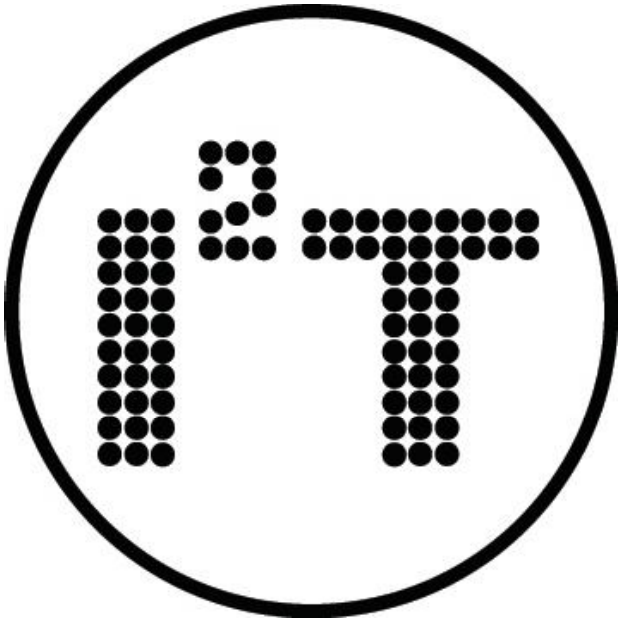
International Scientific – Practical Conference «INNOVATIVE INFORMATION TECHNOLOGIES»



PART 2
INNOVATIVE INFORMATION TECHNOLOGIES IN SCIENCE

Prague – 2014
April 21-25

К 32.97
UDC 681.3; 681.5
I 64

I 64 Innovative Information Technologies: Materials of the International scientific – practical conference. Part 2. /Ed. Uvaysov S. U.–M.: HSE, 2014, 736 p.

ISSN 2303-9728

The materials of The Third International Scientific – Practical Conference is presented below. The Conference reflects the modern state of innovation in education, science, industry and social-economic sphere, from the standpoint of introducing new information technologies.

Digest of Conference materials is presented in 3 parts. It is interesting for a wide range of researchers, teachers, graduate students and professionals in the field of innovation and information technologies.

LBC 32.97
ISSN 2303-9728

12. Konnov N.N., Domnin A.S. Razrabotka modeli algoritma «skol'zjashhego» okna cvetnymi vremennymi setjami Petri / Estestvennye i matematicheskie nauki v sovremennom mire. № 9-10. Sbornik statej po materialam IX-X mezhdunarodnoj nauchno-prakticheskoj konferencii. — Novosibirsk: Izd. «SibAK», 2013 - S 76-81

13. Nikishin K.I., Konnov N.N., Domnin A.L. Modelirovanie trafika seti Ethernet cvetnymi setjami Petri / Sb. materialov I Mezhdunar. nauch.-prakt. konf «Sovremennye problemy komp'juternyh nauk (SPKN-2013)» - Penza : Izd-vo PGU, 2013 - S 120-123

INTELLIGENT SUBJECT SEARCH SUPPORT IN SCIENCE AND EDUCATION

Ivanov V.K, Palyukh B.V., Sotnikov A.N.
Tver, TSTU; Moscow, JSCC RAS

This article presents main results of the pilot study of approaches to the subject information search based on automated semantic processing of mass scientific and technical data. The authors focus on technology of building and qualification of search queries with the following filtering and ranking of search data. Software architecture, specific features of subject search and research results application are considered.

Keywords: genetic algorithm, innovation, data mining, science, education, search query, population, fitness, ranking, relevance, semantic search, filtering, data-warehouse.

Introduction. New efficient scientific knowledge search and synthesis methods (in particular, breakthrough technologies and innovative ideas in economics, science, education) are one of the top research and development targets in the field of information technology. The project Intelligent Distributed Information Management System for Innovations in Science and Education powered by the Russian Foundation of Basic Research (contract No NK13 -07- 00342 \ 13) is to solve this problem. This article presents main results of the pilot study of approaches to the subject information search based on automated semantic processing of mass scientific and technical data.

Specific Features of Subject Search. The major features of subject search tasks which determine the approaches are:

• the required information is often located at the junction of adjacent areas, hence, there is some complexity in the exact wording of the search query.

• along with the information on proper innovation it is desirable to obtain information on applications, risks, specific features, users, authors, producers.

