Features of Data Warehouse Support Based on a Search Agent and an Evolutionary Model for Innovation Information Selection

Vladimir K. Ivanov[1], Boris V. Palyukh[1], and Alexander N. Sotnikov[2]

[1] Tver State Technical University, 22, Quay A. Nikitin, Tver, 170026, Russia,
mtivk@tstu.tver.ru, pboris@tstu.tver.ru
[2] Joint Supercomputer Centre of RAS, 32a, Leninskiy Av., Moscow, 119991, Russia,
asotnikov@jscc.ru

Abstract. Innovations are the key factor of the competitiveness of any modern business. This paper gives the systematized results of investigations on the data warehouse technology with an automatic data-replenishment from heterogeneous sources. The data warehouse is suggested to contain information about objects having a significant innovative potential. The selection mechanism for such information is based on quantitative evaluation of the objects innovativeness, in particular their technological novelty and relevance for them. The article presents the general architecture of the data warehouse, describes innovativeness indicators, considers Theory of Evidence application for processing incomplete and fuzzy information, defines basic ideas of measurement processing procedure to compute probabilistic values of innovativeness components, summarizes using evolutional approach in forming the linguistic model of object archetype, gives information about an experimental check if the model developed is adequate. The results of these investigations can be used for business planning, forecasting technological development, investment project expertise.

Keywords: data warehouse, intelligent agent, subject search, genetic algorithm, innovativeness, novelty, relevance

1 Introduction

The basis model of R&D management involves the competitive analysis and forecasting of the technological development based on scientometric analytical services and semantic systems for searching commercially valuable information. This field features the obvious world-wide trend of using the global and already existing innovative potential. Innovations are the key factor of the competitiveness of any modern business, both in competitive and monopoly markets.

This paper is the first one to give the systematized review of the results of the works performed within the project of *Data Warehouse Support on the Base Intellectual Web Crawler and Evolutionary Model for Target Information Selection*. The objective of the project is to develop the theoretical basis for and pilot

implementation of data warehouse technology with an automated data supply from sources belonging to different subject segments. The data warehouse is suggested to contain technical, economic, social and other characteristics of objects having a significant innovative potential [1]. We suppose that the algorithms and technologies developed would be used for the expert system operation to control the evolution of multistage production processes [2] and [3]. Further, the article considers the tasks and basic results of the project, the architectural solutions of the data warehouse, objects innovativeness indicators, application of the theory of evidence, procedure for objects innovativeness calculation, evolutional approach to forming the linguistic model of the object, and some experimental results.

2 Project Tasks

The project results are supposed to be used in decision support systems (DSS) and expert systems to support solving the following applied tasks: *(a)* determine the characteristics of new domains in business planning; *(b)* forecast the technological development of the business; *(c)* provide an information support for expert groups and individual experts.

The researches focused on the development and experimental testing the model of the evolutional process of query generation and search results filtration. Further, some substantial aspects of the researches done will be detailed.

3 Data Warehouse Support System Architecture

According to [4], the software architecture includes layers of presentation, services, business logic, data access, as well as through functionality to provide the interaction of users and external systems with data sources. Fig. 1 shows the general architecture of the system. Below, there is a brief description of the composition and purpose of the basic components (circles):

1. *User's Applications.* Specialized applied systems for the information support of innovation implementation.
2. *Decision Support Systems. Expert Systems.*
3. *Data Presentation.* Visualization of innovation data warehouse composition, search patterns and results of the latter, objects innovativeness indicators, sets of associated objects.
4. *Services.* Software interfaces for interaction with DSS and expert systems, as well as presentation layer components.
5. *Search Agents.* Innovation solutions information retrieval.
   (a) *Business Processes.* Evolutional generation of the linguistic model of the object archetype and effective multiset of queries. Raw data measurements processing to compute the probabilistic values of the objects innovativeness indicators. Processing of measurements obtained from several sources. Applying the Dempster-Shafer Theory of Evidence. Arranging intelligent search agents interactions.

