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

Boris Palyukh Tver State Technical University, Tver, Russia
pboris@tstu.tver.ru

Vladimir Ivanov Tver State Technical University, Tver, Russia
mtivk@tstu.tver.ru

Alexander Sotnikov Joint Supercomputer Centre of the RAS, Moscow, Russia,
asotnikov@jscc.ru

4rd International Scientific Conference
“Intelligent Information Technologies for Industry”
December 2-7, 2019, Ostrava-Prague, Czech Republic

Project Overview

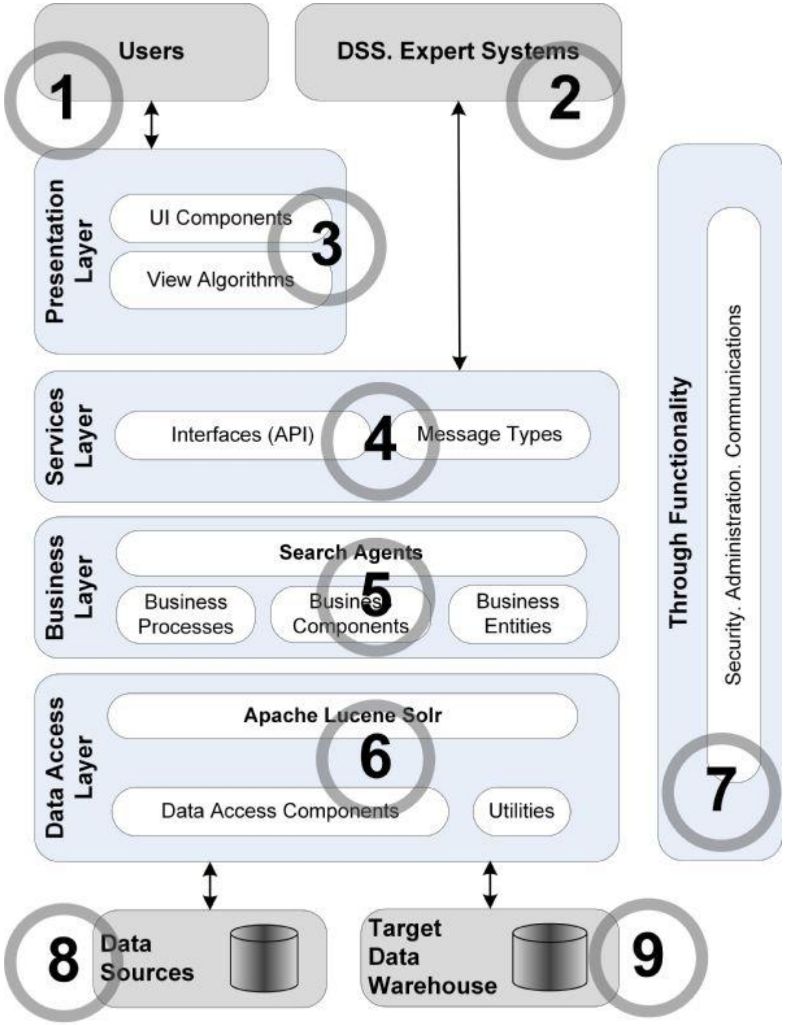
Data Warehouse Support on the Base Intellectual Web Crawler and Evolutionary Model for Target Information Selection

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)

The project results are supposed to support solving the following applied tasks:

- determine the characteristics of new domains in business planning;
- forecast the technological development of the business;
- provide an information support for expert groups and individual experts.

Data Warehouse Support System Architecture



The basic components (circles):

1. User’s Applications.
2. DSS. Expert Systems.
3. Data Presentation.4. Services.
5. Search Agents.
    (a) Business Processes
    (b) Business Components.
    (c) Business Entities.
6. Apache Lucene Solr.
7. Through Functionality.
8. Data Sources.
9. Target Data Warehouse.

Objects Innovativeness Indicators (1)

Innovation is an object (such as an invention, an engineering system component, a technology, a method, etc.) having some properties determining its technological novelty, relevance, and implementability.

The following **innovativeness indicators** have been proposed:

- ***Nov*** is the object novelty.

The object means significant improvements, a new way of using or granting an object or a technology. The object novelty evaluation is based on the normalized integral evaluation of the number of the object information search in the heterogeneous databases.

- ***Rel*** is the object relevance.

The object relevance is a potential producers awareness of the object necessity as a demand. The object relevance estimation is based on the normalized integral estimation of the frequency of users' executed queries.

- ***Imp*** is the object implementability.

The object implementability estimation is based on the normalized estimation of the average restoration period of the object novelty and/or relevance.

Objects Innovativeness Indicators (2)

Object novelty *Nov* is determined as follows:

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

Nov is standardized for [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.