• there is a necessity of available alternatives and different criteria mixing for selecting the most effective practices.

• the information on innovations is fragmentized and heterogeneous; primarily sector-specific character.

In contrast to the search for specific information (facts) on particular aspects of the required content, it is rather difficult to solve a sophisticated problem of searching coordinated information on a target subject. For example, it is required to find the economic performance of mine Raspadskaya JSCo for the first half of 2013. If we use this phrase as a search query, it is possible to get a relevant answer in the first ten search results of Google. But how can one find the information to analyze scientific, technical, economic and social factors affecting the innovative technical, technological, or financial mechanisms of coal-mining in the eastern regions of Russia?

To solve such search problems users have to employ lots of key concept combinations, clarify them in the course of en-route search on the Web or specialized stores such as patent

databases (DB). It is not obvious that for this purpose any reasonable method would be used without fail. Eventually, a large amount of search results would be at the disposal of a user (tens and hundreds of documents), with the found information being more or less relevant to queries. As is quite common, there would be no opportunity to go into details of all the result data. So, the following questions can arise:

• How can one simultaneously assess the relevance of documents found by different queries? Is the relevance of documents determined correctly?

• Is the data ranking in a certain search system correct from the perspective of a user? Do all the results available for direct assessment meet the user's expectations?

• Are all the results that meet the user's expectations available for direct assessment? Are all the required data (e.g. innovative solutions) found at all?

• How one can filter documents extrinsic to the searched subject?

• Is it possible to find any effective solutions relevant in other application fields, but would be successfully used as an innovation in this domain.

• Is it possible to give a visual assessment to lots of found innovative solutions together with linked objects?

There are no clear-cut ways of solving these problems within trivial solutions. Obviously, we need efficient methods of creating and populating the computer-assisted collections of advanced technologies and ideas which would contain not only their descriptions, but selected, classified and associated data. These data can be used to analyze retrospective and prospects of specific innovations, to search current and likely trends. The project in question is an attempt to offer a number of such innovative approaches.

Problem definition. Thus, a project goal can be summarized as: the exploration of new approaches to innovative solution search methods in the database of a data center and its population with Internet data mining results adapted to visual assessment of selected, classified and associated data. We see three key tasks to attain the goal:

• To develop the technology of building and qualification of search queries with the following filtering and ranking of search data.

• To set up methods of cluster analysis to text documents and multimedia objects in order to use them for tagging the links between search results.

• To create a store of innovative solutions for educational and scientific purposes.

Software Architecture. When developing a general software architecture based on mechanisms of direct automated search of innovative solutions the authors determined view layers, those of services, business logic, data access as well as crosscutting concerns (the UML notations and artifacts were applied). In the behavioral model of a system (Fig.1), in a particular session, we can distinguish two periods of user's activation: query formulation (first step) and visualization of the results including the options of the requested and innovative solutions and linked objects (final step). Interim steps are hidden, off-line run and implement the algorithm of interaction between the system components without active participation of a user.

The main functional components:

• Search module. It involves executing a search query in the Internet search systems and the custom directory of innovative solutions; basic search (query by attributes and full-texts), location, data retrieval and summarizing.

• Query qualification module. Selection and ranking of search results: filtration, subject control, qualification of search query.

• Classification module. Classification of search results: selection of a method, cluster analysis of text documents and multimedia objects, data qualification. As a result we obtain a subset of semantically linked data.

• Link identification module. Link start-up: qualitative classification assessment, selection of the best results, interpretation of results; generating the descriptions of solutions with innovative potential in a given subject segment or for a specified object (article, technology, product).

• Visualization module. It involves mapping of search results, procedures of data processing, classification results including semantic links between objects.

• Data-warehouse (DW) management module involves storage and updating of data search and processing results, parameters, and intermediate data; registry of innovative scientific, technological and educational problems. DW is built on the basis of a vector space model, includes document database access libraries and a data indexer.

• Service module. It involves monitoring and analysis of user access to information resources.

