УДК 004.9

TEXT INFORMATION ONTOLOGICAL ANALYSIS IN THE COMPUTER SIMULATION SYSTEMS

Oleksandra Bulgakova

Mykolaiv V.O.Suhomlynsky National University, Nikolska str., 24, Mykolaiv, 54030, Ukraine
sashabulgakova@list.ru

В статті пропонується підхід автоматизації онтологічного аналізу текстової інформації з використанням узагальненого ітераційного алгоритму (ОІА) індуктивного моделювання. Описано технологію збору і сортування інформації, яка включає в себе чотири етапи: формалізація вхідних даних (витягування даних із зовнішніх джерел, їх трансформація та завантаження в сховище); аналіз даних (за допомогою ОІА індуктивного моделювання); визначення інформації до конкретної онтології (екземпляру онтології); створення нових онтологій (екземплярів) на основі проаналізованої інформації.

Ключові слова: аналіз даних, онтологія предметної області, онтологічна інформація, індуктивне моделювання, узагальнений ітераційний алгоритм, структури даних, обробка та зберігання інформації.

The paper proposes an approach for automation ontological analysis of text information using generalized iterative algorithm (GIA) of inductive modeling. The technology for the information collecting and sorting, which includes four phases: input data formalization (extract data from external sources, their transformation and loading in the repository); data analysis (using the GIA inductive modeling); information definition to a specific ontology (ontology instance); new ontologies creation (instances) based on analyzed information.

Keywords: data mining, domain ontology, ontological information, generalized iterative algorithm , inductive modeling,, structures of data, handling and storing of information.

В статье предлагается подход автоматизации онтологического анализа текстовой информации с использованием обобщенного итерационного алгоритма (ОИА) индуктивного моделирования. Описана технология сбора и сортировки информации, которая включает в себя четыре этапа: формализация входных данных (извлечение данных из внешних источников, их трансформация и загрузка в хранилище); анализ данных (с помощью ОИА индуктивного моделирования); определение информации к конкретной онтологии (экземпляру онтологии); создание новых онтологий (экземпляров) на основе проанализированной информации.

Ключевые слова: анализ данных, онтология предметной области, онтологическая информация, индуктивное моделирование, обобщенный итерационный алгоритм, структуры данных, обработка и хранение информации.

Introduction. With the growth of the accumulated information databases requires new data mining methods, algorithms and software for provide access to information, many of which should be classified as artificial intelligence systems systems of knowledge processing. The development of adequate and relatively simple programs that will "extract" the knowledge of the data, will greatly facilitate the work of human.

One of the most effective approaches to the text documents meaning detection and processing is the ontologies [1]. An ontology defines the terms used to describe and represent the knowledge of a particular subject area. Ontologies include computer processing for the basic concepts definition in the domain and the

relationships between them [2]. To obtain a database of ontologies and their models can be used inductive self□organization models based on experimental data (inductive modeling). This approach to modeling instead of the traditional deductive path "from the general laws operation of the facility − a particular mathematical model" is used an inductive approach "from specific observations − to the general model": the researcher hypothesizes about the possible models class and sets the criteria to choose the best models in this class. Computer processing allows to minimize the influence of subjective factors and get the model as an objective result [3□4]. Ontology model is obtained as algorithm result.

1. Domain ontology

Formally, an ontology can be defined as the set of $O = (L, C, F_l, F_C, R_h)$, where $L = \{(w_i, x_i)\}_{i=1,n}$:

L □ glossary domain,

w_i □ term,

x_i □ term rating relative to the other terms,

C □ concepts set,

$F_l(L) \to C$ □ concepts function interpretation that associates each concept a terms set from the dictionary,

R_h □ hierarchy relationship between the concepts [4].

The domain ontology describes the scientific knowledge domain, defined by specific subject. It may include a defined concepts hierarchy built on ontology concepts. All these hierarchies can be linked through associative relationships, some of which will be inherited from the basic technologies, and some will reflect the specifics of the subject area. Introducing concepts formal descriptions and problem domain in the concepts form and relations between them, the ontology should be asking structure for representing real-world objects and their relationships that composes the knowledge base.

Thus, the data will be presented in the form set of information objects different types and the relationships between them. Information object, we assume a data representing a set of text information specific area, relevant to some notion of ontology. To determine the appropriate ontology, the text information will be analyzed using a generalized iterative algorithm of inductive modeling.

