Research of the efficiency of scientific and technical results in the field of chemical safety based on big data analysis

S V Pronichkin1,2,3,4, I B Mamai1,4

1Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Vavilov st. 40, Moscow, Russia;
2Federal Center of Theoretical and Applied Sociology of Russian Academy of Sciences, st. Krzhizhanovsky, 24/35 building 5, Moscow, Russia;
3N.N. Semenov Federal Research Center for Chemical Physics Russian Academy of Sciences, 4 Kosygina Street, Building 1, Moscow, Russia;
4National University of Science and Technology "MISiS", Leninsky Prospect, 4.

Corresponding author’s e-mail: pronichkin@mail.ru

Abstract. The search and extraction of targeted information about promising and breakthrough technologies for ensuring chemical safety is an important element in the analysis of large volumes of unstructured scientific and technical data. Existing approaches to processing large amounts of unstructured data can lead to distortion of the original information. New approaches to the search and extraction of target information based on the typification of the display of visualized large volumes of data of scientific and technical programs are proposed. It is proposed to overcome the disadvantages of existing approaches by using the representation of multi-attribute objects based on the multiset formalism, which allows one to simultaneously take into account all combinations of attribute values, as well as a different number of values for each of them. Multi-feature objects presented as multisets are proposed to be divided into relevant and irrelevant in terms of similarity to the reference multiset based on various metrics. This approach makes it possible to level the features of the initial data and opens up opportunities for solving new problems of studying large volumes of unstructured information of various nature. The results of the computational experiments in the chemical engineering field have shown the effectiveness of the proposed methodological approaches to the search and extraction of target information from large volumes of unstructured data of scientific and technical programs.

1. Introduction
The unevenness of world economic development leads to the redistribution of raw materials and energy resources in favor of more developed countries that are better able to use new knowledge and new technologies for economic growth. The economy of modern industrialized countries is increasingly based on knowledge, which is embodied in the results of scientific and technical activities. Russia's relative lag is caused not only by the economic downturn, but also by the structure of its real sector, which is dominated by the production of raw materials, energy, and an uncompetitive manufacturing industry. Today, the raw material orientation of the Russian economy has almost completely outlived its usefulness, a transition to an innovative development path is necessary.

Innovation is impossible without research and development, sometimes even of a fundamental nature. The positive social effect of innovation can be the basis for budgetary support for the
implementation of relevant R&D. The introduction of the most dynamic and innovative ideas of scientific teams should be provided with effective means of support, which may include tax preferences, loans, subsidies and state guarantees, etc.

Currently, the complex of strategic tasks of the Russian Federation is being implemented through scientific and technical programs (STP). Scientific and technical programs are a set of activities linked in terms of timing and objectives. The end result of the program is the product of the intellectual activity of its participants. It is advisable to make decisions on the further use of the results of their commercialization on the basis of a comprehensive assessment of their absorption potential. There is a need to collect and systematize information on the implementation of program activities, to automate the analysis of big data on their effectiveness. One of such programs is the federal target program “National system of chemical and biological safety of the Russian Federation” (Program). When processing large amounts of data of the Program, various problems arise related to their storage, search, transmission, analysis and visualization.

Methods for searching and extracting target information are an effective tool that allows you to effectively analyze the content of large volumes of semi-structured information without significant investment of time. Searching for and extracting target information involves reducing the amount of information with subsequent typing, which stores the most relevant content [1, 2]. The relevance of solving the problem of reducing the dimension is due, on the one hand, to the rapid growth of information, and on the other hand, to the need for its quick and convenient analysis, in particular, to ensure the possibility of analyzing unstructured descriptions of the content and results of scientific projects of the Program.

In this article, using the example of the Program, a systematic analysis of multi-criteria expert assessments of the absorption potential of scientific and technical results is carried out using methods of verbal analysis of decisions. A systematic analysis of the absorption potential made it possible to identify a number of dysfunctions of the national innovation system in the field of diffusion of the results of state STP. Proposals for their mitigation have been formulated.

2. Existing approaches
Science and technology programs are an important chemical safety tool. The programs create a knowledge-producing "environment" driven by the concentration of research resources. There are many definitions of the concept of "environment". Hyuk and Jing [3, 4] proposed a definition in which the environment is understood as the totality of all objects, the properties of which are essential for a particular subject. Such a subject is an actor - a generalized type of an economic entity performing economically significant actions (acts). By economically significant actions we mean absorption. The concept of "absorption" is used in the general sense of Cohen and Levinthal [5] as the use of new knowledge. The results of STP are considered as objects whose properties are essential for actors. Moreover, it is precisely state programs, since in modern conditions of the national innovation system of Russia, the main subject of research and development funding is the state, and the priority mechanism of such support is the project approach, implemented within the framework of scientific and technical federal targeted programs.