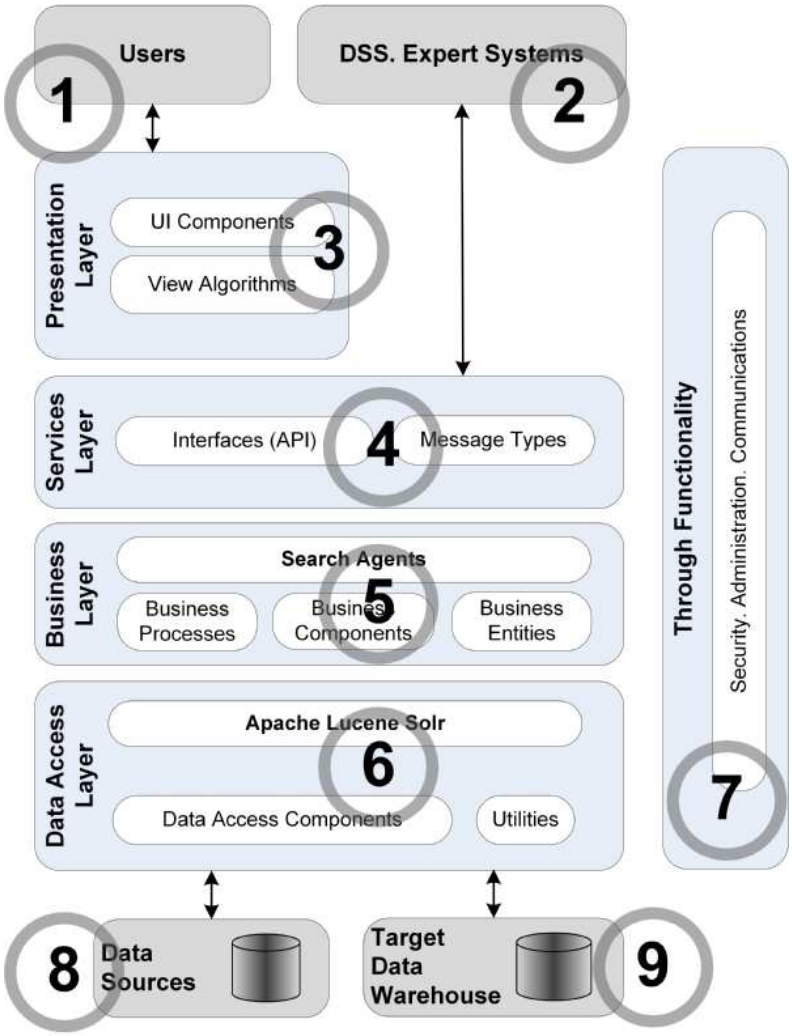

Fig. 1. The general architecture of the data warehouse support system

(b) *Business Components.* A genetic algorithm to produce the effective multiset of queries. An algorithm to filtrate the search results. A model to calculate the objects innovativeness indicators. An algorithm for a group processing of the object innovativeness level measurements.

(c) *Business Entities.* A concepts basis. The linguistic model of the sought-for object archetype. Search pattern. Objects innovativeness indicators. Fuzzy indicators of the probability of specified innovativeness properties. Limitations on the reference information-model in the specified domain.

6. *Apache Lucene Solr* (http://lucene.apache.org). A software implementation of the model of the vector space of the documents: an object model, a software library for document data warehouse access, data indexer, data storage.
7. *Through Functionality.* Security. Administration. Network communications.
8. *Data Sources.* Internet resources. Specialized data warehouses and databases.
9. *Target Data Warehouse.* A register of innovative solutions.

4 Objects Innovativeness Indicators

We introduced concepts of technological novelty, relevance and implementability as the components of the object innovativeness criterion. *The novelty* means significant improvements, a new way of object usage or presentation (novelty subjects are potential users or a producer oneself). *The relevance* is a potential producer's recognized need for an object formed as the demand. And *the implementability* determines the technological validity, physical feasibility and integrability of an object into the system in order to obtain the effects desired.

The linguistic model of the sought-for-object archetype has been proposed. The model terms are classified as key properties describing the object structure, application conditions or functional results. A marker determines the archetype definition domain. The queries are constructed as terms-and-marker combinations. The genetic algorithm of queries generation and results filtration is used to obtain a quasi-optimal queries set.

