1,Ip_2,...,Ip_m\}$ – a set of information products. The function of fuzzy binary relation is defined $F_R:X\times Y\to[0,1]$.

Then for all $x\in X$, $y\in Y$, the function $F_R(x,y)$ is the degree of importance of the characteristic y according to the evaluation of the individual x when determining the advantage of a particular InP. The function of fuzzy binary relation H is defined as $d:Y\times\boldsymbol{Ip}\to[0,1]$. For all $y\in\boldsymbol{Y}$, $Ip\in\boldsymbol{Ip}$, $d_H(y,Ip)$ is equal to the degree of influence of the characteristic y on the information product *Ip*, then we form the matrix of characteristics:

$$H = \begin{array}{c} \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \begin{array}{c} \begin{array}{cccc} Ip_1 & Ip_2 & \dots & Ip_m \end{array} \\ \begin{bmatrix} d_H(y_1,Ip_1) & d_H(y_1,Ip_2) & \dots & d_H(y_1,Ip_m) \\ d_H(y_2,Ip_1) & d_H(y_2,Ip_2) & \cdots & d_H(y_2,Ip_m) \\ \dots & \dots & \dots & \dots \\ d_H(y_n,Ip_1) & d_H(y_n,Ip_2) & \dots & d_H(y_n,Ip_m) \end{bmatrix} \end{array}, \quad (30)$$

elements of which are determined by the function of belonging to a certain sphere of using:

$$\mu A_i(y,Ip_i) = \frac{\sum F_R(x,y)\cdot d_H(y,Ip_i)}{\sum F\ (x,y)}$$

for all $x \in \boldsymbol{X}$, $y \in \boldsymbol{Y}$ and $Ip \in \boldsymbol{Ip}$, (31)

that is, the basis for constructing a classification rule.

Then the importance of the characteristics Y is determined by the vector R:

$$R = \begin{bmatrix} \mu A_1(y_1,Ip_1) \wedge \mu A_2(y_1,Ip_2) \dots \mu A_{m-1}(y_1,Ip_{m-1}) \wedge \mu A_m(y_1,Ip_m) \\ \mu A_1(y_2,Ip_1) \wedge \mu A_2(y_2,Ip_2) \dots \mu A_{m-1}(y_2,Ip_{m-1}) \wedge \mu A_m(y_2,Ip_m) \\ \dots \\ \mu A_1(y_{10},Ip_1) \wedge \mu A_2(y_{10},Ip_2) \dots \mu A_{m-1}(y_{10},Ip_{m-1}) \wedge \mu A_m(y_{10},Ip_m) \end{bmatrix}. \quad (32)$$

By taking into account the rank of the characteristic, the Kendall concordance coefficient was modified to match the cardinal weights of the experts:

$$w = \frac{12}{m^2(n^3-n)}S, \quad S = \sum_{j=1}^{n}\left(\sum_{i=1}^{m} R_{ij} - \frac{m(n+1)}{2}\right)^2, \quad (33)$$

where n is the number of analyzed InPs, m is the number of experts, R_{ij} is the rank of the j-th characteristic of the IP, assigned to it by the i-th expert.

Step 3. The viability of the InP is defined as an integral measure

$$G = \sum_{i=1}^{k} w_i y_i, \sum_{i=1}^{k} w_i = 1, \quad G \to [0...1]. \quad (34)$$

As a result of the analysis, we combine the decisions S of experts. $\boldsymbol{Ek}_1$ is the set of experts who recognized the information product Ip_1 the decision was made on low viability, and Ek^2 is the set of experts, who recognized the InP as necessary and decided on high viability.

Designation. If $Ek_1 \cap Ek_2 = \varnothing$, $Ek_1 \cup Ek_2 = \{1,...,N\}$, then the decided S defined as:

$$Ip = \begin{cases} unviable if \prod_{i\in I_1} G_i \prod_{i\in I_2}(1-G_i) > \lambda \prod_{i\in I_2}(1-W_i)\prod_{i\in I_1} W_i, \\ viable if \prod_{i\in I_1} G_i \prod_{i\in I_2}(1-G_i) < \lambda \prod_{i\in I_2}(1-W_i)\prod_{i\in I_1} W_i. \end{cases} \quad (35)$$

Also, in this article, the risk factors of InPs and their effect on the viability of PIs at various stages of the life cycle are determined.

Definitions. The InP risk factor is a situational characteristic of the InP, which leads to an uncertain outcome and the occurrence of adverse consequences due to distortion of information or non-relevant search results (Table 3).

