Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090412

Model of Data Warehouse with Uncertain Consolidated Data

Nataliya Shakhovska and Rostyslav Strubytskyi*

Department of information systems and networks, Lviv Polytechnic National University, Bandery str, 12, 79013 Lviv, Ukraine

Received: 13 Oct. 2014, Revised: 13 Jan. 2015, Accepted: 14 Jan. 2015
Published online: 1 Jul. 2015

Abstract: Problems which arise up during work with separate sources with depositories information using and databases are analyzed. There is formalized model of consolidation datawarehouse and there is built the methods for uncertainty elimination. Data processing problem under conditions of uncertainty is analyzed. The methods of representation in respect of consolidated data repository, elimination of unknown and incomplete data and reduce inaccuracies, particularly, multiplicity of interpretations are developed.

Keywords: datawarehouse, heterogeneous data, uncertainty

1 Introduction

To implement data warehouses and dataspaces using database management systems, means of data exchange and integration are necessary. Data sources such as spreadsheets, multimedia information, etc, that are used in energy systems, can have their own means of storing and processing, and then the task of integration is the recognition of these information resources and access to them. When talking about data storage, the structure of sources is known in advance, and the main challenge is clearing and loading data itself. For dataspaces it is necessary to provide the opportunity to work with the software, which theoretically might not be at the user's workstation. If such a possibility is not foreseen, it is necessary to predict the development of such data storage structure so that it can retrieve data from data sources to provide answers to user queries.

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

Indeterminacy can arise at the level attribute tuple and relation (indeterminacy in the circuit description). The appearance of indeterminacy at the attribute and tuple due to multidimensional display leads to the spread of uncertainty in all copies of a particular concept. Since the data space of millions of data items subject area, the traditional means of handling indeterminacy(interval maths, multivalent logic) becomes ineffective because of the large number of operands.

Thus, the specificity of data space (the presence of diverse set of sources, data doubling, ambiguity describing data sources) leads to the fact that the indeterminacy in traditional relational databases 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 data space. Therefore, for processing indeterminacy in the data space must use a different approach, the need for the use of which has not had in relational databases and data warehouses.

* Corresponding author e-mail: natalya233@gmail.com

© 2015 NSP
Natural Sciences Publishing Cor.

2 The causes and sources of uncertainty in data warehouses

Since the data fall into local datawarehouse from multiple sources, there is the problem of duplication of data uncertainty and imprecision.

Let us analyze the causes of uncertainty in the storage and data space.

1. Uncertainty in circuit mediator.

Mediator is software component interacts with user of integrating systems and information sources. It provides a single ”point of entry” (API) to user requests, performs basic stages of processing the request:

–identify sources that may contain the query result;
–decomposition of requests to specific sources (based on their descriptions);
–optimization plan execution;
–and so on.

Scheme mediator is a set of schemes of terms found in queries. It does not necessarily cover all the attributes of any of the sources, but includes information about the domain data source. Uncertainty in circuit mediator may be several reasons. Firstly, if the agent schema automatically determined from the data source at run time, there is uncertainty about the results of a query. Secondly, when the domain is broad, there is uncertainty regarding compliance schemes or their data overlap.

This uncertainty leads to inaccurate mapping schema source and a source for other uncertainties. The reasons for this uncertainty is external (attack), software, hardware disturbance in the process of selecting and loading data.

2. Uncertainty in circuit mapping.

Usually this type of uncertainty appears in a dictionary of synonyms *Dic*. This is the particular case of uncertainty in the mediator scheme. As thesaurus defines semantic relations between terms in the sources of data that are completely independent, and many primary reflections schemes will be automatically received is, the display may be inaccurate.

An example of such uncertainty may be a case where one term identifying different objects (polysemy).

3. Uncertainty of data consolidated data repository (data warehouse).

This type of uncertainty appears as result of automatical data integration from different sources. In addition, systems that incorporate many sources contain false or contradictory information. Uncertainty can arise even when the raw data were accurate, since the reflection properties of one can be used a variety of domains.

An example of the domain, which clearly demonstrates this type of uncertainty is a system of authentication events. In this case, an important role is played by the degree of confidence in the data source.

4. Uncertainty requests.

Uncertainty queries arising from the presence of different data models and their expressive power. System itself transforms the request received from the user, for example, based on keywords. When converting this type of query in SQL-query to a structured sources of uncertainty may be the results of a query.

Uncertainty requests clearly demonstrate retrieval system, where the user requests given too many search results, and only some of them actually satisfy the user.