The following expressions to compute the values of the innovativeness indicators have been proposed:

$$Nov = 1 - \frac{1}{S}\sum_{k=1}^{S}[1 - exp(1 - \frac{R_k}{R_{min}})] \tag{1}$$

where Nov is the object novelty (the value is normalized for range [0;1]), S is the total number of executed queries; R_k is the number of documents found in the database as a result of the k-th query; R_{min} is the minimum number of the documents found among all queries.

$$Rel = \frac{1}{S}\sum_{k=1}^{S}[1 - exp(1 - \frac{F_k}{F_{min}})] \tag{2}$$

where Rel is the object relevance (the value is normalized for range [0;1]), S is the total number of executed queries; F_k is the frequency of users' executed queries similar to the k-th query; F_{min} is the minimum frequency of the query execution among all query.

A hypothesis of the adequate representation of real processes in the information space. The object novelty evaluation is based on the normalized integral evaluation of the number of results of the object information search in the heterogeneous databases. It is suggested that the number of the search results which are relevant to the search pattern would be less for new objects rather than for

long-time existing and well-known ones. The object relevance estimation is based on the normalized integral estimation of the frequency of users' executed queries similar to the queries generated from the search pattern. Taking into account the direct quantitative evaluation of the innovativeness, we suppose that this approach can be complementary to the conventional ones [5] and [6].

5 Applying the Theory of Evidence

Since an obviously incomplete and inaccurate object information is expected to be obtained from different sources, fuzzy indicators for the probability of the technological novelty and relevance for the object have been introduced. To calculate the above probabilities, application of the Dempster-Shafer Theory of Evidence is validated [7], [8] and [9]. So, the basic probability m of the fact that the object innovativeness indicator (Nov or Rel) measurements belong to the interval A can be evaluated from the following:

$$m : P(\Omega) \to [0,1], m(\emptyset) = 0, \sum_{A \in P(\Omega)} m(A) = 1 \tag{3}$$

where Ω is the indicator measurements set, $P(\Omega)$ is the set of all Ω subsets. Further, the belief

$$Bel = \sum_{A_k : A_k \subseteq A} m(A_k) \tag{4}$$

and plausibility

$$Pl = \sum_{A_k : A_k \cap A \neq \emptyset} m(A_k) \tag{5}$$

for specified k intervals are calculated. These functions determine the upper and lower boundaries for the probability that the object has the property specified. This is the way to estimate the values of indicators Nov or Rel in the conditions of the incomplete and inaccurate information about the objects. We also studied the applicability of Theory of Evidence for solving tasks in complex technical systems diagnostics and optimum control over the evolution of multistage processes in the fuzzy dynamic medium [3] and [10]. The objective of these investigations is to suggest a new architecture (for interactions among the intelligent search agents in the system of the heterogeneous data warehouses) based on the concept of an "abnormal" agent. The abnormal state (AS) of a search agent can be interpreted as the presence of a challenger for innovation as a result of the search by this agent. AS can be diagnosed as an exit of the objects innovativeness indicators beyond their characteristical values. The following necessary and sufficient condition can be used to indicate the search agent AS, $s \in S$:

$$(\forall s \in S)(F = 0) \leftrightarrow P^* \neq 0 \tag{6}$$

where F is the indicator function, P^* is the set of registered AS.

Taking into account the potentially great number of information sources, testing diagnostic hypotheses provides enhanced efforts to avoid missing a true

innovation and indicating a false one. The indicator function allows the quantitative evaluation how "reasonable" or "useful" is the search agent.

6 Evolutional Approach

The evolutional approach is validated for and applied to solving the task of forming the linguistic model of the object archetype. The main idea is to use a special genetic algorithm (GA) to arrange an evolution process generating the stable and effective set of queries to obtain the most relevant results. The basic concepts of the approach used in the development of the GA proposed by the authors are described in [11]. Thus, the initial interpretations are as follows: *a query* is *an individual*, *an encoded set of query terms* represents *a genotype*, the replacement of a query term with another term is defined as *crossover*, and the replacement of a query term with its synonym is *mutation*. The procedure of a fitness function calculation consists in executing a query by a search engine and getting the set of relevant documents found – *a phenotype*.

