The concept of forecasting based on semantic analysis of text information

A I Ivanus1, V P Nevezhin2, A A Pustynkin3, I S Vladimirov3 and E I Chernova4

1 Department of System analysis in Economics, Financial University under the Government of the Russian Federation, 49 Leningradsky Prospekt, 125993, Moscow, Russia
2 Department of Mathematics, Financial University under the Government of the Russian Federation, 49 Leningradsky Prospekt, 125993, Moscow, Russia
3 Federal state unitary enterprise "All-Russia scientific and research institute "Center", 11 Sadovaya-Kudrinskaya str., 123242, Moscow, Russia
4 Faculty of international economic relations, Financial University under the Government of the Russian Federation, 49 Leningradsky Prospekt, 125993, Moscow, Russia

E-mail: VPNevezhin@fa.ru

Abstract. The main issue, which may be practically impossible to solve, is that in making a long-term forecast, all available data usually relate to the past or present, and there is not enough exact and reliable information for future issues. The following feature is observed: the larger the forecast horizon, the amount of information about the future state not only decreases, but also changes its structure. So, if all information will be presented as consistence of digital and text components, the relative amount of digital information for forecasting decreases, and the text information increases. In the limit, with a significant forecast horizon of decades, we will not have digital information at all, but only text information. The article considers the method of using text information to solve a wide range of problems in long-term forecasting. The method is based on K. Goedel's idea of semantic truth in relation to the consistency of digital and textual statements. It is theoretically justified and practically shown that from these positions textual information after the implementation of the semantic analysis procedure, as a result of which when semantic cores of truth of statements are formed, they can be used for forecasting. This approach is methodically new and does not require the use of complex software tools. An algorithm for implementing the procedure for semantic analysis of information is proposed, as a result of which, a concept map is formed that allows analyzing the object of forecasting more systematically, without ignoring its elements and interconnections between them. Using the concept map, generation of new knowledge about the future States of the forecast object can be proceed and an "array" of possible options can be created. Subsequently, the most probable forecast data is selected from this "array". The presented approach is based on using only text information without taking into account digital information, but in real forecast tasks, of course, digital information must also be taken into account. Obtained results illustrate the method and also interest wide class of long-term forecasting problems.
1. Introduction
A long-term forecasting is one of the most difficult problems that people have tried to solve at all times. The desire to look into the future and anticipate events has always been considered as an issue that deserves universal respect and recognition. Therefore, solving problems related to the generation of theory and methods of forecasting, is always a priority.

Modern science has developed many forecasting methods based mainly on the use of mathematics, statistics and the availability of appropriate databases. This is truth beyond doubt. However, the main issue here is that all available data, as a rule, relate to the past or present, and for future events there is not enough exact and reliable information. And this issue is fundamentally unsolvable.

At the same time, we can observe one interesting feature. Thus, as the forecast horizon increases, the amount of information about its future state decreases. This is a natural process. The structure of this information is also changing. If all the available information will be presented in two components – digital and text, the relative amount of digital information decreases, and text – increases. In the limit, with a forecast horizon of decades, we will not have any digital information at all, but only text information [1, 2].

What is the difference between digital and text information? First of all, consistency, reliability, correctness, credibility, or in the general understanding – truth. Even these qualities, localized in digital information, show its significant advantage. Text information does not have this property. The truth and reliability of text claims cedes numerical pointing’s and mathematical dependencies.

If we consider these two components of information from the point of view of K. Goedel’s incompleteness theorem [3], then the consistency condition is provided in digital information [4, 5]. In the text information, consistency of all available statements is not provided [6]. And this is a fundamental point. This means that if the text contains a statement A that is considered to be true, then it can also contain and be true its negation $\bar{A}$. Thus, these two statements can exist in the same time. In order to determine which of the statements is true and which is not, there is necessity to form a metasystem – B, by K. Goedel’s theorem. Only from the point of view of this metasystem, it will become clear which of the two statements A or $\bar{A}$ will be true and which is not [7].

Therefore, if we consider the generation of forecasts from the truest statements, then our approach should use this circumstance.

As a result of the research, the following algorithm for solving long-term forecasts based on text information is proposed.

2. Scheme for solving the problem of generating forecasts
The proposed scheme for forming long-term forecasts includes 9 main units that determine the algorithm for their solution, figure 1.