3 Analysis of the literature source and formulation of the problem

Classify types of indeterminacy by the nature of their manifestation in the data space. One of the first works in this direction is the work of L. Zade [1], G. Tselmer [2] emphasize that indeterminacy, as the objective form of life surrounding of the real world, is conditioned, on the one hand, 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 states in which an object changes in dynamics, may be in future time.

F. Knight under conditions indeterminacy understands the insufficient of learning and the need to act upon opinion rather than knowledge [3], VV Cherkasov treats that indeterminacy as a constant changeability of conditions, fast and flexible re-orientation of production, the actions of competitors and market trend analysis, etc. He calls indeterminacy the most characteristic reason of risk is administrative activity. He distinguishes two classes of indeterminacy: 1.¡¡good¿¿ indeterminacy when for unknown factors are statistic or probabilistic characteristics, and ¡¡bad¿¿ indeterminacy when such characteristics, in principle, can not be obtained, and there are methods of definition of both kinds of indeterminacy, which arising in real tasks (Wentzel, 1980).

In (Moiseev, 1975) contains the following classification indeterminacies [4]:

–the degree of indeterminacy: probabilistic, linguistic, interval, full of indeterminacy;
–the nature of indeterminacy: parametric, structural, situational;
–obtained by using in the course of managing information: surmountable and incorrigible.

In Diev V.S. and Truhacheva R.I. [5] suggest more detailed classification of indeterminacies in the current business system.

In [6] there are defined such types of indeterminacies, the nature of which is:

–Value is unknown (missing).
–Incompleteness of the information

© 2015 NSP
Natural Sciences Publishing Cor.

–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 undefinability.

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

Uncertainty of types 3-8 categorize 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 most often meets 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 sources at which one 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 arises in connection with inexact analysis or ambiguous mapping of objects in relation. In relation is figured with the help of padding attribute, the characterizes values which measure of confidence to a tuple or subset of values of attributes in a tuple.

The multivalence of interpretation is by one of sources of originating of inconsistencies.

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

Such type of uncertainty meets 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 occurrence one another.

Nowadays time the methods of elimination are missing, inexact and indistinct data [1,2,3] are designed. Therefore it is necessary to elaborate methods, which can work with all types of uncertainty.

Uncertainty of these types may be in database, datawarehouse and dataspace (Fig. 1).

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

Consider more specified types of indeterminacies and shows for their appearance in the datawarehouse and simple data [6]. Analysis that indeterminacy is resulting from the consolidation of data into a single source (local

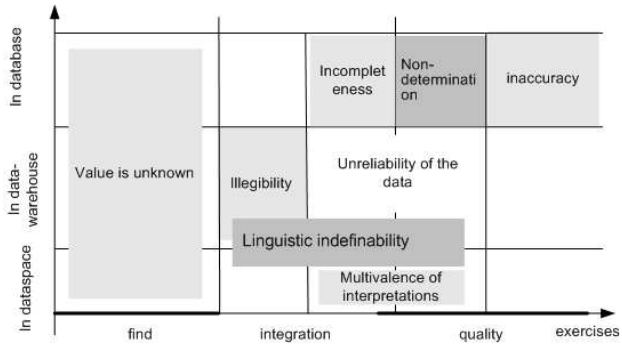

Fig. 1: Types of indeterminacy in the consolidated data in dataspace and levels at their withdrawal

or virtual), and, therefore, will have to deal with structured data. For present of a single source will use a relational model.

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

The inaccuracy of data occurs in the level of characteristics (attributes in the relational databases) and means that the importance inherent in the object, but unknown.

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

where s — object that describes the characteristics of a procession of consolidated data, unk — lacking of importance, A — remaining attribute importance characteristics procession of consolidated data.

$$unk \cup A = s,\ unk \cap A = \varnothing.$$

Presenting this type of indeterminacy is identical to the datawarehouse. In case of indeterminacy in the level of directory data leads to noise in all the information obtained from the source data with 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 procession. 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 of dictionary as a result of failures of intelligent agent determine the structure of the source:

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

© 2015 NSP
Natural Sciences Publishing Cor.

Modeled as well as in the datawarehouse, but, unlike datawarehouse arises at the level of the relation (catalogue of data).

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

Lack of precision occurs due to incomplete or ambiguous studying display characteristics essence. Simulated by addition schemes related additional attribute (attributes) whose values contain the level of confidence in the validity of a subset of the values of non-key attributes.

$$s = \{A, unk_1, unk_2, \dots, unk_n\}, A \in K, A', 1 \le n \le |A'|,$$
$$unk_{attr} = P^{attr}(i, j), |A| \ge \{unk_1, unk_2, \dots, unk_n\}$$

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

The level of confidence can be marked using a numerical scale linguistic assessments, illegible magnitude.