The GA search pattern K is a set of terms related to a certain subject area. Each search query is represented by a vector $\overline{q} = (c_1, c_2, \dots c_n, \dots c_m)$, where $c_n = \{k_n, w_n, z_n\}$, $k_n \in K$ is a term, w_n is a term weight, z_n is a set of term synonyms k_n, m is the number of terms in a query. The result of the query is a set of documents R, $|R| = D$. The initial population of S queries is a set Q_0, where $|Q_0| = S$, $S < |K|/2$, $q \in Q_0$. The fitness function for the query population is calculated as follows:

$$\overline{W}(Q) = \frac{1}{S} \sum_{j=1}^{S} \frac{1}{R} \sum_{i=1}^{R} w_{ji}(g, f, s) \tag{7}$$

where $Q = (q_1, q_2, \dots q_S)$ is the population of S queries; w_{ji} is the fitness function of i-th results of j-th query. Here, w_i depends on position g in the search engine results list, frequency f of this search result in all S query result lists, similarity measure s of the short result text and search pattern K.

To understand how GA works, Holland's Schema Theorem plays a key role. It was stated for the canonical GA and proves its convergence. It is obviously reasonable to check if the theorem of schemes works for any modifications of the canonical GA. Our investigations specify conditions for the correct check of the theorem of schemes. So, a new encoding method (geometric coding) has been proposed. To code individuals, we suggest using distance $Dist(\overline{q}_i, \overline{q}_0)$ between vector $\overline{q}_i$ and initial vector $\overline{q}_0$. In the case of a cosine measure we have:

$$Dist(\overline{q}_i, \overline{q}_0) = \frac{\overline{q}_i * \overline{q}_0}{\| \overline{q}_i \| \cdot \| \overline{q}_0 \|} \tag{8}$$

Encoding method applicability criterion based on the uniform continuity of the fitness function has been suggested too. Fitness function $\overline{w}(q_j)$ is called uniformly continuous on the set Q, if $\forall \epsilon > 0$ $\exists \lambda > 0$, such that $\forall q'$, $q'' \in Q$ satisfying the condition $|q'' - q'| < \lambda$, the inequality $|\overline{w}(q'') - \overline{w}(q')| < \epsilon$ is valid. It means

that small changes of individual code q_j lead to small changes of fitness function $\overline{w}(q_j)$. Also, it means that the value of λ limiting the deviation of individual code q_j only depends on the value ϵ of the deviation of fitness function $\overline{w}(q_j)$ and does not depend on the value of individual code q_j, i.e. it is constant on the whole domain of the function.

7 Method for Objects Innovativeness Calculation

In the framework of the project, a version of a method for raw indicators measurements processing has been processed to calculate the probabilistic values of the objects innovativeness indicators. The main steps of the method are as follows:

1. Execute the specified number of quasi-optimal queries generated by the GA from the search pattern. From the viewpoint of the Theory of Evidence, such queries are observed subsets or focal elements. For all retrieved documents R the number of group intervals is determined as $I = S^{1/2}$. In terms of measuring Nov the mentioned intervals correspond to the nominal scale "It is novel", "It is evidently novel", "It is evidently not novel", "It is not novel".
2. Compute the basic probability $m(A_k)$ of the appearance of innovativeness indicators according to (3) for each of the subsets observed. Note, $m(A_k)$ can be estimated as follows:

$$m(A_k) = q_k/S, \quad \sum q_k = S \tag{9}$$

   where q_k is a number of observed subsets (queries).
3. Compute the belief Bel and plausibility Pl for each A_k according to (4) and (5).
4. Processing the measurement results retrieved from different search engines. The combined base probability m_{12} for two search engines:

$$m_{12}(A_k) = \frac{1}{1-K} \sum_{A_i^{(1)} \cap A_j^{(2)} = A} m_1(A_i^{(1)}) m_2(A_j^{(2)}) \tag{10}$$

$$K = \sum_{A_i^{(1)} \cap A_j^{(2)} =} m_1(A_i^{(1)}) m_2(A_j^{(2)}) \tag{11}$$

   where K – the conflict factor.
5. Evaluating source credibility. It can be considered with the introduction of discount factor α for base probability $m(A)$. Discounted base probabilities are estimated as follows:

$$m^{\alpha}(A) = (1-\alpha) m(A) \tag{12}$$

An own algorithm for the group processing of the objects innovativeness level measurements has been developed.The combining is executed recursively, from couples of sources: two evidence sources form a single conditional one, the evidences of which are combined with the next actual source.