Object relevance is estimated as follows:

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

Rel is standardized for [0;1]; 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.

Object implementability is estimated as follows:

$$Imp = 1 - \frac{1}{2}(\overline{LN}_{01}^{\max}(Nov(t)) + \overline{LR}_{01}^{\max}(Rel(t)))$$

$\overline{LN}_{01}^{\max}$ and $\overline{LR}_{01}^{\max}$ are average distances between two consecutive time series points t_i, t_i+1 ∈ [t_0;t_m] of the local maxima of the functions *Nov*(*t*) and *Rel*(*t*), respectively.

Applying the Theory of Evidence (1)

$$\boldsymbol{Bel(A_i) = \sum_{A_j \subseteq A_i} m(A_j)}$$

A belief function shows the degree of confidence in that fact that objects have the given property

A_i is an event proving an object (object sets) has the given property); $A_i \subseteq C$; C is an exhaustive event set (a problem solution result);
$m(A_j)$ is a basic distribution of event probabilities A_j; $m\left(A_j\right) \in [0,1]$.
$Bel(\emptyset) = 0$, $Bel(A_i) \in [0,1]$ и $Bel(C) = 1$.

$$\boldsymbol{Pl(A_i) = 1 - Bel(\overline{A_i}) = 1 - \sum_{A_i \cap A_j = \emptyset} m\left(A_j\right)}$$

A plausibility function shows the degree of plausibility in that fact that objects have the given property

$$\boldsymbol{Bel(A_i) \leq p(C_i) \leq Pl(A_i)}$$

$p(A_i)$ is true probability of the given property of the objects from $A_i \subseteq C$.
$Bel(A_i)$ and $Pl(A_i)$ are lower and upper limits of $p(A_i)$.

This is the way to estimate the values of indicators *Nov*, *Rel* and *Imp* in the conditions of the incomplete and inaccurate information about the objects:

p(Nov) − the probability of an object having technological novelty.
p(Rel) − the probability of an object having relevance (in demand).
p(Imp) − the probability of an object having implementability.

Applying the Theory of Evidence (2)

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:

- The objective is to suggest a new architecture for interactions among the intelligent search agents in the system of the heterogeneous data warehouses.
- It is based on the concept of an "abnormal" agent.
- The abnormal state (AS) of a search agent is 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 AS:

$$(\forall s \in S)(F = 0) \leftrightarrow |P^*| \neq 0$$

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.

Evolutional Approach (1)

- The evolutional approach applied to 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.
- Thus, the initial interpretations are as follows: ***a query is an individual***, ***an encoded set of query terms is 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 afitness 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 a vector $q = (c_1, c_2, \dots c_n, \dots c_m)$, 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:

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

$Q = (q_1, q_2, \dots q_S)$ is the population of S queries; w_{ji} is the fitness function of i-th results of j-th query; 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.

Evolutional Approach (2)

1. Our investigations specify conditions for the correct check of the theorem of schemes known as Holland's Schema Theorem. To code individuals, we suggest a new encoding method (geometric coding) using distance $Dist(q_i; q_0)$ between vector q_i and initial vector q_0. In the case of a cosine measure we have:

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

2. Encoding method applicability criterion based on the uniform continuity of the fitness function has been suggested too.

Fitness function $w(q_j)$ is called uniformly continuous on the set Q, if $\forall\ \varepsilon > 0\ \exists\ \lambda > 0$, such that $\forall q', q'' \in Q$ satisfying the condition $|q'' - q'| < \lambda$, the inequality $|w(q'') - w(q')| < \varepsilon$ is valid.

It means that small changes of individual code q_j lead to small changes of fitness function $w(q_j)$. 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 $w(q_j)$ and does not depend on the value of individual code q_j, i.e. it is constant on the whole domain of the function.

Method for Objects Innovativeness Calculation

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.

2. For all retrieved documents the number of group intervals is determined. 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*".

3. Compute the basic probability m(A_k) of the appearance of innovativeness indicators:

$m(A_k) = q_k/S, \quad \sum q_k = S$ where qk is a number of observed subsets (queries).