The inexactitude getting from the application of mathematical operations on numeric data (this type is also indefinite resulting from the work of interval values). This type of indefinite is modeled by an additional attribute and can occur due to lack of precision in data dictionary .

Unlike datawarehouse, in data space occurs quite often in connection with the processing of data stored on different platforms used to solve different classes of problems.

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

Non–determination procedures output decisions (chance) occurs when the need to store intermediate or final results of the procedure for withdrawal or decision, and with regard to the facts at the level of values aggregated attributes. Modeled by expanding circuit and data storage occurs exclusively in the consolidated database:

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

Unreliability is a type of indeterminacy, which is considered one of the characteristics of the object. Although the nature of this feature is uncertain with regard to both its domain using traditional numerical scales and applied to its traditional values of mathematical operations. Arises from the definition of trust reference to data source. Modeled by addition schemes additional attribute data directory. The value of this attribute is changed as a result of the data space. It appears as a characteristic of the inverse value of trust in the data source.

$$s = s \cup [unk_j], unk_j \notin A, unk_j = \frac{1}{P(j)}$$

The multivalence interpretations is 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 relation scheme complement additional attribute that contains a degree of confidence in the validity of the data procession. The type of ambiguity wherein the injected level attitude.

Linguistic indeterminacy is connected with use of a natural language in information resources (in text files and web resources) which have qualitative character, and there can be owing to misunderstanding (lack of knowledge) a word meaning or misunderstanding of sense of the offer. Such type of indeterminacy meets in systems of formulating of textual information (the machine translation system, system for self-training, etc.). In a context of data spaces 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 attribute is considered as a class mark, and the problem of elimination of indeterminacy is transformed to a problem of reference to a class. Use of this method allows to eliminate indeterminacy like ¡¡unknown¿¿ and ¡¡imperfection¿¿ at the level of value of attribute and a subset of attributes. However, unlike datawarehouse, it is necessary to consider still trust level to data source, that is work with indeterminacy at the level of the relation.

The problem of elimination (reduction) of indeterminacy is a construction homomorphic display of a set of the consolidated data which were stored in storage, in a set of the data used for maintenance of decision-making (fig. 2) for the purpose of improvement of quality of consolidated data and acceptance on their basis of effective leading strategic decisions, considering probability of emergence of attacks. Attack — addition to data space of the source, which structure of data causes a polysemy of interpretations in the dictionary of synonyms Dic.

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

© 2015 NSP
Natural Sciences Publishing Cor.

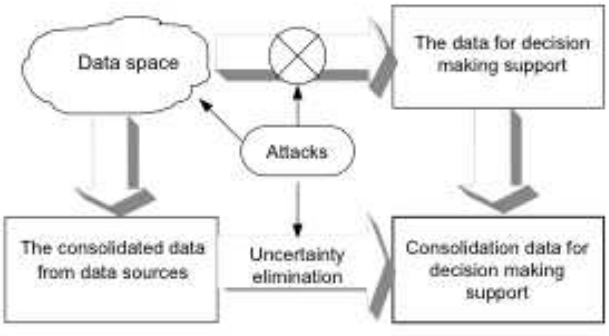

Fig. 2: The task stating of elimination indeterminacy in datawarehouse and in data space

4 The main material

4.1 The model of consolidated data

The scheme of consolidated data *Cg*' is a final set of attributes $\{A_1, A_2, \ldots, A_n\}$, set of attributes $\{A\_unk_1, A\_unk_2, \ldots, A\_unk_p\}$ with indistinct or non-determinate definitions and set of attributes $\{Unk_1, Unk_2, \ldots, Unk_m\}$, which domains are the numerical data, that are modeling 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 scheme of the data catalog *Cg*:

$$Cg' = \langle \{C_1, C_2, \ldots, C_n\}, \{C\_unk_1, C\_unk_2, \ldots, C\_unk_p\}, \{Unk_1, Unk_2, \ldots, Unk_m\}, Dic, Cg \rangle$$

The attributes in set *CUnk* are uncertain, and trust to them is stored in a set of attribute values *Unk*.