8 Data Warehouse Support System Functioning

Fig. 2 shows the chart of UML sequences that describes the general functioning of the Data Storage Support System. The figures correspond to the system components described in Section 3 above. It shows the sequence of communications among the interacting objects (components and actors). Note two important activity periods: data presentation (component 3) and search agents functioning (component 5) including obtaining variants of an innovation solution and its associated objects. The intermediate steps reflect the algorithmic aspects of the interactions among the system components.

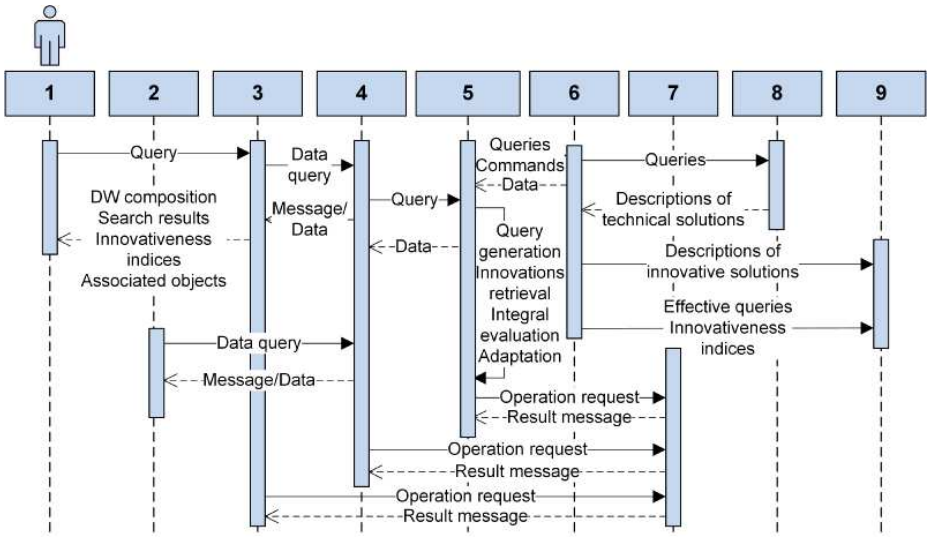

Fig. 2. Data warehouse support system components functioning

9 Experimental Investigations

To experimentally check if the computational model developed is adequate, we stated the following tasks:

1. Approve the computation procedure for objects innovativeness indicators.
2. Compare the computed values of the innovativeness indicators to the expert-estimated ones.
3. Compare the computed values of the innovativeness indicators obtained after data processing from different search engines.
4. Evaluate the dynamics of the object innovativeness indicators in time.
5. Validate the feasibility of the measured innovativeness indicators for further processing.

The following search engines were selected as objects information sources: http://new.fips.ru, https://elibrary.ru, https://rosrid.ru, https://yandex.ru, https://wordstat.yandex.ru, https://google.com, https://adwords.google.com, https://patents.google.com, https://scholar.google.ru.

The objects to analyze were ten top inventions made in 2017 and selected by Rospatent (Russian patent authority) experts, and ten random inventions made in 2017. The search patterns were prepared by experts. The document bodies to analyze were formed.

Our experiments proved the validity of the methods of the Theory of Evidence for processing of the measured innovativeness indicators. Despite anticipated differences in the absolute values of the measurements from different data sources, our model can adequately evaluate relative changes in the values of the object novelty and relevancy indicators (combined values of the indicators show similar results). The general conclusion is: the average novelties of the objects estimated as the best objects by the experts are greater than the average novelties of the random objects.

Our experiments involved an analysis of the objects novelty evaluated for twenty years. Fig. 3 give as examples Optic Nerve Electric Stimulator object archetype novelty and relevancy plots (it is specified by the corresponding linguistic model).

Approximation of the values obtained (solid trend lines) validates the hypothesis that the object novelty lowers in time. However, the object relevancy increase in time. It is clear that the object becomes more and more popular among the users, so it enjoys a growing potential interest to it.

Our experiments validated the well-known cyclic regularities (revealed in the analysis of correlation between the innovations and economic growth). In a quite long interval, the values of the innovativeness indicators show cyclic changes (dotted trend lines on the plots). Though the computation results as a whole are ambiguous, we do identify the cycles which require testing the hypothesis of the innovation cycles in the particular usage domain.