Various methods are used to assess the results of scientific and technical programs. Direct methods are based on the selection of quality indicators of an intellectual property object, measured on the basis of special methods. One of the techniques was proposed in [6], the technique is based on the construction of general identification tables. Then this approach was developed in the works of other authors [7, 8].

A characteristic feature of the existing direct methods of assessment is the use of quantitative equivalents for qualitative characteristics such as the completion of work, its novelty, etc. For such a transition, special scales are used. Then, on the basis of already quantitative indicators, special coefficients are formed, for which cost characteristics are established. Moreover, the monetary value of intellectual capital obtained in this way is quite arbitrary and such information is very difficult to analyze in the case of large volumes of unstructured data.
The analysis of existing approaches to reducing the dimension of large volumes of unstructured scientific data revealed two approaches, namely, extraction and abstraction. When extracting, a subset of phrases or sentences is selected from the original scientific text [9]. With abstraction, a generalized idea of the original scientific information is built on the basis of internal semantics. Then, using the methods of generating natural language constructions, a conclusion is formed, which may contain phrases and sentences that were clearly absent in the original scientific data [10]. Currently, most research focuses on extraction methods, while abstraction methods remain less developed.

Extraction methods are based on a statistical approach, abstraction methods - on a linguistic one [11]. The statistical approach is the easiest to implement, but less effective than the linguistic approach, which allows you to get a result closer to natural language. At the same time, the linguistic approach depends on the quality of the implementation of the generator of natural language constructions and, in contrast to the statistical approach, the result obtained may contain grammatical errors. It was found that the main problem of existing methods of dimensionality reduction is the sparseness of big data.

3. Proposed scientific and methodological approaches

To reduce the dimension of large arrays of semi-structured data, a more effective approach is proposed that takes into account the strengths and weaknesses of the known methods of extraction and abstraction. In existing approaches, multi-character objects are usually represented as a set of tuples. The set of such objects has a rather complex structure, which is rather difficult for analysis. It is more convenient to use another way of representing multi-attribute objects, based on the multiset formalism, which allows one to simultaneously take into account all combinations of attribute values, as well as a different number of values for each of them.

Typification of the display of visualized large amounts of data is carried out by introducing a reference multiset $I$. In the tasks of processing expert opinions on scientific projects of the Program, the elements of the reference multiset can be keywords. For chemical safety newsletters, keywords can be taken from publication titles. Since the title of an article is often an abstraction of the entire article, the abstract of the article, sentences similar to the title, can also become elements of the reference multiset. Fundamental and applied research news feeds are usually deductive. Initial sentences can also be considered as elements of the reference multiset.

Multi-feature objects, presented as multisets, are proposed to be divided into relevant and irrelevant according to their similarity to the reference multiset. To this end, classes of metric spaces of multisets $(A, d)$ are introduced on the family of multisets $A \equiv \{A_1, \ldots, A_n\}$, which are defined by the following metrics:

$$d_{1p}(A, I) = [m(A \Delta I)]^{1/p}; \quad d_{2p}(A, I) = [m(A \Delta I) / m(Z)]^{1/p};$$

$$d_{3p}(A, I) = [m(A \Delta I) / m(A \cup I)]^{1/p}; \quad d_{4p}(A, I) = [m(A \Delta I) / m(A + I)]^{1/p},$$

where $p$ is the degree of the metric, $m$ is the measure of the multiset on the algebra $LZ$, $Z$ is the maximal multiset [12]. The main Hamming-type metric $d_{1p}(A, I)$ characterizes the difference between the two multisets $A$ and $I$. The fully averaged metric $d_{2p}(A, I)$ characterizes the difference between the two multisets $A$ and $I$ depending on the maximum possible distance. The locally averaged metric $d_{3p}(A, I)$ determines the difference between the two multisets $A$ and $I$ associated with the joint "common part" of these two multisets in the original space. The averaged metric $d_{4p}(A, I)$ determines the difference between the two multisets $A$ and $I$, associated with the maximum possible "common part" of these two multisets in the original space. The generalized characteristic of the proximity of each multiset to the reference multiset is calculated based on the specified metrics. Then a threshold value is set, and based on it, multisets that satisfy the constraint are marked as relevant, and the rest are considered irrelevant.