To show relationships between sets of attributes *CUnk* and *Unk* there is used binary relation *Meta*, whose value is determined based on a sample representing the source and the data directory *Cg*:

$$Meta = |meta_{ij} \cdot \sigma_{arg(i)}(Cg)|, \forall i = 1 \ldots p, \forall j = 1 \ldots m$$

$$meta_{ij} = \begin{cases} 1, Unk_j \Leftrightarrow C_u nk_i \wedge \sigma_{arg(j)}(Dic) \\ 0, otherwise \end{cases}$$

The sum of the rows of a binary relation is equal to 1, since we assume that the credibility of the attribute not be displayed two or more attributes from the set *Unk*:

$$\forall i = 1 \ldots p, [\sum_{j=1}^{n} .meta_{ij} = 1.]$$

The usage of to *Meta* relation allows to model any type of uncertainty is not expanding domains of attributes.

Cortege consolidated data *dc* is the information object description *t* data source *S*, disclosed in a set (tuple) values of characteristics (attributes), a subset of attribute values which contains information about the object data source object names and synonyms, and these information may be incomplete, unclear or non-determined data. The object modeled in the data source of this tuple is existing, but the information about it is missing, vague, incomplete, determined, and so on. Here are some examples tuple consolidated data for different types of information resources.

1. Relational database. In this case we used the extended relational tuple:

$$dc = t_{rel} \cup Unk, t_{rel} = \{c_1, \ldots, c_n\} \cup \{c\_unk_1, \ldots, c\_unk_n\},$$

where $\{c_1, \ldots, c_n\}$ are exact attributes and $\{c\_unk_1, \ldots, c\_unk_n\}$ are attributes with uncertainty.

2. Data warehouse is combining data from the relations of facts and dimensions. The set of values of the measurements and specifications are presented as facts tuple:

$$dc = t_{dw} \cup Unk,$$
$$t_{dw} = \{c_{11}, \ldots, c_{1n}\} \cup \cdots \cup \{c_{k1}, \ldots, c_{kn}\} \cup$$
$$\cup \{c_{rf_1}, \ldots, c_{rf_l}\} \cup \{c\_unk_{11}, \ldots, c\_unk_{kn}\} \cup \cdots \cup$$
$$\cup \{c\_unk_{11}, \ldots, c\_unk_{kn}\} \cup \{c\_unk_{rf_1}, \ldots, c\_unk_{rf_l}\},$$

where c_{ij} is the value of exact j-th characteristic in the i-th measurement, c_{rf_j} is the value of j-th characteristic in the facts relation, c_unk_{ij} is the value of j-th attribute with uncertainty in the i-th measurement, $c\_unk_{rf_j}$ is the value of j-th characteristic with uncertainty in the facts relation.

3. The tuple of semi-structured text source describes the importance of semantic networks vertices and the degree of membership of values to objects whose names are written in the dictionary of synonyms:

$$dc = t_{text} \cup Unk,$$
$$t_{text} = \{c_1, ..., c_n\} \cup \{c\_unk_1, ..., c\_unk_n\}.$$

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

Importances attributes of procession the consolidated data we will divide into groups.

© 2015 NSP
Natural Sciences Publishing Cor.

1.Exact (known) is the importance of the primary key, external key (may be absent). Mark them through C.
2.Absence is no information physically. Put them through.
3.Indeterminacy is for subsets of attributes introduced a set of attributes *Unk*, indicating a degree of truth values of these attributes. The default value of the attribute *Unk* assign value, which means the highest degree of truth.

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

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

$$dc = \langle C, C\_unk, Unk, \{dic\}, \{cg\}\rangle,$$

where C is a subset of attribute values with distinct values, C_unk — a subset of attribute values with lack of precession and non-deterministic values, Unk — a subset of attribute values with degrees of truth values of attributes C_unk, dic — the set of values of the data dictionary, cg — set values directory data.

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

Model of consolidated data contains data from all types of sources of data space.

4.2 Development operations on model of consolidated data

As the datawarehouse of the consolidated data is widening datawarehouse of the data constructed on the basis of relational model, we will improve operations for work with it further [6].

For processing and analysis of indeterminacies using query relational operators should exercise selection procession by the values of a set attributes *Unk*. In datawarehouse there is similar cut operation. Let r and s — related to the scheme R, r' and s' — 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 $r' - s'$ — relation to the scheme.