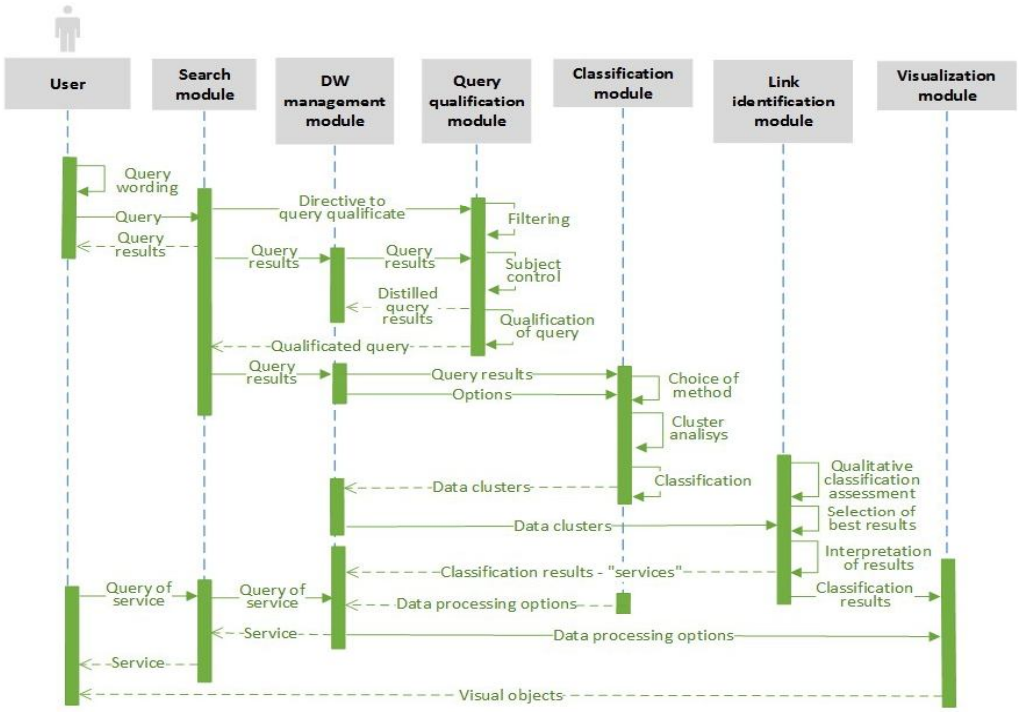


Fig.1. Behaviour of software components

It is particularly remarkable that the developed original object model is oriented to work with any text objects related to the subject of the processing: queries, search results, text documents. Over 30 entity classes specify a document processing environment, a set of documents, methods of calculating the package document similarity measures as well as search functions in document package, types of reports, a collection of document words, lemmatization, a document structure and its specific parts.

The detailed architectural solutions are described in [1].

General Search Algorithm. One of the elements of the presented above architecture is a generalized heuristic algorithm for filtering and rating the search results, which is based on available search engines; the algorithm is supposed to provide a background for search modules and inquiry qualification, as well as for retrieval schedule and search procedures in general.

The algorithm under consideration uses search results of known search engines being in service; it is invariant to them; with various degree of automation; it uses the search engine rating results.

The algorithm instruments a multistep process of sequential filtrations of search results and the analysis of semantic similarity of the found object content to adaptively generated reference texts (*k*-patterns). For an assessment of quality of ranking executed in compliance with algorithm, a modified *DCG* measure was used. The ways of generating effective *k*-patterns were investigated as well.

Let us briefly run through the algorithm operation (Fig.2). The description of a generalized request Q_0 includes the initial set of key concepts of the target document subject.

The generation of the set Q of search queries $q \in Q$, $|Q| = N$ is automated with an adaptive genetic algorithm searching for an effective total pertinence of the resulting document sampling under given evolutionary process depth constraints (see below).

The execution of queries q_{ij} is accompanied with filtering search results R_{qs} rated by a search engine and generating total results R. Filtering provides for the exclusion of some documents which subject area is formally pertinent but should not be the subject of the search for some reasons. It is done by hand or with a classifier which learning set is updated during the analysis of found texts.