2. Collecting and sorting information technology

Collecting and sorting information technology includes the following steps:

1. Formalizing the input data (extract data from external sources, their transformation and loading in the repository);
2. Data analysis/mining (using inductive modeling GIA)
3. Information determination to a specific ontology (ontology instance).
4. Create a new ontology (instances) on the basis information analyzed.

Step 1. Formalization submitting input data

Data is a presentation of facts and ideas in a formalized form suitable for transmission and processing of information in some process [6].

At this step, each document is represented as a set of terms, the set of documents is divided into subsets of documents similar topic (clusters), this results in terms of one subjects group. This allows to establish a relationship between terms and concepts. Each term is characterized by the frequency of occurrence (weight).

Problem is solved using the algorithms of inductive modeling. To solve problems using algorithms inductive modeling inputs must be strictly formalized and reduced to a tabular form. To solve this problem, the data need to be extracted from external sources, transformed and downloaded into the repository. Deleting data is a copying from the operational systems, documents and other sources, providing data integrity and uniqueness. Transformation involves the transformation of data to overall appearance, delete the errors, bind to dimensions. Transfer of transformed data storage is performed on the stage image. Integrated into the system can be used as data for the construction of direct reports, and further analysis using data mining algorithms.

Then we analyze input data characteristics in the inductive modeling tasks on various parameters.

In [7] used set theory to formalize the presentation of data at each construction models stage using GMDH algorithms. We have the following components (built using analysis method of structural identification [8]):

$W = (X, Y)$ – data set (sequence N values random variable Y, that characterized M features X)

$\mathrm{W} = \{w_j\}, j = \overline{1, J}, J = n \cdot m, n = \overline{1, N}, m = \overline{1, M}$;

NW – norm data set $NW = \{\overline{w}_j\}, j = \overline{1, J}$;

F – classes of models set $F = \{f_k\}, k = \overline{1, K}$;

G – generators structures models set $G = \{g_l\}, l = \overline{1, L}$;

P – set of parameter estimation structures methods $P = \{p_r\}, r = \overline{1, R}$;

CR – models criteria set $\mathrm{CR} = \{cr_q\}, q = \overline{1, Q}$;

V – classification models set $\mathrm{V} = \{v_t\}, t = \overline{1, T}$.

Then constructing set process of all possible models can be represented as a direct product of components sets $Z = W \times NW \times F \times G \times P \times CR \times V$. Some set elements Z, as described $z_i = \{w_j, \overline{w}_j, f_k, g_l, p_r, cr_q, v_t\}$, $j = \overline{1, J}, k = \overline{1, K}, l = \overline{1, L}, r = \overline{1, R}, q = \overline{1, Q}, t = \overline{1, T}$, $i = \overline{1, I}, I = J \cdot K \cdot L \cdot R \cdot Q \cdot T$, will be considered as specific data that have been stored in an environment at a particular passage full cycle simulation.

Step 2: Analysis of data

Documents clustering will be made on the basis of generalized iterative algorithm inductive modeling (GIA).

Let us briefly consider the iterative structure of algorithm used for solving the general problem of search for a better model under such formulation:

$$f^* = \underset{f \in \Phi}{\operatorname{argmin}}\, CR(y, f(X, \hat{\theta}_f)), \tag{1}$$

where $\hat{\theta}_f$ is an estimation of parameters for any partial model $f \in \Phi$, CR is a model quality criterion for selection of optimal model.

The set Φ of models being compared can be formed by various generators of model structures of diverse complexities. All structure generators developed within the GMDH framework naturally divided into two main groups – sorting-out and iterative ones which differ by techniques of variants generation and organization of search of a given criterion minimum. For simulation will be used the generalized iterative algorithm, GIA GMDH, fig.1 [9].