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

$$r' = r \cup \sigma_{max(P(\Pi(Cg)))}(Dic) \cup Cg$$

Addition to the relation is correctly to work in case of assignment to all values of the *Unk* attribute of the lowest degree of trust (a priori it is considered, what this information which is brought in the relation is truthful and full, and about the rest information of anything doesn't know to us). Election of such method of representation of degree of the 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*.

$$slice : \sigma^{cons}\left(\begin{array}{r} Unk\Theta unk) \cup (C\_unk\Theta c\_unk) \cup \\ \cup \sigma_C(Dic) \cup \sigma_C(Cg) \end{array}\right)(cg') =$$

$$= \left\{\begin{array}{r} r \in dc | t(Unk)\Theta unk,\ t(C\_unk)\Theta c\_unk, \\ meta_{Unk,C\_unk} = 1 \\ \sigma_C(Dic) \notin \varnothing,\ \sigma_C(Cg) \notin \varnothing,\ unk = P(cg') \end{array}\right\},$$

where Θ – set of symbols (marks) binary relations on pairs of values domains. It is believed that for each attribute used C_unk comparison operations. As a rule, will be used only compare such signs on one domain: $=, \neq, <, \leq, \geq, >$.

Statement: Improved operator cutting as the operator sampling preserves communicative properties and relatively distributional Boolean operations.

Proof. Let $r'(R')$ is a relation, $R \leftarrow R \cup Unk \cup Dic \cup Cg$, and are the attributes in R', and let $a \in dom(A)$, $b \in dom(B)$. Then equality holds: $\sigma^{cons}{}_{A=a}(\sigma_{B=b}(r')) = \sigma^{cons}{}_{B=b}(\sigma_{A=a}(r'))$.

Advanced operator of a cut distributive rather binary Boolean operations:

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

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

Analogue operation clotting in a datawarehouse, based on the relational model is the operation of projection. Through the projection ratio of processions consolidated data should track connection attributes *Unk* subset of a subset of attributes C_unk and check whether for the name attribute is C_unk synonym in the dictionary of synonyms Dic. Therefore, improved operator clotting presented as follows:

$$\text{drill} - \text{down} : \pi^{\text{cons}}_{\text{X}}(\text{cg}') =$$

$$= \text{IIF}\left(\begin{array}{r} \neg\text{ISNULL}\left(\sigma_{\text{Cg=R}\cup\text{C\_unk=X}}(\text{c\_unk})\right); \\ \pi_{X \cup \pi_{Unk}(\sigma_{Cg=meta(C\_unk,Unk)=1}(c\_unk))}(dc); \\ \text{IIF}\left(\sigma_{\text{C}\cup\text{C\_Unk=X}}(\text{Dic}); \pi_{\sigma_{\text{C}\cup\text{C\_Unk=X}}(\text{Dic})}(\text{r}); \pi_{\text{X}}(\text{dc})\right) \end{array}\right)$$

where $\text{IIF}(\text{condition}; \text{operation1}; \text{operation2})$ — 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 ratio r operand does not contain processions and defect — in that case. Also the search attribute synonym in the dictionary of synonyms $Dic(\sigma_{C \cup C\_Unk=X}(Dic))$ and replacement needs $(\pi_{\sigma_{C \cup C\_Unk=X}(Dic)}(r))$.

Statement: Enhanced coagulation operator maintains its traditional projection operator.

© 2015 NSP
Natural Sciences Publishing Cor.

Proof. If $X_1 \subseteq X_2 \subseteq \cdots \subseteq X_m \subseteq R'$, then $\pi^{cons}_{X_1}\left(\pi^{cons}_{X_2}\left(\ldots\left(\pi^{cons}_{X_m}(cg')\right)\ldots\right)\right) = \pi^{cons}_{X_1}(cg')$.

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

Traditional connection operator can not be used for dataspace and datawarehouse with consolidated data, since statistical analysis necessary connection related facts relational dimensions, and if non-empty subsets of attributes *Unk* in respect of the facts and dimensions of such a connection is incorrect. Also on operator connections affected by the fact that there is a need not only connect on those attributes specified as input parameters, but also check for synonyms in a dictionary of synonyms *Dic*. For improving service connection 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* contain 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* for procession subordinate tables that do not fall in the ratio will be equal to the highest degree of confidence [6].

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

where r is traditional relation, cg' is 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 $Unk(Cg' = Cg \cup Unk)$, is the set of attributes with S, which are not covered in terms $r = (B \subset Cg, B \not\subset Cg \cap R)$, *min* is the importance, which means the lowest level of faith, $NVL(Unk, min)$ is the operation that assigns min for all importance *Unk* for connecting related processions $cg', \triangleright\triangleleft$ is the left connection. At first it is checked, it is necessary carry out connections on set by attributes, and behind synonyms $\sigma_{C\cup C\_Unk=X}(Dic)$. If not, then the operation left connection for relations with schemes S' and R and the projection of the attributes-synonymous.

