Natalya SHAKHOVSKA, Mykola MEDYKOVSKY[1], Liliana BYCZKOWSKA-LIPINSKA[2]

Lviv Polytechnic National University (1), Technical University of Lodz (2)

One approach to dataspace modelling

Abstract. *The paper analyses the problems arising in processing the separate information sources and databases. The physical model of dataspace is presented, covering the formalization of search methods of the unstructured, semi-structured and strictly structured data.*

Streszczenie. *Praca analizuje problemy przetwarzania danych pojawiające się podczas pracy z oddzielnych źródeł informacji. Prezentowany jest model fizyczny przestrzeni, nacisk kładący na formalizm metod przeszukiwania różnego rodzaju danych.* (***Modelowanie przestrzeni danych***)

Keywords: : Dataspace, intelligent agent, integration methods, consolidation, information product.
Słowa kluczowe: Przestrzeń danych, inteligentne agenta, metody integracji, konsolidacji, Informacje o produkcie.

Introduction

The physical system is dynamic and its elements have evolved at different rates. It complicates the collection and processing of information on elements of such a system. To work with different types of information from different sources, we can apply dataspace. One element of the dataspace is an information product.

Information Product (IP) is a documented information resource, prepared according to the needs of users and submitted as product.

Catalogue of Information Product – metadata about information products – describes the location of an information product, its structure, methods of access to the information resource, etc.

Traditionally, experts used usual for them sources of information for solving tasks [1,2]. Apparently, this approach has incomplete information, which is processed. Many sources of data and services that exist on the Internet are causing the need for a radical change in the methods of getting data. The adoption of adequate solution require the data, coming from different sources to satisfy the following requirements: be complete, consistent and received on time; be informative, because they should be applied for decision support; be of uniform structure for the opportunity of being downloaded in single datawarehouse and analyzed; kept in uniform models of data and be independent of the development platform for the opportunity of using this data in other means. But today there are no data processing methods that would satisfy all of the requirements for data processing [3,4].

Another problem, that persist in the process of consolidation is the uncertainty of data, the result of duplication, inaccuracies, data absence, contradictions of the data. Another problem is the determination and approval of the schemes of the data of information resources. The existing methods are working either on the known data schemes, the data source are under strict control, which makes it impossible to set various semantic relationships [4,5].

The main part

Dataspace is a set of all information product domain (databases, datawarehouse, web pages, text files, spreadsheets, image data respectively). The databases, datawarehouses, text files, spreadsheets that are described in different formats are used.

Talking about an IP, we mean its content (information resource, IR), and the set of information about it (accommodation, access scheme, speed of information update, etc.). We are also interested in the operations which are carried out over IR depending on its DSIR. The main task of dataspace is to allow the user to work with data sources without knowing its DSIR, accommodation, access methods, etc.

Consolidated data is derived from multiple sources and systematically integrated heterogeneous information resources, which together are have such features as completeness, integrity, consistency and adequacy. This consolidated information is model of the subject area for its analysis and processing efficiency in the processes of decision making.

Data sources such as spreadsheets, multimedia information, etc, 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.

DSIR determination is carried out by using intelligent agent and means the addition into the *Cg* the new data about IP DSIR

(1) $$f_{Ip}(DS) \xrightarrow{Agent} Cg \cup Ip.Cg$$

where: *Cg* is data space catalogue, *Ip.Cg* – IP catalogue Ip.

The agent is: *Agent = < Cg, M, Dic, MB, Dif, H >*

where: *Cg* is information about sources that are already in the DS; *M* – a component of the agent responsible for the perception of the environment, which is the environment of model management; *Dic* – synonymic terms that indicate the sources of the same properties; *MB* – the base of agent experience containing "the history of impacts" on the agent from the environment and the corresponding agent reaction; *Dif* – the component that is responsible for training (provides a list of differences which showed an agent); *H* – the component responsible for the actions of the agent.

At the heart of intelligent agent is information about the sources, which are already in dataspace. His task is to compare the data structures of source that will include in the dataspace and data structures in the dataspace, and determine the difference.

Firstly, agent choices the method of access to data source:

$$MB:\begin{cases} struct(Ip.Rl), M=1 \\ parse(Ip.Rl), M=2 \\ S(Ip.Rl), M=3 \end{cases},$$

where source can be structured (database, datawarehouse, xml), semi-structured (spreadsheet, html) or unstructured (text).