![Figure 1](image)

Figure 1. Flowchart of the solution of the forecast problem based on text information.
2.1 Description of flowchart units

Unit 1. Database of text information
In developing a method of long-term forecasting (taking into account the increase of forecast horizon), we will proceed from the following obvious assumptions:

- the amount of information about the future state is formed mainly from digital and text information. We can also take into account the existence of such types of information as emotional and sensory. But in this research, we do not consider them;
- the total amount of digital and text information naturally decreases, but the share of the total amount of text information, as noted earlier, increases.

In such conditions, the role of text information becomes more important, and the necessity to develop methods of forecasting based on text information only increases.

Unit 2. Image of the future
The following important point should be noted first. The forecast, whether it is short-term or long-term, begins with the generation of a certain prototype of the future that we predict. Therefore, there may be the following options for it [8, 9]:

- there is some mathematical model for predicting the process we are interested in;
- there is some text description of this process in the future;
- the complete absence of any idea of the future, as well as the absence of any mathematical model or text description.

Unit 3. Ontology construction
An ontology of the forecast object is compiled as a regular graph based on the available text information.

The algorithm of ontology construction based on an array of text information is following:

- At the first stage, a thesaurus (dictionary of key terms) is formed, which serves to be information material for ontology construction. All concepts in the thesaurus must be linked through their own definitions. If it is not available, temporarily can be used an unstructured array of text information.
- At the second stage, the received array is processed for the presence of key concepts [12, 13]. To simplify and speed up this procedure, semantic analysis programs, such as "TextAnalyst" can be used. After downloading the information array into this program, the following window will be displayed (figure 2). In the upper-left corner of the window, the most important and frequently encountered thesaurus words are listed in order of hierarchy. Since the thesaurus itself is part of the forecasting methodology, so the key term is "forecasting", and under it are three dependent concepts: "forecasting object", "forecasting method" and "development", which, in turn, can unfold and contain the next level of the hierarchy.
Figure 2. Interface of the program "TextAnalyst" (a fragment of the example is given in Russian version of the program).

- When all the interconnections found by the program have been determined, they are displayed on the concept map using the CmapTools computer program [14]. An example of creating such a concept map based on 100 definitions from the thesaurus is shown on figure 3. Thus, we get a map of the subject area, where all the terms are precisely defined without using expert methods and that are in a strong interconnection. Figure 3 shows an example of an ontology for solving tasks of the forecast we are interested in. The given ontology allows us to evaluate the boundaries of the studied area of knowledge, which will reduce the probability of missing any important details in making a forecast.

As a rule, the ontology contains redundant information as extra branches of its graph. This is not difficult to see if we analyze well-known examples of constructing ontological graphs. It seems that there is a certain competitive mood among the authors in terms of increasing the dimension of ontologies. In practice, it is difficult to use such a graph structure precisely because of the redundancy of information. Although it would be more correct to direct these possibilities in the direction of solving the problem of searching for interdisciplinary connections and dependencies rather than simply increasing the dimension of the graph. Excessive ontology also introduces a certain level of ambiguity in the understanding of the subject area under study. For example, in the preprint of the Higher school of Economics (Moscow, Russia) "Ontological modeling of the economy of enterprises and industries in modern Russia" [15] one of the disadvantages of existing automated systems for constructing ontology, among other things, is the small number of possible interconnections, namely: "genus-species", "partitive", "associative". According to the authors of the article, a larger number of interconnections is often redundant, and compiling an ontology, it is advisable not to go beyond the three basic interconnections between concepts [15, 16].
Figure 3. Ontology of the subject area of the forecasting process (a fragment of the example is given in Russian version of the program).

Unit 4. Semantic analysis and semantic cores constructions
It is preferable to implement the simplification procedure using semantic cores of knowledge truth (terms) - SCKT, which are included in the structure of the ontology and formed using a specially developed procedure presented in [7]. As an example, let’s consider the process of forming the semantic core of truth of the term “science”.

At the first stage, variants of the definition of the word "science" are selected. There were 44 of them in total. The following are only the first five of them.

1. Science is an area of human activity, the function of which is development and theoretical systematization of knowledge about reality; includes both the activity of obtaining new knowledge
(NK), and its result – the sum of knowledge underlying the scientific picture of the world;

2. Science is a special type of cognitive activity aimed at developing objective, systemically organized and justified knowledge about the world.

3. Science is an area of human activity, the function of which is the development and theoretical systematization of objective knowledge about reality.