Otherwise operation of the left connection on the general is carried out by attributes, and then over the relation received from the previous operation of a projection on which the empty value of a subset of the Unk attributes formed as a result of connection appropriates min value is carried out.

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

Statement: Advanced operator of connection commutative and associative.

Proof. For these relations: q', r' and s'

$$\left(q' \bowtie r'\right) \bowtie s' = q' \bowtie \left(r' \bowtie s'\right)$$

Let's enter designations for some repeated connections. Let $s_1'(S_1'), s_2'(S_2'), \ldots, s_m'(S_m')$ are relations, $R' = S_1' \cup S_2' \cup \cdots \cup S_m'$ and S' is sequence of $S_1', S_2', \ldots, S_m'$. Then let $t_1, t_2, \ldots, t_m$ is the sequence of procession, where $t_i \in s_i', 1 \le i \le m$. Processions connecting to S', if there is a procession $t \in R'$, that $t_i = t(S_i'), 1 \le i \le m$. Procession is the result of combining processions $t_1, t_2, \ldots, t_m \in S'$.

4.3 Reducing indeterminacy 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 is much more common data sets in which important dependencies defined only on a subset of the values of key attributes, call the following dependencies partial functional dependencies. That is, a partial functional dependency - a FD defined in some fixed ratio selection.

$$F_p : K = \{a_i\}, a_i \in A, D = \{a_j\}, a_j \in A, \\ R' \subset R : K \to D | R' \quad (1)$$

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

Probabilistic productive relationship — this production rules in the selection of the basic attitude that holds significant number of objects for this selection. The threshold of significance should be determined by expert, or based on calculations of the probability of false selection of this relationship.

$$F_I : K = \{a_i\}, a_i \in A, D = \{a_j\}, a_j \in A, \\ P(k \in K \to d \in D) = p \quad (2)$$

Here k and d — procession of values of certain groups of attributes K and D, respectively.

The main indicator of the reliability of such dependence is the ratio number of objects for which there is probabilistic productive relationship the number of objects in the selection:

$$P(F_I) = \frac{|\sigma_{k\in K \wedge d \in D}(R)|}{|\sigma_{k\in K}(R)|} \quad (3)$$

© 2015 NSP
Natural Sciences Publishing Cor.

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

It will construct a classification rule based training data set cg', in which the value tag class (meaning a subset of attributes Y) are known. Classified generally built for the scheme cg', and will therefore not be affected by the new tuples arriving in the ratio of consolidated data repository (independence from the test set).

Mark of class called linguistic variable characteristic habitual or objects that are the values 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 get out from in advance known sets of values (within studied area are fixed), and reference to a class of objects information about which just arrived in datawarehouse with the consolidated data, is carried out on the basis of classification rules.Additions of marks are carried out automatically, as receipts of new data sources in space of data - also dynamic.

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

$$P(s \in S \to t \in T) = \sum_{t_i \in T} P(s \in S \to t = t_i) = \sum_{t_i \in T} \frac{\sum_j |s = s_j \wedge t = t_i|}{\sum_j |s = s_j|} \quad (4)$$

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 axioms for functional dependencies for classification rules.

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

Proof. $P(s \in S \to s \in S) = \frac{|\sigma_{s \in S \wedge s \in S}|}{|\sigma_{s \in S}|} = \frac{|\sigma_{s \in S}|}{|\sigma_{s \in S}|} = 1$

Replenishment: If $P(s \in S \to t \in T) = p$, then $P(s \in S \wedge w \in D(W) \to t \in T) = p$

Proof.

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

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

Eliminating indeterminacies occur among attribute values Y relation r, is classified using the modified algorithm chase.

The point of the method:

1. search for procession that have the same values in the set of attributes X;
2. search for procession that have the same values in the set of synonyms attributes X;
3. calculation of the level of confidence in the source procession obtained in steps 1) and 2);
4. calculation of confidence to attribute sources of procession obtained in steps 1) and 2).
5. determining the procession with the highest level of confidence.

If we are able to classify the objects its necessary to build classification functions. Generally, in the space of data can be stored information about several types of classes, 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 call the modified functional relationships that are performed for a specific subset of procession related consolidated data repository.