Secondly, agent sets the type of relationship H to determine the characteristics that have to get in the query result.

1) Equivalence relation: $x = y \Rightarrow H(x,y)=1$; $P^{Y}(IP_{new})=1$,

where $P^{Y}(IP_{new})$ is the trust to attribute in IP_{new}.

2) Synonymous relation:

$\forall x \in X, \exists y \in Y : X \neq Y, X \subset Dic, Y \subset Dic \Rightarrow H(x, F(x)) = 1$,
$P^Y(IP_{new}) = 1$

3) Conversion relation: $y = F(x)$.

$\forall x \in X, \exists y \in Y : X \neq Y, X \subset Dic, Y \subset Dic \Rightarrow H(x, F(x)) < 1$,
$P^Y(IP_{new}) = 0{,}5$

4) Generalization relation: *Y* is the generalization of *X*:

$F : X \to Y, y \neq F(x). \forall x \in X, \exists y \in Y : X \subset Dic, Y \subset Dic \Rightarrow H(x, y) < 1$,
$P^Y(IP_{new}) = 0{,}5; P^X(IP_{new}) = 0 \cdot Dic \setminus X \cup Y \to Dic$

5) Isomorphic of domains:

$F : X \to Y, y \neq F(x) \exists x \in dom(X), \exists y \in dom(Y) : X \subset Dic$,
$Y \not\subset Dic \Rightarrow H(x, y) < 1, P^Y(IP_{new}) = 0{,}25; P^X(IP_{new}) = 0{,}5;$
$Dic \cup Y \to Dic$

6) Relation of polarity $F : X \to Y, F^{-1} : Y \to X, X \subset Dic, Y \not\subset Dic$,
$P^Y(IP_{new}) = 0; P^X(Cg) = 0,\ X \cup Y \to Dif$; $H(x, y) = 0$.

$$\left.\begin{array}{c} F : X \to Y, y = F(x) \\ \forall x \in X, \exists y \in Y : X \subset Dic \\ P^X(Cg) = 0, Mb <> 0 \end{array}\right\} \to \begin{cases} X \cup Y \to Dif \\ P^Y(IP_{new}) = 0 \end{cases}.$$

To evaluate the quality of consolidated data we used general methodological approach to the selection of an adequate range of standardized in ISO 9126 basic features and sub-characteristics.

Functionality is relative number of objects that hit the storage of consolidated data cg', to the total number of objects available in the IP.

(2) $$z_1 = \frac{|cg'|}{|Ip_i.Ir|}.$$

To ensure the functionality of attached IP we match SDIR with isomorphic of domains and generalization (the result of intelligent agent).

Correctness or reliability of the data is relative number of descriptions of objects with IP, which do not contain defects and errors to the total number of objects in dataspace:

(3) $$z_2 = \frac{|\sigma_P(Ip)|}{|cg'|}.$$

To ensure the correctness of the enclosed PI matched SDIR equivalence.

Usability determines the usefulness of the application of consolidated data for specific users. In the DS estimation practicality is implemented by the utility function of the decisions $v(r)$:

(4) $$z_3 = \frac{v(r)}{|cg'|}.$$

Moreover, this characteristic takes into account the dependence of the decision on the level of confidence. To provide usability we add IP with equivalence and conversion relations.

Portability is convenience and efficiency corrections, improvements or adaptation of the structure and content description of data depending on changes in the external environment of use. In the dataspace portability associated with the change of IP data in the catalog:

(5) $$z_4 = \frac{|\sigma_{meta\_upd}(Ip_i.Cg)|}{|cg'|}.$$

To ensure portability we add IP with synonymous relation.

Now we determine the usefulness of data from the IP on a decision based on them. Assessing the usefulness of data carried as follows. There is a set of controlled variables $Z = (z_1, z_2, z_3, z_4)$. We defined a continuous function Q. The statement: objective function as the existing restrictions on a global maximum:

(6) $$Q(z_1, \ldots, z_4) = \sum_{i=1}^{4} \left(\sum_k r_k z_i \prod_{j=1} P_{ij} \right) \to \max,$$

where: *j* is the number of information product, P_{ij} – the trust to information product *j* for solution *k*, r_k – evaluate of the solution *k*, C_1 – total cost of loading facilities, C_2 – total cost of modifications of descriptions, T – the total load time, t_s – the average download time of one object, c – the average cost of loading (modification) of one object.

$$\begin{array}{l} 1 \geq z_1 \geq 0.75; z_3 \geq 0; \\ 0.25 \geq z_1 - z_2 \geq 0; \\ 1 \geq z_4 \geq 0.5; z_1 t_s \leq T, z_2; \\ z_1 c \leq C_1; z_4 c \leq C_2 \end{array}$$

This nonlinear optimization problem with linear constraints, which is solved by certain methods.