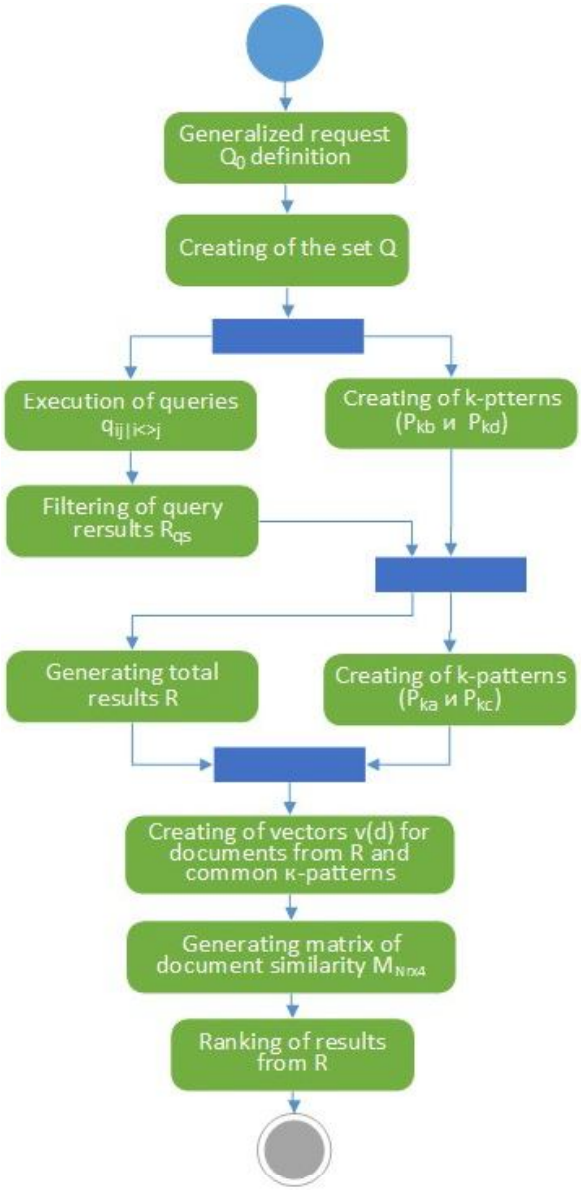


Fig. 2. General pattern of a generalized heuristic algorithm for filtering and rating the search results

The examples of documents being filtered are tutorials, student's papers, training programs, tests and notes, site promotion materials, company's sites, shopping sites, social networking sites; blogs; advertisements; virus-infected resources; nonexistent resources. The generation of k –patterns or reference texts is done simultaneously. They are used for calculating document similarity measures (P_{ka} is a text combination based on the first positions of rated search results, P_{kc} is the most pertinent result, P_{kb} is the text constructed from authority dictionary entries and P_{kd} is a text constructed from Q_0). Further the model of document vector space is used, i.e. each document d (the search results from R and k –patterns) is interpreted as vector $\overline{v}(d) = (w_{1,d}, w_{2,d}, \dots, w_{1,N_r})$, where $w_{t,d}$ is determined with a common metric $tf_{t,d} * idf_{t,d}$. A matrix $M_{N_r x4}$ of document semantic similarity from set R with common k –patterns is generated and the rating of documents from R in accordance

with their similarity $Sim(d_1 d_2)$ to k –patterns is done. The algorithm is described in detail in [2].

Note that the project provides for the usage of Internet search engine work results. The proposed search algorithms will be added to authoritative decisions – classical approaches to search result ranking (HITS, PageRank, BrowseRank, MatrixNet) which are based on the combination of document semantic pertinence and authority as well as user's behaviour and experience.

Generation of Search Queries. The project proposes and investigates the approach to search result generation based on a genetic algorithm. The approach is used to specify a semantic kernel of a document desired set and generate sets of effective queries. The problem definition provides for the organization of an evolutionary process generating a stable and effective query population forming a relative search image of a document. A target set of search results is to be formed by such document addresses which are (a) in the first positions of a ranked list constructed by a search engine; (b) present in the result lists of multiple queries; (c) semantically similar to reference texts generated during evolutionary queries; (d) adequate to the environment given to a crawler by a user profile.