4. Compute the belief *Bel* and plausibility *Pl* for each A_k.

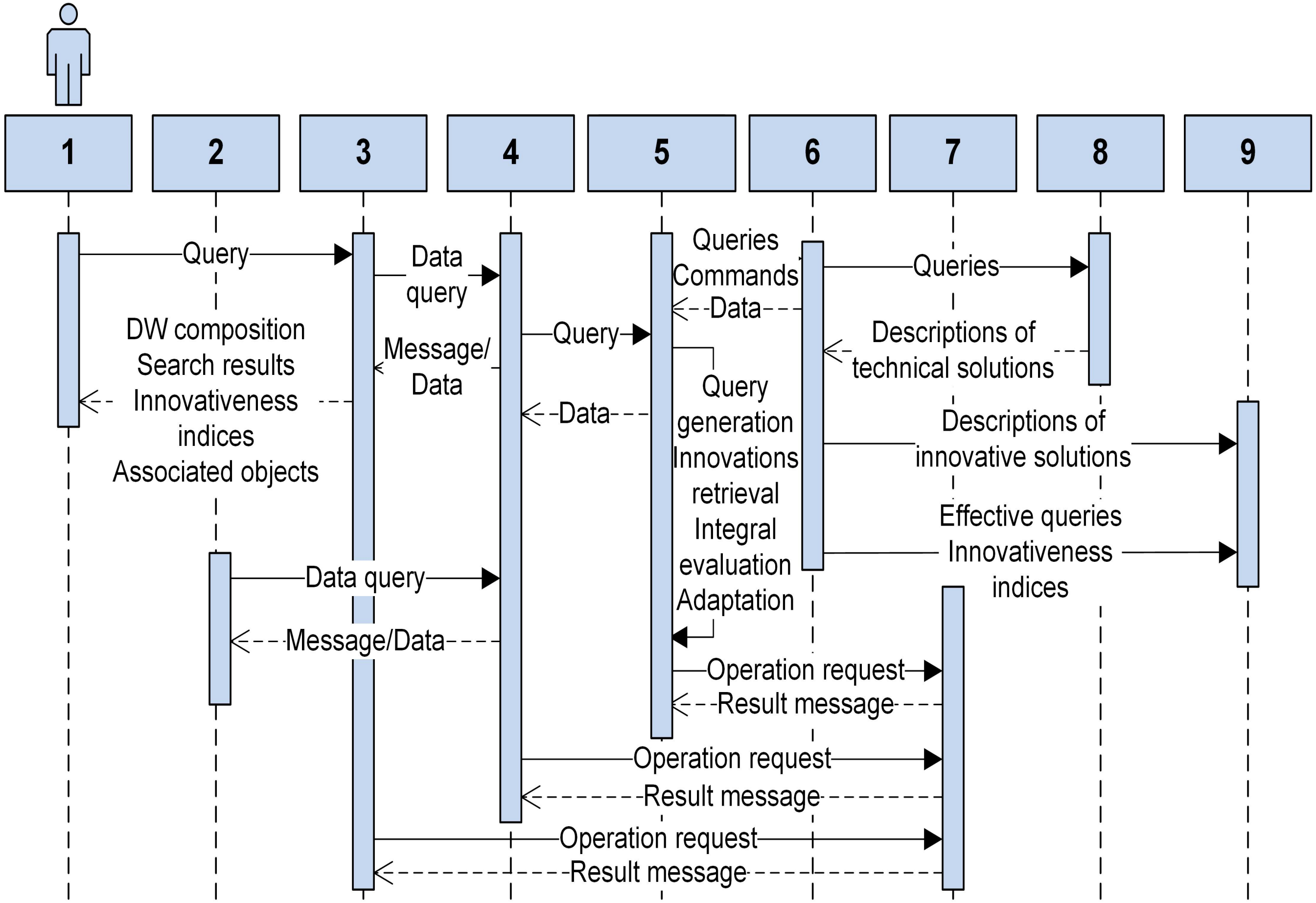
5. Processing the measurement results retrieved from different search engines. The combined base probability m_{12} for two search engines (*K* – the conflict factor):

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

6. 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: $m^{\alpha}(A) = (1-\alpha)m(A)$.

7. 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.

Data Warehouse Support System Functioning



1. User's Applications.
2. DSS. Expert Systems.
3. Data Presentation.4. Services.
5. Search Agents.
   (a) Business Processes
   (b) Business Components.
   (c) Business Entities.
6. Apache Lucene Solr.
7. Through Functionality.
8. Data Sources.
9. Target Data Warehouse.

Note two important activity periods: data presentation (3) and search agents functioning (5) including obtaining variants of an innovation solution and its associated objects.

Experimental Investigations (1)

1. The goals

- Approve the computation procedure for objects innovativeness indicators (OII).
- Compare the computed values of the OII to the expert estimated ones.
- Compare the computed values of the OII obtained from different search engines data.
- Evaluate the dynamics of the object innovativeness indicators in time.
- Validate the feasibility of the measured innovativeness indicators for further processing.

2. 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*.