Along with the actual evaluation of the quality of the consolidated information (7) is necessary to evaluate the quality of the reference sample (8), reflecting the best decision. Then the actual evaluation is performed normalization by the formula (9).

(7) $$Q_i^e = \sum_i n_i z_i^e, Q'^e_{const} = \sum_i k_i Q_i^e,$$

where k_i – the rank of importance, $k_i \in [0; 1]$,

(8) $$Q_{const} = Q'_{const} \Big/ Q'^e_{const}.$$

Formula (9) used to make decisions on: reducing uncertainty; determine if adding PI.

A method for managing elements of DS based on the function of the quality and levels of confidence is given (10).

Set $norm : [\min, \max]$

Set ex

$Sm(P <= ex) \to Ip$

(9) $$f_{man} : \begin{cases} Q > \max(norm) \to \text{Set } norm : [\min, \max], \text{Set } ex \\ Q > \text{Avg}(norm) \to profile_{us}\left(\pi_{P_{Ip} <= ex}(Cg)\right) \\ \exists(P_{Ip}) = 0, Q \approx \text{Avg}(norm) \to f_{Ip}(Cg) \\ Q < \text{Avg}(norm) \to DS - O_{P0}\left(\pi_{P_{Ip} <= ex}(Cg)\right) \\ Q < \min(norm) \to \begin{array}{c} f_{Ip_{new}}(Cg) \\ Se(Ip_{new}) \end{array} \end{cases}.$$

Scenario management:

- set a higher value thresholds as a function,
- override user rights,
- conversion levels of trust, redefinition of conformity,
- removal businessman, whose confidence is less than specified,
- search for new IP, which would improve data quality.

The result of intelligent agent is consolidated data from information products.

Example of domain for which we must build dataspace is University. This object is characterized by hierarchical structure, large volumes of information. It must solve tasks of complex analysis. Information for high school usually needs to make data integration, because at the time of developing a single datawarehouse has many information product, which must share information among itself, and provide some information in the corporate datawarehouse with the purpose of analytical processing.

Analytical tasks that should be solved by university analysts include:

- search dependencies between estimates of students in subjects and results of entry;
- search of courses in which performance “Progress”, “Quality” are very high or very low;
- search of dependencies between results scientific activities of students and their practical results in the form of practice, participate in competitions, contests work, etc.

Existing software products perform tasks which are entrusted to it and don’t always allow to solve new tasks. For example, “Lviv Polytechnic” developed such information systems to automate and support the learning process (Fig. 1).

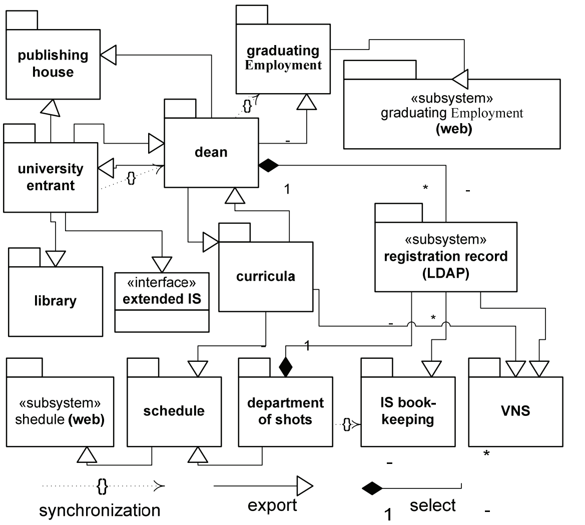
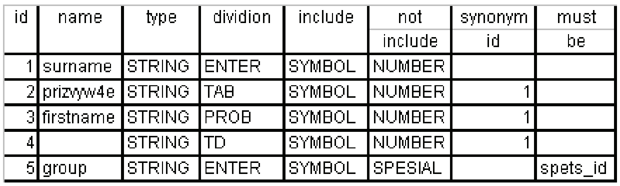


Fig.1. Interaction scheme between the main university databases.