Algorithm referring to the class:

1. If $\sigma(cg') = \{dc_1(X_1)\downarrow, \ldots, dc_1(X_n)\downarrow\}$ and $\{dc_2(X_1)\downarrow, \ldots, dc_2(X_n)\downarrow\}$
   And
   $\{dc_1(X_1)\downarrow, \ldots, dc_1(X_n)\downarrow = dc_2(X_1)\downarrow, \ldots, dc_2(X_n)\downarrow\}$
   And $\{dc_1(Y)\downarrow\}$
   And $\{dc_2(Y) = \perp\}$
   And If $\sigma_{X_1}(Dic) = \emptyset$
   Then replaced $\perp$ by $dc_1(Y)$ and $dc_1(P) = dc_1(P)/(\sum_i {}^{m_{1i}}/_{n})$
2. If $\{dc_1(X_1)\downarrow, \ldots, dc_1(X_n)\downarrow\}$
   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 importance $dc_1(X^m)\downarrow = dc_2(X^m)\downarrow\}$
   And $\{dc_1(Y)\downarrow\}$ and $\{(Y) = \perp\}$,
   Then replaced $\perp$ in r $dc_1(Y)$ and $dc_2(P) = dc_2(P)/(\sum_i {}^{m_{2i}}/_{n})$

© 2015 NSP
Natural Sciences Publishing Cor.

3.If $\{$ in $dc_i m_i$ with n importance of attributes $-\downarrow, m_j \leq n\}$
And $\{$ in $dc_j m_j$ with importance of attributes $-\downarrow, m_j \leq n\}$
And $\{$ on certain importance $dc_i(X^m)\downarrow = dc_2(X^m)\downarrow\}$
And $\{m_i/n \leq m_j/n\}$ and $\{P \geq 1 - m_i/n\}$
And $\{dc_i(Y)\downarrow\}$ and $\{dc_j(Y)\downarrow\}$ and $\{dc_2(Y) = \perp\}$
Then change $\perp$ in $dc_j(Y)$ and $dc_2(P) = dc_2(P)/(\sum_i {}^{m_{2i}}/_n)$

Example:

Let we have a database of research part and scientific reports of two chairs. It is clear that between the specified sources there is dependence: hit given in a database of research part through scientific reports of chairs (departments). Hierarchical organization of scientific reports processing is presented in the fig. 3.

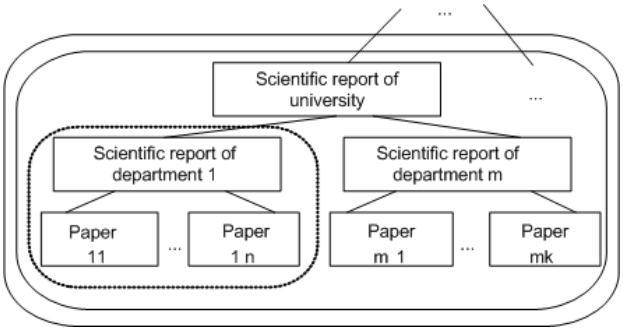

Fig. 3: Hierarchical organization of scientific reports processing

Let to datawarehouse with the consolidated data as a result of integration got two processions with such contents:

File with the first department's scientific report:

...
1.2. Articles in foreign journals included in scientometric databases
1. *Shakhovska N.* Algebraic system of data space **International Journal of Science and Technology Education Research** Vol. 4(2), pp. 30 - 37, February 2013; http://www.academicjournals.org/IJSTER/PDF/pdf2013/Feb/Schakhovska.pdf
...
1.3. Conference Proceedings (abstracts) in the publications included in scientometric databases

2. Articles in professional journals Ukraine:
1. *Shakhovska NB Model consolidated data and their processing under conditions of uncertainty, Computer Science and Information Technology, №715, 2013. - S.203-213*
2. *Shakhovska NB The method of transformation of algebraic operations of the "data space" in the real model data sources, Proceedings of the National University "Lviv Polytechnic", "The Information Systems and Networks", № 743,2012, S.161-175*

Fig. 4: The first department's scientific report

File with the second department's scientific report:

And we have such database schema of scientific department:

As result we have two similar tuples with information about same publication:

1.2. Articles in foreign journals included in scientometric databases
1. *Natalya Shakhovska, Mykola MEDYKOVSKY, Petro STAK.Application of algorithms of classification for uncertainty reduction .PRZEGLĄD ELEKTROTECHNICZNY, ISSN 0033-2097, R. 89 NR 4/2013 pp 284-286.*
2. *Natalya Shakhovska, Mykola Medykovskyj, Vasyl Lytvyn Dataspace Class Algebraic System for Modeling Integrated Processes .JOURNAL OF APPLIED COMPUTER SCIENCE. - Lodz Vol. 20 No. 1 (2012), pp. 69-80.11.*
3. *Shakhovska NB The method of transformation of algebraic operations of the "data space" in the real model data sources, National University "Lviv Polytechnic", "The Information Systems and Networks", № 743,2012, S.161-175*