To reduce the dimension of the feature space, the multi-method technology PAKS-M (Sequential Aggregation of Classified Situations by many Methods) is proposed, which operates with multi-feature objects specified by multisets, based on the knowledge of experts and / or the preferences of the decision
maker [13]. The technology provides hierarchical granulation of information by reducing the dimension of the feature space and sequential aggregation of a large number of initial (numeric, symbolic or verbal) data into a small number of composite indicators or a single integral indicator (index) with verbal scales. Gradations of scales of indicators are constructed using different methods of verbal analysis of decisions [14]. The totals or index represent the initial performance in a compact form.

4. Researching the proposed approaches
Computational experiments were carried out to test the proposed approaches on the example of the scientific and technical results of the Program. The typification of the display of visualized large amounts of data of expert opinions on the projects of the Program was carried out using a reference multiset. The elements of the reference multiset were the keywords of the target scientific and technical program. A generalized draft of the program was formed, characterizing promising and breakthrough technologies. Such a project, financed from budgetary funds, should have the following characteristics.
Research should be carried out in the field of rational use of natural resources. The implementation of the results, which the organizer of the project is independently engaged in, has not begun, however, these results may be in demand by production companies, and the recoupment of costs will be up to three years. Based on the processing of expert assessments, the results of the Program were classified according to various criteria of absorption potential, characterizing the contribution of the result to solving the main tasks of the Program, practical feasibility and relevance of the result. The analysis of expert assessments of the Program's results made it possible to highlight strengths and areas for improving their absorption potential.

Strengths of the Program results:
- high relevance - 82.36% of the results require application immediately or in the near future;
- a high level of scientific and technical significance - 79.41% of the results exceed or correspond to the known results;
- a high level of competitiveness - 76.47% of the results exceed or correspond to the known results in terms of their consumer characteristics / properties;
- high level of possible demand - 76.47% of the results are in demand or demand is possible;
- wide scope of use - 97.06% of the results can be applied in many industries and areas or in one industry and direction;
- large scale of application - 88.23% of the results can be applied at the national, regional and / or sectoral level;
- great opportunities for replicating the result - 82.35% of the results allow replication;
- short terms of practical implementation - 88.21% of the results can be implemented within a period not exceeding 5 years;
- experience in mastering - for 85.29% of the results there is a certain experience in mastering the obtained results.

Areas for Improving Program Outcomes:
- a low degree of readiness for mastering - 73.53% of the results require additional work to prepare the result for mastering;
- low level of legal protection - 67.65% of the results do not allow assessing their patentability;
- low readiness of the production base for development - only for 26.47% of the results, the production base is fully ready for their development;
- low readiness of production personnel for development - only for 14.71% of the results, production personnel are fully ready for their development;
- low readiness of the material and technical support / supply system for development - only for 17.65% of the results, the material and technical support / supply system is fully ready for their development;
- low level of funding - only for 11.76% of the results, financial resources for their development are available in full.
The classification of the results of the Program according to the criteria indicates a fairly high level of their relevance and practical feasibility. At the same time, the overwhelming majority of the results do not have a significant, but only a certain impact on the implementation of the Program objectives. There are difficulties with the allocation of financial resources for mastering the results; the readiness of the systems of material and technical support of consumers to master the results is relatively low.

The ranking of the results of the Program was carried out using the methods of verbal analysis of decisions [14], designed for group ordering of multi-feature objects. Rankings of the results of the Program were built for each group of criteria and all criteria in general. In particular, three rankings were constructed for each criterion: ascending, descending, and complex.

The ascending ranking of the results characterizes the closeness of the expert assessments of the results to the highest gradation on the criterion scale. The descending ranking of the results characterizes the remoteness of the expert evaluations of the results from the lowest gradation on the criterion scale. The relative ranking of the results shows the closeness to the highest gradation on the criterion scale relative to the ascending and descending rankings.

In order to neutralize the influence of different criterial characteristics of the results on their rankings, when analyzing the results of the Program, the method of group decision making was used - the Borda procedure, with the help of which different rankings of the results were combined into one final ranking.

Results rankings for all criteria (Pmn - result code) were as follows:
- upward ranking
  P58> P61 = P62> P52> P38 = P55> P5 = P53 = P59> P42> P4 = P26 = P27 = P41> P54> P32 = P50 = P64> P25 = P28 = P43 = P48> P57 = P60> P45,
- top-down ranking
  P58 = P61 = P62 = P38 = P42 = P4 = P32 = P64> P52 = P59 = P50> P55 = P5 = P41 = P25> P53 = P26 = P27 = P28 = P43 = P60> P45> P57 > P48,
- relative ranking
  P58> P61 = P62> P52> P38 = P55> P42 = P59 = P5 = P53> P4 = P26 = P27 = P50 = P25> P54> P32 = P64 > P57 > P48.