- “Entrant” – accounting admissions to first year Bachelor and Master degrees formation orders to enrol, etc.;
- “Curricula” – development and accounting curricula in the specialties;
- “Dean” – accounting students, accounting individual curricula, accounting and analysis student success;
- “Schedule” – accounting audience fund formation of schedule lessons and exams;
- “Graduate Employment” – analysis quality of graduates, accounting practice and diploma projects students.

The systems save on the server with SQL Server 2005. The consolidation and replication technologies are used as a method of integration. This method copies certain data from one system to another. Features of these systems are: hierarchy of objects; they use various means of processing and data analysis; data processing from sources doesn’t adapted for integration; they need the processing of streaming data and data that arrives late.

The result of intelligent agent is the synonymic dictionary *Dic* (Fig. 2).



id	name	type	dividion	include	not	synonym	must
					include	id	be
1	surname	STRING	ENTER	SYMBOL	NUMBER		
2	prizvyw4e	STRING	TAB	SYMBOL	NUMBER	1	
3	firstname	STRING	PROB	SYMBOL	NUMBER	1	
4		STRING	TD	SYMBOL	NUMBER	1	
5	group	STRING	ENTER	SYMBOL	SPESIAL		spets_id

Fig. 2. The Synonymic dictionary

Further attributes are defined, in which the agent attempts to record inputs. This displays the first record of each block of data sources. If the agent does not find attribute value, then it leaves it blank. Meaning you can fill yourself.

The data completeness and usefulness of accumulated decisions are affecting the quality of the dataspace. Completeness collected descriptions of objects is the relative count of objects or documents available in the data source to the total count of objects, which hit the local store (illustrated in Fig. 3).

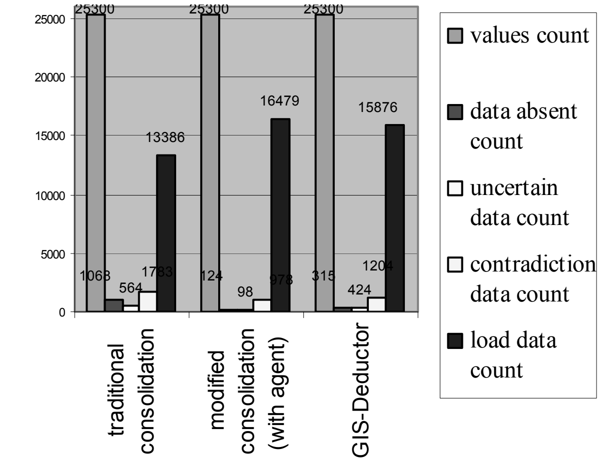


Fig. 3. Analysis of accumulated complete descriptions of objects.

This algorithm is compared modified integration algorithm which used in the Oracle Data Integrator, GIS-Deductor and our method. Count of input records databases that get in the local data repository is 15000. Without a prior determination of the structure of the data sources was not loaded in the repository of consolidated data (count of absent data in the chart), some data could not be loaded because of the discrepancy between the structure of local storage and structure of sources (uncertain data in the chart). Apparently, the number of tracks using the consolidation with agent is the largest.

Conclusion

In this article the physical model of dataspace is presented. Further investigation will cover the formalization of search methods of the unstructured, semi-structured and strictly structured data and building the appropriate algorithms.

REFERENCES

[1] Turban, E. (1995) Decision support and expert systems: management support systems. -Englewood Cliffs, N.J.: Prentice Hall.

[2] Data Warehousing: Similarities and Differences of Inmon and Kimball (2005), http://www.b-eye-network.com/view/743

[3] D. Kossmann, J.-P. Dittrich. Personal Data Spaces. (2007) http://www.inf.ethz.ch/news/focus/res_focus/feb_2006/index_DE.

[4] Рогушина Ю.В., Гладун А.Я. (2007) Формирование тезауруса предметной области как средства моделирования информационных потребностей пользователя при поиске в Интернете //Вестник компьютерных и информационных технологий.– М.: 2007. – № 1. – С.26–33.

[5] ETH - Databases and Information Systems – iMeMex, (2007) www.dbis.ethz.ch/research/current_projects/iMeMex

Authors: *PhD, Assoc.prof. Natalya Shakhovska, natalya233@gmail.com; doctor, prof. Mykola Medykovsky; doctor, prof. Liliana Byczkowska-Lipinska*