Fig. 5: The second department's scientific report

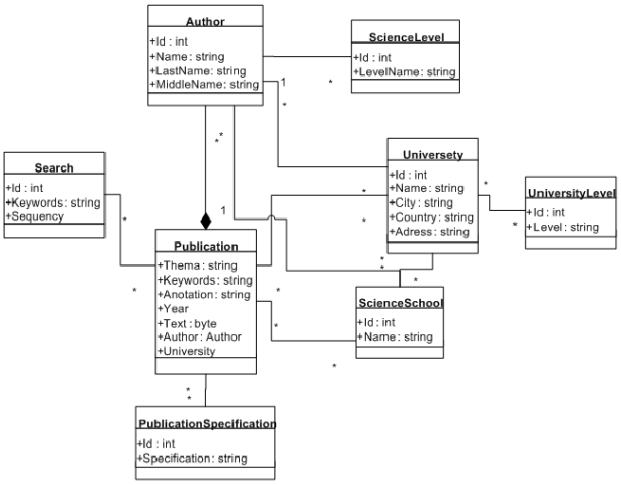

Fig. 6: The database schema of scientific department

Received indeterminacy of a look 'a polysemy of interpretations' (distinctions value of the Publisher attribute).

In the catalog of data is specified that Text32 is data source for DB1. Then, without looking that trust level to the first is higher than a procession, than in the second, the second procession will get resultant selection.

5 Conclusion

1.There is introduced consolidated datawarehouse model as an improved model of the uncertainty relation. It is simulated the physical object – a datawarehouse, which one indicating it attributes with clear and undefined values – to reduce uncertainty and taking into account the existence of public confidence in the data source to increase the effectiveness of management decisions. There is analyzed the causes of uncertainties in storage and data spaces. Among them are highlighted: the uncertainty in the scheme of the mediator; uncertain schema mapping, data uncertainty consolidated data repository.

2.It is improved operations over the relation with indeterminacy for the purpose of their application in datawarehouse with the consolidated data that allowed realizing unary operations of data space.

© 2015 NSP
Natural Sciences Publishing Cor.

Table 1: The result of data integration

ID	Author	Title	Publisher	Source	Trust
1	Shakhovska N.	The method of transformation of algebraic operations of the ”data space” in the real model data sources	Lviv Polytechnic National University	DB1	0,7
2	Shakhovska N.	The method of transformation of algebraic operations of the ”data space” in the real model data sources	Proceedings of the National University ”Lviv Polytechnic”	Text32	0,4

3.A 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.

4.Considered methods are useful also for decision making, as it provides a search for hidden relationships between the characteristics of consolidated data repository. Such dependence should be considered when making decisions on the basis of consolidated data.

References

[1] Zadeh, L. The concept of a linguistic variable and its application to the adoption of approximate solutions, L. Zadeh-M., New York, 1976. - 166 p.

[2] Tselmer G. Risk consideration in management decisions, G.Tselmer, Problems of ICSTI, 3, p.94-105, 1980.

[3] Knight F.KH. Risk, uncertainty and profit., F.Kh.Nayt M., Business, 358 pages, 2003.

[4] Moiseyev N. N. Elements of the theory of optimum systems, N.N.Moiseev M, Science, 528 pages, 1975.

[5] Trukhachev R. I. Decision-making models in the conditions of uncertainty. M, Science, 151 with. 1981.

[6] Shakhovska NB Model of Consolidated Data and Its Formulating with Uncertainty. Natalya Shakhovska, Mykola Medykovskyj, Vasyl Lytvyn, Janusz Lipinski, Journal of Applied Computer Science, Vol. 22 No. 1 (2014), Lodz, pp.213-221, 2014.

Nataliya Shakhovska received the doctor of information technologies in Lviv Polytechnic National University of Lviv, Ukraine. She is professor in Information Systems and Networks Department, dean of Institute of computer sciences and information technologies.Current research interest: Database and datawarehouse integration, distributed systems, integrated systems and dataspaces. She has published more than 160 scientific papers, 2 monographs, 4 textbooks.

Rostyslav Strubytskyi is master of computer sciences of Ternopil State University, PhD student in Lviv Polytechnic National University. Current research interest: cloud computing, parallel processing, datawarehouse, integration.

© 2015 NSP
Natural Sciences Publishing Cor.