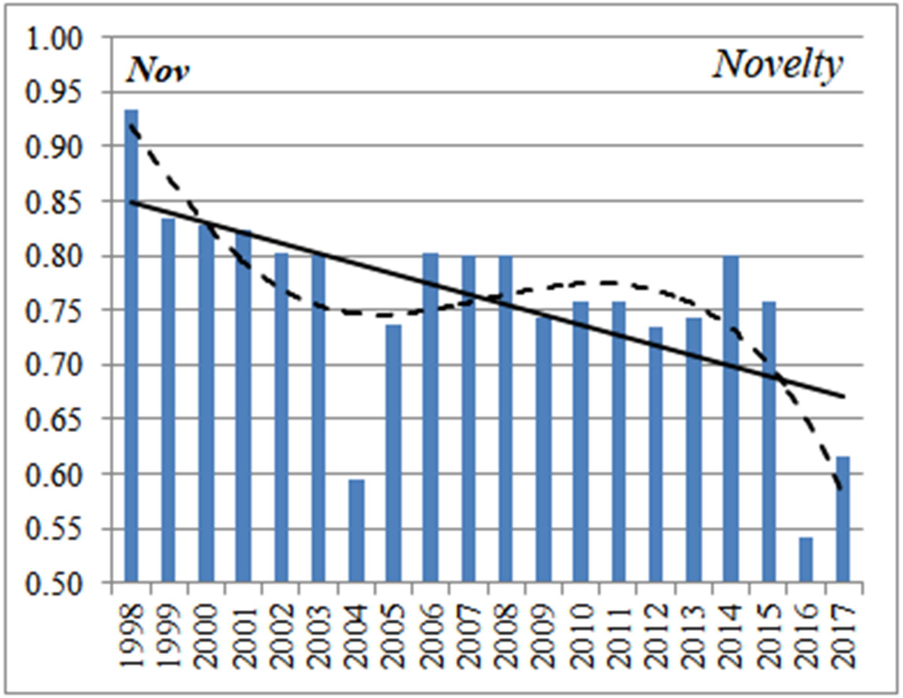
3. The objects to analyze were ten top inventions made in 2017 and selected by Rospatent (Russian patent authority) experts, and ten random inventions made in 2017.

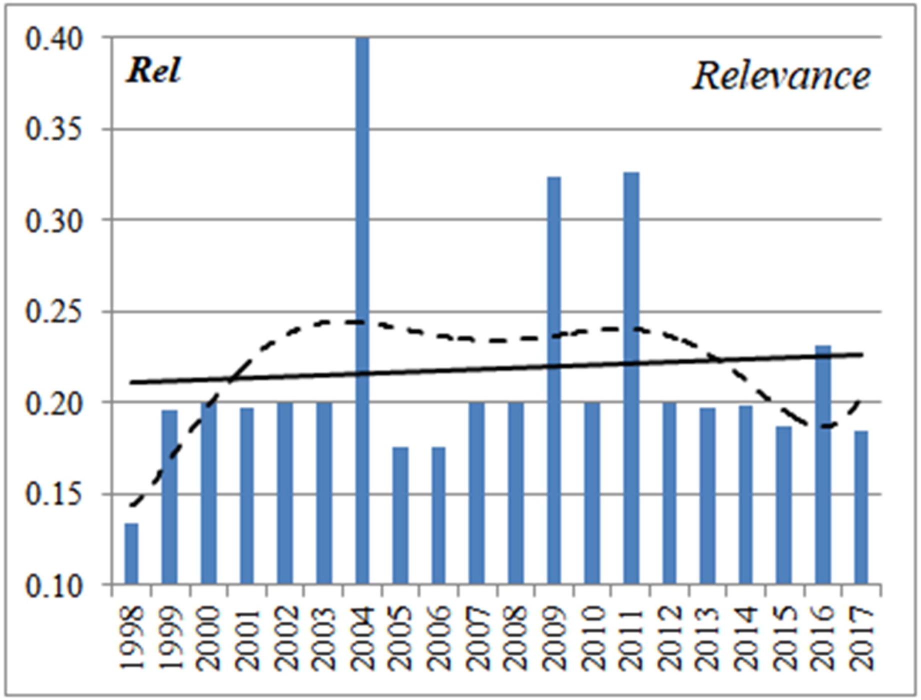
4. Our experiments proved the validity of the methods of the Theory of Evidence for processing of the measured OII. Our model can adequately evaluate relative changes in the values of the object novelty and relevancy indicators (combined values of the indicators show similar results).

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.

Experimental Investigations (2)

1. Optic Nerve Electric Stimulator objectarchetype novelty and relevancy plots.
2. An analysis of the objects novelty evaluated for twenty years.
3. The object novelty lowers in time.
4. The object relevancy increase in time. The object becomes more popular among the users.
5. In a quite long interval, the values of the innovativeness indicators show cyclic changes. We do identify the cycles which require testing the hypothesis of the innovation cycles in the particular usage domain.





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.

Thank you.