Table 3

Risk factors of InP

Type of risk factor	Example of risk factor
Risk factor for creating	Incorrect relationships and relations Selection of means of implementation Determination of ownership/tenure
Risk factor of use	Unauthorized familiarization and use (in particular, copying) Unauthorized linking Unauthorized modification (modification) Deliberate destruction of information
Risk factor of spreading	Correctness of InP application Choice of format

The usage of these risk factors allows us to predict the availability of InP information for the whole Big data catalogue.

7. Discussion and future work with uncertainty reduction in Big data

So, the new model of consolidated data developing for Big data catalogue was created. This approach allows us to collect data from duplicated sources. The method of uncertainty reduction can be used for different issues in decision support systems. For example, it should be the first step in the data cleaning process and classification. The model of consolidated data can be used for system documentation and automative metadata creation. The advantage of the method for determining the viability of the information product is the possibility to find useful information products in Big data catalogue in case of duplicated data and find the value of the risk factor. In contrast to the method of Learning from Uncertainty for Big Data, which allows large-scale missing values of big data only, the proposed method works also with indeterminacy. However, it is very difficult to reduce linguistic uncertainty. The future work is finding of the correlation between the value of the risk factor and the type of uncertainty.

8. Conclusions

1. The model of consolidated data, which is an extension of the model related to the indeterminacy was given. It allowed us to process data with different types of uncertainty. The operations over the relation with indeterminacy for the purpose of their application in the data warehouse with the consolidated data that allowed realizing unary operations of Big data catalogue are improved. It allows us to preprocess all types of uncertainty in Big data and Big data catalogue.

2. The method for reducing the indeterminacy of data available in the repository of consolidated data as a basis for further evaluation of the quality of consolidated data was created. The considered method is useful also for decision making. It provides a search for hidden relationships between the characteristics of the consolidated data repository. Such dependence should be considered when making decisions based on consolidated data. The result of this work is to reduce the uncertainty for assessing the viability of the information product.

3. The method for determining the viability of the information product was created. It allows us to find useful information products in Big data catalogue in case of duplicated data and find the value of the risk factor.

References

1. Shakhovska N. B., Bolubash Y. J., Veres O. M. Big data federated repository model // The Experience of Designing and Application of CAD Systems in Microelectronics. 2015. doi: 10.1109/cadsm.2015.7230882
2. Zadeh L. The concept of a linguistic variable and its application to the adoption of approximate solutions. New York, 1976. 166 p.
3. Tselmer G. Risk consideration in management decisions // Problems of ICSTI. 1980. Issue 3. P. 94–105
4. Knight F. K. Risk, uncertainty and profit. Moscow: Business, 2003. 358 p.
5. Moiseyev N. N. Elements of the theory of optimum systems. Moscow: Science, 1975. 528 p.
6. Trukhachev R. I. Decision-making models in the conditions of uncertainty. Moscow: Science, 1981. 151 p.
7. Shakhovska N., Medykovsky M., Stakhiv P. Application of algorithms of classification for uncertainty reduction // Przeglad Elektrotechniczny. 2013. Vol. 4, Issue 89. P. 284–286.
8. Learning ELM-Tree from big data based on uncertainty reduction / Wang R., He Y.-L., Chow C.-Y., Ou F.-F., Zhang J. // Fuzzy Sets and Systems. 2015. Vol. 258. P. 79–100. doi: 10.1016/j.fss.2014.04.028
9. MRPR: A MapReduce solution for prototype reduction in big data classification / Triguero I., Peralta D., Bacardit J., García S., Herrera F. // Neurocomputing. 2015. Vol. 150. P. 331–345. doi: 10.1016/j.neucom.2014.04.078
10. Karami A. A Framework for Uncertainty-Aware Visual Analytics in Big Data // In AIC. 2015. P. 146–155.
11. Wang X., He Y. Learning from Uncertainty for Big Data: Future Analytical Challenges and Strategies // IEEE Systems, Man, and Cybernetics Magazine. 2016. Vol. 2, Issue 2. P. 26–31. doi: 10.1109/msmc.2016.2557479
12. Taming Uncertainty in Big Data / Bendler J., Wagner S., Brandt T., Neumann D. // Business & Information Systems Engineering. 2014. Vol. 6, Issue 5. P. 279–288. doi: 10.1007/s12599-014-0342-4
13. Veres O., Shakhovska N. Elements of the formal model big date // In Perspective Technologies and Methods in MEMS Design (MEMSTECH). 2015 XI International Conference. 2015. P. 81–83.
14. Vovk O. B., Shakhovska N. B. Formation of the factors influencing the behavior of the information product // Radioelectronics, informatic, management. 2015. Issue 2. P. 43–53.