Formally, in the general case for layer *r* define the GIA GMDH as follows:

1) the input matrix is $X_{r+1} = (y_1^r, ..., y_F^r, x_1, \ldots, x_m)$,

2) apply the operators:

$$y_l^{r+1} = f(y_i^r, y_j^r),\ l = 1,2,\ldots, C_F^2,\quad i, j = \overline{1,F} \tag{2}$$

and

$$y_l^{r+1} = f(y_i^r, x_j),\ l = 1,2,\ldots, Fm,\ i = \overline{1,F},\ j = \overline{1,m} \tag{3}$$

with a quadratic partial description

$$\begin{gathered} z = f(u,v) = a_0 + a_1 u + a_2 v; \\ z = f(u,v) = a_0 + a_1 u + a_2 v + a_3 uv; \\ z = f(u,v) = a_0 + a_1 u + a_2 v + a_3 uv + a_4 u^2 + a_5 v^2. \end{gathered} \tag{4}$$

3) for each description is the optimal structure (an example for the linear partial description):

$$f(u,v) = a_0 d_1 + a_1 d_2 u + a_2 d_3 v, \tag{5}$$

where $d_k, k = 1,2,3$, $d_k = \{0, 1\}$ are structural elements of the binary vector $d = (d_1\, d_2\, d_3)$ taking values 1 or 0 (inclusion or not a relevant argument). Then the best model will describe: $f(u, v, d_{opt})$, where

$$d_{opt} = \arg\min_{l=\overline{1,q}} CR_l,\ q = 2^p - 1,\ f_{opt}(u,v) = f(u,v,d_{opt}) \tag{6}$$

4) the algorithm stops when the condition $CR^r > CR^{r-1}$ is checked, where CR^r, CR^{r-1} are criterion values for the best models of (*r*–1)-th and *r*-th layers respectively. If the condition holds, then stop, otherwise jump to the next layer.