The original population from N search queries may be a set of $Q = \{q_i\}$, $|Q| = N$, $N < |Q_0|/2$, $q_i = (k_1, k_2, \ldots, k_m)$, where $(k_1, k_2, \ldots, k_m)$ is a random combination of key concepts of a search image Q_0. The value of an objective function must determine the query quality (population individual fitness). For each i-th query result the value may be calculated as $w_i(f, p, s, a)$, where f is determined by a result position in a ranked result list made by a search engine; p is determined by entering the result in the result lists of most queries; s is determined by a semantic similarity to k –patterns formed adaptively during the algorithm execution; a is determined by a user profile as an environment factor (values f, p, s, a are normalized for the range from 0 to 1). The value of a target function for each query is calculated as an averaged weight of query results $\overline{w} = \frac{1}{p}\sum_{i=1}^{p} w_i$, where w_i is a weight of each result calculated after executing all queries; P is a number of document addresses seen as the query result. The value of a objective function is interpreted as the capability of a search query to generate the results to be in the next population generation.

To choose parent couples the method of genotype outbreeding is proposed. It can provide for the most complete participation of all current queries in generating the next query population (the first parent individual is chosen randomly and the second individual is the "farthest" from the first one, the distance can be calculated as $\Delta\overline{w} = \overline{w}_1 - \overline{w}_2$). The evolutionary operator of crossover is done with discrete recombination which corresponds to the exchange of key words (genes) between queries. The peculiarity of the proposed implementation is that the key word of a parent query is not substituted for the other parent query key word but its synonym. It allows generating considerably more child queries, with properties (semantics) of parent queries being preserved.

The essence of the most adequate mutation operation of the approach under study is the probabilistic change of a key query word (gene) chosen randomly. The essence of mutation of the approach under study is the change of a key query word (gene) chosen randomly. Since the number of key words in a query $q_i = (k_1, k_2, \ldots, k_m)$ is fixed, it is not possible to use such mutation operators as a new gene addition, new gene insertion, gene deletion. Besides, there is no sense in gene place exchange in the context of executing search queries.

To generate a new population an elite selection denying the loss of best solutions is used. An intermediate population is generated. It includes both the parents and their children. N with the best values of a objective function $\overline{w}$ is chosen from all the population members. They will go in the next population. Generally, the condition of terminating the algorithm is

considered to be population stability. For example, when a mean-square deviation of objective function values (query fitness) reaches some threshold specified by an algorithm parameter. The genetic algorithm is described in detail in [3].

The methods of semantic text comparison are used here. They are the computation of key concept weights and the construction of document vectors and not the known approaches (e.g. shingling) based on detecting direct adoptions in the text.

The research of some approaches to data centre different information systematization should be noted. As a result, a multistep algorithm of alternative search in an information catalogue with a target step number to be a base of a desired solution selection is developed [5, 6].

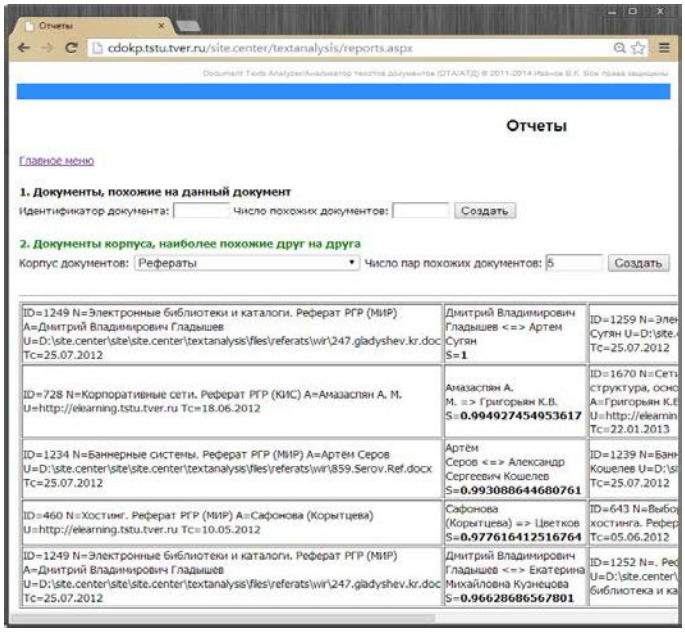


Fig. 3. The appearance of one of the reports of experimental software platform DTA

Data Warehouse. The possibilities of Data Warehouse (DW) generation with realizing a document vector space model to use it as a base of a data-centre information support are researched in the project. A software platform Document Text Analyzer (DTA) for semantic document analysis (their metric similarity computation) is developed within DW.