The final ranking of the results of the Program for all criteria:
  P58> P61 = P62> P52> P38> P55> P42> P4> P5> P53> P32 = P64> P41> P50> P26 = P27> P25> P54> P28> P43> P60 = P48> P57 > P45.

The analysis of the ordering of results shows that the ranking of results in terms of relative proximity to the highest gradation of criterial assessments and the final ranking, which combines the three rankings, almost completely coincide. This fact indicates good consistency of the rankings constructed.

The most significant results of the Program identified in this way were subjected to subsequent analysis in order to more in-depth assessment of their absorption potential. The experts made proposals to increase the investment attractiveness, and to ensure the availability of the best results of the Program for the mass consumer. The experts also noted not only the possibility, but also the need to commercialize the results. We noted that there are good prospects for the commercialization of the results, including with access to the international market and other production areas. According to experts, the payback period for the results will be up to 3 years. The performers have the necessary knowledge and experience to commercialize the results. We noted that there are good prospects for the commercialization of the results, including with access to the international market and other production areas. According to experts, the payback period for the results will be up to 3 years. The performers have the necessary knowledge and experience to commercialize the results. The experts also noted that the implementation and practical implementation of the results requires significant material costs, pointed out organizational difficulties associated with the readiness of the end consumer to master the results.

The noted areas for improvement reflect the general dysfunctions of the national innovation system of Russia (NIS) in the field of diffusion of the results of state scientific and technological progress.

A sufficient number of works [15, 16] are devoted to the performance of the NIS of its functions, in which dysfunction is understood as the impossibility of the NIS to perform its functions.

By dysfunction of the NIS, we mean the deviation from the optimal level of functioning of a particular NIS function, which shows that there is an objectively established socio-economic contradiction. We will call deviations from the optimal level of dissemination and economic application of the results of STP as dysfunction of the STP in the field of diffusion of STP results.
Conceptually, NIS dysfunctions are largely related to the practical impossibility of adequate accounting by market means of external effects, costs and benefits from the implementation of the results of scientific and technological progress. The limitations of the existing market mechanism for the implementation of the results of scientific and technical progress do not allow solving many socio-economic problems, to the solution of which the measures of the programs are aimed.

The most important reason for this is asymmetric information. NIS actors, who are end users of R&D, may not be aware of the risks associated with their further development and subsequent implementation in production, and the executors of STP projects have little interest in informing them about these risks. In modern conditions of the implementation of measures of state scientific and technical progress, information essential for the implementation is not available to all interested parties and is at the predominant disposal of performers and experts who accept the results of scientific and technical progress within the framework of government contracts. This situation creates prerequisites for increasing the number of commercially effective results of scientific and technological progress that will not be mastered by the actors of the NIS.

For the correct functioning of the NIS in the field of diffusion of the results of scientific and technical progress, their effective support by the relevant legal institutions, in particular, the system of property rights, is necessary. Expert assessments characterizing the level of legal protection of the results of the Program activities are not high. At the same time, the intellectual property right to the results of work obtained within the framework of scientific and technical progress belongs to the customer of the work - the state. This situation leads to low motivation of performers to work in the field of subsequent patenting and implementation of the results obtained.

A significant cause of NIS dysfunction is the specificity of taking into account the time factor. The short-sightedness of the NIS actors is manifested in their orientation towards obtaining quick results, profits while underestimating long-term investments in research and development. This is reflected, in particular, when using the quantitative cost-effectiveness method [17]. Most NIS actors usually assume an investment period of 3-5 years. The scientific and technical results considered in this work were created during the entire period of implementation of the Program activities, which was 5 years. According to experts, 88.21% of the results obtained can be implemented within a period not exceeding 5 years. In this case, most of the results are “ineffective” in terms of traditional market approaches. It is important to note that the majority of government programs are precisely focused on obtaining such “ineffective” results in the social sphere, in the field of environmental protection, national defense and, in general, the traditional spheres of state responsibility.

The low level of actors' competence is also the cause of NIS dysfunction. The experts noted the low readiness of the production base, personnel and the logistics / supply system to master the results of the Program. In paper [18], the main reason for this dysfunction is the lack of a sufficiently qualified management that allows you to quickly master advanced technologies.