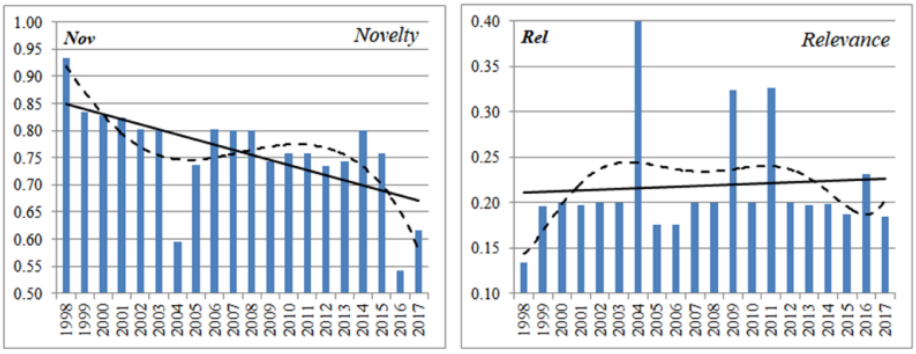

Fig. 3. Object innovativeness indicators behavior (example)

10 Conclusion

The works on the project directions discussed herein are finished. Further we are planning to carry out investigations on the following:

- Formalize the description of the linguistic model of the object archetype, including aiming the search pattern at the innovativeness of the sought-for objects and specifying limitations on the reference information model.
- Develop a behavior model for the intelligent search agent working with a data source in a multiagent system with the heterogeneous data warehouses.

This work was done at the Tver State Technical University with supporting of the Russian Foundation of Basic Research (projects No. 18-07-00358 and No. 17-07-01339) and at the Joint Supercomputer Center of the Russian Academy of Sciences – Branch of NIISI RAS within the framework of the State assignment (research topic 065-2019-0014).

References

1. Ivanov, V. K.: Computational Model to Quantify Object Innovativeness. CEUR Workshop Proceedings 2258, pp. 249-258 (2019). http://ceur-ws.org/Vol-2258/paper31.pdf
2. Palyukh, B. V., Vinogradov, G. P., Egereva, I. A.: Managing the evolution of a chemical engineering system. Theoretical Foundations of Chemical Engineering, vol. 48, issue 3, pp. 325-331 (2014).
3. Palyukh, B. V., Vetrov, A. N., Egereva, I. A.: Architecture of an intelligent optimal control system for multi-stage processes evolution in a fuzzy dynamic environment. Software & Systems. vol. 4, pp. 619-624 (2017).
4. Microsoft Application Architecture Guide, 2nd Edition, October 2009, 529 p. (2009). www.microsoft.com/architectureguide)
5. Tucker, R. B.: Driving growth through innovation: how leading firms are transforming their futures. 2nd ed. Berrett-Koehler Publishers, San Francisco (2008).
6. Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data. The Measurement of Scientific and Technological Activities, 3rd Edition (2005). https://www.oecd-ilibrary.org/science-and-technology/oslo-manual/9789264013100-en
7. Shafer, G.: A Mathematical Theory of Evidence. Princeton, N.J.: Princeton University Press (1976).
8. Yager, R., Liping, Liu.: Classic Works of the Dempster-Shafer Theory of Belief Functions, London: Springer (2010)
9. Ivanov, V. K., Vinigradova, N. V., Palyukh, B. V., Sotnikov, A. N.: Modern Directions of Development and Application Areas of Dempster-Schafer Theory (review). Artificial intelligence and decision making, vol. 4, pp. 2-42, Moscow (2018).
10. Palyukh, B., Ivanov, V., Sotnikov, A.: Evidence theory for complex engineering system analyses. In: 3rd International Scientific Conference on Intelligent Information Technologies for Industry, IITI 2018. Advances in Intelligent Systems and Computing, vol. 874, pp. 70-79 (2019).
11. Ivanov, V. K., Palyukh, B. V., Sotnikov, A. N.: Efficiency of genetic algorithm for subject search queries. Lobachevskii Journal of Mathematics, vol. 37, no. 3, pp. 244–254 (2016). http://dx.doi.org/10.1134/S1995080216030124