The prototype of the DW was tested successfully when associated technologies of the integral electronic document quality assessment and document pertinence in different contexts analysis were employed [4]. In particular, the debugging of software shell and interface of the TSTU specialized electronic teaching pack database, data centre warehouse components, was done. The database is used to test and apply the project research results. The pioneering technology of the students' work uniqueness assessment (course and design-graphic papers, semester tasks, reports, essays, tests) is put to use.

Application Areas. The list of application areas of the approaches under discussion in the paper, the research results and technologies is given below:

- A competitive analysis and competitive intelligence. A survey of commercial, scientific and technical, social information sources in a target field. A search of business valuable information. A client information acquisition (in CRM systems). A characterization of new fields and directions in business planning. A search of sector innovation decision descriptions.
- Educational technologies. An analysis of students' paper works (graduation, course papers), theses. A selection and expert examination of teaching materials (books, articles, papers, essays, surveys, etc., including web-resources). Scientometrical analytical services.
- The work of competition committees and sponsoring agencies. An expert examination in venture and other investment funds, the work of councils and groups of experts. An analysis of applications, information cards, competition documentation, expert examination rules and conditions. Normalizing and metrological control of technical

documentation. An analysis of project design documentation, standards, norms, rules, regulations, manuals.

- Patenting, novelty expert examination. Materials selection for patent investigations. A documentation analysis of intellectual property objects, license contracts. Technological development forecasting.
- A content analysis of document texts in sociological surveys.
- Staff recruitment at enterprises and in organisations. An analysis of applicants' resumes vacancy descriptions.
- Rubrication of personal digital documents. PC text document (files) classification and grouping.

It should be noted that the project made some patent research which aim was to find analogs of the system designed and establish its novelty. At the moment of the research result report preparation any data of direct project analogs or its components realized are not discovered. The search of the FGBU Federal Institute of Industrial Property's document database did not show any matches of the project results with technologies recorded in official publications of the titles of protection.

Conclusions. One of the R&D management reference models include a competitive analysis and technological development forecasting based on scientometrical analytical services and semantical systems of business valuable information search. A relatively new world trend is evident: an effective use of global knowledge dataflow. Widely-known solutions (illumin8, NetBase, Orbit) may be examples. With all the differences of these and similar systems the major search pattern is selecting materials on demand, highlighting key concepts in the desired area and grouping materials respectively, filtering and semantic result processing, generating analytical reports. In this sense, the project tasks the results of which were discussed in the article are timely and urgent, and on the appropriate level of the problem interpretation.

References

1. Ivanov, V.K., Palyukh, B.V., Sotnikov, A.N. Arkhitektura intellektual'noy sistemy informatsionnoy podderzhki innovatsiy v nauke i obrazovanii // Programmnyye produkty i sistemy. – Tver', 2013. – № 4. – P. 197-202.

2. Ivanov, V.K., Vinogradova, N.V. Evristicheskiy algoritm fil'tratsii i semanticheskogo ranzhirovaniya rezul'tatov poiska dokumentov // Vestnik Tverskogo gosudarstvennogo universiteta: nauchnyy zhurnal: Seriya "Prikladnaya matematika" / № 41. - Tver. gos. un - t. –Tver', 2013. – № 3. – P. 97-107.

3. Ivanov, V.K. Osnovnyye shagi geneticheskogo algoritma fil'tratsii rezul'tatov tematicheskogo poiska dokumentov: stat'ya // Innovatsii v nauke. – Novosibirsk, 2013. –№ 25. – P. 8-15.

4. Ivanov, V.K., Mironov, V.I. Osobennosti analiza skhodstva dokumentov v razlichnykh kontekstakh zaimstvovaniya pri podgotovke tekstovykh materialov: stat'ya // Otsenka kachestva vysshego professional'nogo obrazovaniya s uchetom trebovaniy FGOS i professional'nykh standartov: materialy dokladov zaoch. nauch.-prakt . konferentsii . – Tver', 2013. – P. 19-26.

5. Palyukh, B.V., Yegereva, I.A. Mnogoshagovaya sistema poiska al'ternativ v informatsionnom kataloge // Programmnyye produkty i sistemy. – Tver', 2013. – № 3. – P. 291-295.

6. Paliukh, B., Egereva, I. Multistep algorithm of alternatives search in an information catalogue // 10th International Conference on Interactive Systems: Problems of Human-Computer Interaction. – Collection of scientific papers. – Ulyanovsk : USTU, 2013. – P. 129-132.