4. Science is an area of human activity, the function of which is the development and theoretical schematization of objective knowledge about reality.

5. Science is the most important element of spiritual culture, the highest form of human knowledge; a system of developing knowledge that is achieved through appropriate methods of knowledge, expressed in precise concepts, the truth of which is verified and proved by public practice.

The collected versions of definitions are loaded into the "Miratext" semantic analysis program. As a result, we get the following information about the occurrence of the term "science", table 1.

Columns 4 and 5 of table 1 show the results of processing of values the received word occurrence, namely:

Ln (gr. 1) - logarithm of the sequence numbers of column 1;

Ln (gr. 3) - the logarithm of the meaning of the occurrence of words in the term "science".

According to the data of the last two columns, a graph of the profile of semantic cores of truth knowledge of term "science" is constructed, figure 4.

| № | Words in definition | The frequency of occurrence (in descending order) | Ln (gr. 1) | Ln (gr. 3) |
|---|---------------------|-----------------------------------------------|------------|------------|
| 1 | knowledge           | 39                                            | 0.00       | 3.66       |
| 2 | activities          | 34                                            | 0.69       | 3.53       |
| 3 | production          | 33                                            | 1.10       | 3.50       |
| 4 | world               | 24                                            | 1.39       | 3.18       |
| 5 | objective           | 20                                            | 1.61       | 3.00       |
| 6 | people (person)     | 13                                            | 1.79       | 2.56       |
| 7 | area                | 13                                            | 1.95       | 2.56       |
| 8 | system              | 13                                            | 2.08       | 2.56       |
| 9 | to direct           | 11                                            | 2.20       | 2.40       |
| 10| human               | 11                                            | 2.30       | 2.40       |
| 11| society             | 10                                            | 2.40       | 2.30       |
| 12| form                | 10                                            | 2.48       | 2.30       |
| 13| nature              | 9                                             | 2.56       | 2.20       |
| 14| which               | 9                                             | 2.64       | 2.20       |
| 15| scientific          | 8                                             | 2.71       | 2.08       |
| 16| foundation          | 7                                             | 2.77       | 1.95       |
| 17| theoretical         | 7                                             | 2.83       | 1.95       |
Based on the graph, we can say that the first three words form a so-called "plateau", which is the direct meaning of the term "science". In fact, it is a set of interrelated words with a common meaning "knowledge-generating activities". In this case, SCKT quite short, it consists of only three words, but nevertheless everyone understands the meaning inherent in it. Despite the fact that there are many opinions in society about the essence of scientific activity, however, the absolute majority of them agree with the written above.

After working through each term in the concept map in this way, more comprehensive and up-to-date version of it can be created, because some terms will fall out of it due to rare use or too large a spread of interpretation.

**Unit 5. Generation of new knowledge**

The concept map of a forecast object allows to create a number of options for its future states based on the analysis of the development of its elements. The generation of new knowledge is possible only if the basic principles of science are observed, one of which is the "principle of consistency". In other words, the concept map allows you to think systematically about the future states of the object under study. The following methods and algorithms are possible for artificially creating a new semantic core of knowledge truth [17]:

1) violating current cause-and-effect interconnections of the existing SCKT:
   - method of alternatives,
   - method of "provocations",
   - random selection method (for children),
   - method of analogies,
   - special mathematical algorithms (game [18], etc.);
2) creating new cause-and-effect interconnections of new SCKT ("Medici Effect" [19]).

**Unit 6. Selection of options**

After creating an "array" of options for future states of an object, begins its analysis and selection the most probable states, which will be further refined.

**Unit 7. Expert adjustment of image of the future**

Expert adjustment of the received forecast is a necessary step in case of forecasting unstable systems, or if the forecast is long-term. It is based on the latest information about the object or its components (concepts on the concept map) [20].
Unit 8. Enrichment of extra information
In case, if none of the previously selected options for the development of the forecast object is acceptable, there is a necessity of searching an additional information that will allow to make the necessary adjustments to image of the future and get a more accurate forecast result.

Unit 9. Semantic thesaurus
The semantic thesaurus is intended for intellectual unification of terms and definitions. It is created as a repository of the semantic structure of the language in order to form a unified system of terms and definitions [10, 11].