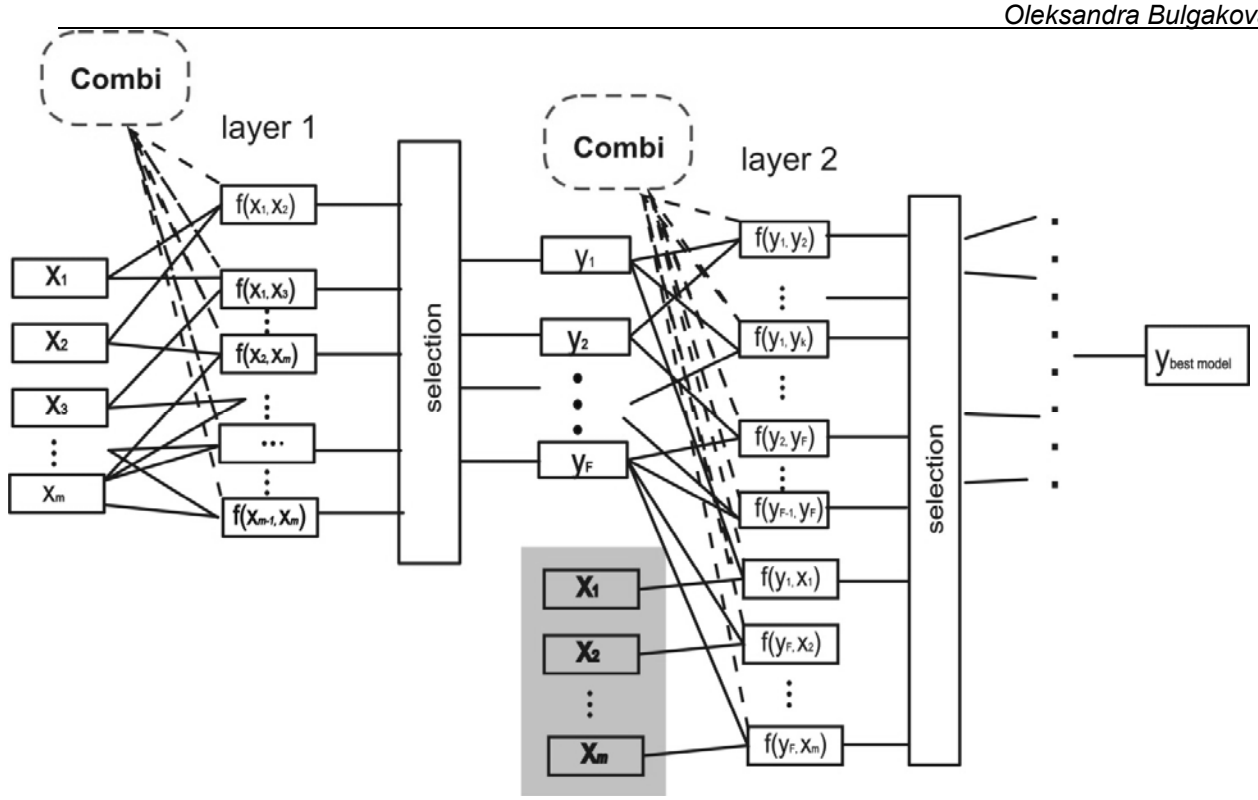

Fig.1. The generalized iterative algorithm schema

Define the GIA GMDH as many iterative and iterative combinatorial algorithms, described by vector of three elements DM (Dialogue Mode), IC (Iterative□Combinatorial), MR (Multilayered□Relaxative), ie any iterative algorithm is defined as a special case of a generalized: GIA (DM, IC, MR). This is possible with the help of specialized program complex of modeling based on iterative algorithms group method of data handling, which implemented the following features: automatic and interactive options for organization of user interface, management through the web interface, ensuring multiaccess. Constructed best model are presented by system for the graphic and semantic analysis, determined the effect of the arguments on the target factor, as well as analyzes and selects the most informative arguments [10].

Step 3: Definition of information to a specific ontology (ontology instance)

After GIA finished will be obtained "ontology model". At this step, the text information will be analyzed with the help of the models obtained for each ontology (ontology instance) sorted. Each model will have its own threshold (minimum and maximum) value based on the error simulation. Thus, as a result of the phase is determined not only set the partition areas of knowledge, which will include text, but also the conformity degree of the relevant sections document, which gives reason to stop or continue the analysis.

Step 4: Creation of the new ontology

At this stage, we have the opportunity to create new instances of ontologies, which are not in the current knowledge base. After the formalization of the input data and analysis can remain documents that were not related to any category. Such documents will be stored in a special data warehouse and analyzed at regular

intervals on the basis of which will constitute a glossary of terms. The dictionary will be stored semantic information, which will link elements of the dictionary, highlighting at the same time a new class of problem and domain.

3. Conclusion

The article describes the approach for automation ontological analysis of text information using generalized iterative algorithm (GIA) of inductive modeling. The technology for the information collecting and sorting, which includes four phases: input data formalization (extract data from external sources, their transformation and loading in the repository); data analysis (using the GIA inductive modeling); information definition to a specific ontology (ontology instance); new ontologies creation (instances) based on analyzed information.

References

1. Baeza□Yates R., Ribeiro□Neto B. Modern Information Retrieval. ACM Press, 1999.

2. T.R. Gruber. A translation approach to portable ontology specifications.

1. Acquisition, 5(2), 1993.

2. Степашко В.С. Теоретические аспекты МГУА как метода индуктивного моделирования // УсиМ. – 2003. – №2.– С.

3. Bulgakova O., Kordik P. Methods of true data mining model selection – with experimental results // Proceedings of 3rd International Workshop on Inductive Modelling IWIM□2009, 14□19 September 2009, Krynica, Poland. – Prague: Czech Technical University, 2009. – P. 23□27.

4. Zakharova I.V., Melnikov A.V.,Vokhmitsev J.A. «An approach to automated ontology building in text analysis problems».//Workshop on computer Science and Information Technologies CSIT'2006, Karlsruhe, Germany, 2006. P.177□178.

5. http://wikipedia.org/

6. Щербакова Н.В Формалізація структур зберігання інформації в задачах індуктивного моделювання // Моделювання та керування станом еколого□економічних систем регіону. Збірник праць. К.: МННЦІТС, 2009. – С. 229□ 234.

7. Ефименко С.Н., Степашко В.С. Имитационный эксперимент как средство для исследования эффективности методов моделирования по данным наблюдений // УСІМ. –2009. – №1. –С. 69□78.

8. Stepashko V.S., Bulgakova O.S. Generalized iterative algorithm of the group method of data handling // USiM. – 2013. – № 2. – P: 5□18.

9. Bulgakova O.S., Zosimiv V.V., Stepashko V.S. Program complex modeling of complex systems based on iterative algorithms with the ability of GMDH network access: 14□th International conference SAIT 2012, Kyiv, Ukraine, 176□178 p.