The experts noted the low level of funding - only for 11.76% of the results of the Program there are funds for their development in full. This is due to the underdeveloped interactions between business and the public sector of research and development. Performers of STP events do not have the opportunity to include new participants in their composition, establish external relations with business structures and enter into network interactions with other performers. The economic mechanism of interaction between NIS actors participating in the implementation of scientific and technical progress activities is limited by a small percentage of work that is allowed to be performed under construction contracts.

Any dysfunction of the NIS is translated into each of its elements. The considered NIS dysfunctions are transformed into mediocre results of STP. To overcome the existing causes of dysfunction within the framework of a separate project and government programs as a whole, effective proposals for their mitigation are needed.

The main results of the formation of an environment favorable for increasing the innovative and investment attractiveness of the Program results should be: creation of effective tools for identifying promising projects and state support for further promoting the program results; intensification of efforts to improve the flexibility and development of risk sharing mechanisms between the state and
manufacturing companies; high focus on stimulating links between various participants in innovation processes, as well as on the formation and development of research and production partnerships.

5. Conclusion
The successful practical implementation of the results of scientific and technological progress is determined by their absorption potential. A systematic analysis of the absorption potential of the Program made it possible to identify a number of areas for improvement, which are associated with the fact that the implementation and practical implementation of the results requires significant material costs and their payback takes time. In addition, there is a need for production and auxiliary areas, solving personnel training issues and a number of others, which NIS actors cannot always foresee and take into account. Some issues of practical implementation of the results of the Program are difficult to solve even with sufficient funds.

The noted areas for improving the absorption potential reflect the general dysfunctions of the NIS in the field of diffusion of results in the chemical engineering field. To achieve effective practical implementation of the results of scientific and technical progress, it is necessary to provide a number of organizational and economic measures aimed at mitigating dysfunctions. Without ensuring information transparency at all levels of decision-making, the number of abuses in the economy and in society will grow so rapidly that the transition to a competitive market and sustainable economic growth will be impossible.

To stimulate the development of high-tech sectors of the economy, it is necessary to exempt from taxes all profits used in the production of small and medium-sized businesses. This tax practice has had a positive effect in the South-Eastern countries [19, 20]. Institutional reforms are also needed to ensure the mutually coordinated work of all NIS participants. This will lay the foundation for a competitive innovation market and increase the number of potentially lucrative long-term investment projects in clean technologies.

In addition, the Russian economy must become more open than the government envisions. In particular, there is no need to create obstacles for foreign investment in strategic sectors, since this weakens the influence of global technological progress on them. Effective practical implementation of numerous results of state scientific and technological progress will allow the country to obtain long-term competitive advantages in the world economy.

Acknowledgments
The article is prepared with the financial support of the Russian Science Foundation, project № 19-78-10055.

References
[1] Ritwik M and Tirthankar G 2018 Procedia Computer Science 135 178
[2] Ignacio A., Arturo C. and Carlos-Francisco M. 2019 Knowledge-Based Systems 180 1
[3] Hyuk J. and Yong T. 2009 Expert Systems with Applications 36 8986
[4] He Z. and Jing M. 2017 Physica A: Statistical Mechanics and its Applications 494 225
[5] Cohen W. and Levinthal D. 2000 Strategic Learning in a Knowledge Economy 3 39
[6] Gargate G. and Momaya K. 2018 World Patent Information 52 29
[7] List J. 2020 World Patent Information 60 101
[8] Hoock C. and Brown A. 2020 World Patent Information 61 48
[9] Wu K., Li L., Li J. and Li T. 2013 Information Sciences 224 118
[10] Hirao T., Nishino M., Yoshida Y., Suzuki J., Yasuda N. and Nagata M. 2015 IEEE/ACM Transactions on Audio, Speech and Language Processing 23 2081
[11] Radev D., Hovy E. and McKeown K. 2002 Computational Linguistics 28 399
[12] Petrovsky A. 2018 Communications in Computer and Information Science 934 125
[13] Petrovsky A. 2015 Scientific and Technical Information Processing 42 470
[14] Petrovsky A., Pronichkin S., Sternin M. and Shepelev G. 2019 Scientific and Technical Information Processing 46 1
[15] Datta S., Saad M. and Sarpong D. 2019 Technological Forecasting and Social Change 143 27
[16] Reale F. 2019 Technology in Society 59 174
[17] Jena A. and Philipson T. 2008 Journal of Health Economics 27 1224
[18] Reinhardt R. and Pronichkin S. 2018 MGIMO Review of International Relations 58 94
[19] Muhl S. and Talpsepp T. 2018 The Quarterly Review of Economics and Finance 68 226
[20] Dierick N., Heyman D., Inghelbrecht K. and Stieperaere 2019 Journal of Economic Behavior & Organization 163 190