Tasks solved by the semantic thesaurus to achieve a given goal:
1) overcoming diversity in the interpretation of terms and definitions included in scientific, industrial and social turnover, based on the processes of their unification;
2) creating a logical structure for an array of terms and definitions in accordance with the criteria of completeness, accuracy, relevance, reliability, timeliness, consistency, etc.;
3) informational support of innovation processes in terms of providing users with unified terms and definitions included in the semantic thesaurus database;
4) semantic analysis of terms and definitions of a wide profile;
5) providing users with high-quality information in accordance with the specified subject;
6) identification of the presumed area of new knowledge existence based on semantic analysis of terms and definitions of a given scientific innovation topic in order to identify both the most and the least significant areas of predicted processes.

3. Conclusions
To sum up:
1. Method for analyzing text information about the object of forecasting in the framework of solving long-term forecasting problems is proposed. The basis of the method is to identify and establish interconnections between main concepts and their reflection in the concept map. The current level of technology development allows to automate the process of analyzing text information using appropriate software.
2. The received concept map allows to analyze the object of forecasting more systematically, without ignoring its elements and interconnection between them. Having it, there is a possibility to start a new knowledge about the future states of the forecast object and create an "array" of possible versions of them. Subsequently, the most probable forecast options are selected from this "array".
3. The presented approach is based on using only text information without taking into account digital information. However, in real forecast tasks, of course, the last one must be definitely taken into account.

References
[1] Taussky-Todd O 1987 Membrances of Kurt Gödel, Godel remembered Salzburg 10-12 July 1983 29 – 41
[2] Grattan-Guinness I 1979 In memoriam Kurt Gödel: his 1931 correspondence with Zermelo on his incompletability theorem Historia Mathematica 6(3) 294-304
[3] Dawson J W 1986 The papers of Kurt Gödel Historia Mathematica 13(3) 277
[4] Novak J D and Cañas A J 2015 The theory underlying concept maps and how to construct and use them Institute for Human and Machine Cognition: http://cmap.us
[5] McAleese R 1998 The Knowledge arena as an extension to the concept map: reflection in action Interactive Learning Environments 6(3) 251–272
[6] Androsov A Yu and Shkilev V D 2017 Philosophical approach to the interpretation of the extended Gödel incompleteness theorem Electronic journal Science, technology and education 12 267-671 (in Russian)
[7] Ivanus A I 2019 Cognitive methods and technologies of economic management in conditions of uncertainty (Moscow: Prometheus) 151 (in Russian)
[8] Guriano N, Gangemi A, Pisanelli D M and Steve G 1999 An overview of the ONIONS project:
applying ontologies to the integration of medical terminologies Data & Knowledge Engineering 31

[9] Guriano N 1997 Understanding, building, and using ontologies : using explicit ontologies in kbs development International Journal of Human and Computer Studies 46 (2-3)

[10] Hearst M A 1992 Automatic acquisition of hyponyms from large text corpora In Proceedings of the 14th International Conference on Computational Linguistics 539–545

[11] Cardoso J 2007 Semantic web services: theory, tools and applications IGI Global, Hershey PA

[12] Talmy L 1977 Rubber-sheet cognition in language Papers from Regional Meeting of the Chicago Linguistic Society

[13] Carey S 1999 Knowledge acquisition: enrichment or conceptual change? In E Margolis and S Lawrence, concept (Massachusetts: MIT press) 459–489

[14] Cañas A J and Novak J D 2014 Concept mapping using cmaptools to enhance meaningful learning / Knowledge cartography: software tools and mapping techniques Advanced information and knowledge processing (2nd ed) (New York: Springer-Verlag) 23–45

[15] Efimenko I V and Khoroshevsky V F 2011 Ontological modeling of the economy of enterprises and industries in modern Russia Part 3 Russian research and development in the field of ontological engineering and business ontologies preprint WP7 / 2011/08 / House of the Higher School of Economics(Springer) 68 (in Russian)

[16] Bubareva O A 2019 Investigation of mechanisms for automatic construction of ontologies over a set of unstructured data South Siberian Scientific Bulletin 22 77-82 (in Russian)

[17] Chow G C 1960 Tests of equality between sets of coefficients in two linear regressions Econometrica 28 (3) 591-605

[18] Nash J 1951 Non-cooperative games Annals of Mathematics (Princeton New Jersey: Princeton University) 54 (2) 286–295

[19] Johansson F 2004 The medici effect: Breakthrough insights at the intersection of ideas, concepts & cultures (Harvard Business School Press Boston)

[20] S H T E M Forecasting, step 7: adjusting the forecast: https://shtem.ru/manual-correction-forecast