Natural Language Processing for Forecasting Innovative Development of the Energy Infrastructure

Alex Kopaygorodsky[1,2,*]

[1]Melentiev Energy Systems Institute SB RAS, Irkutsk, Russia Federation
[2]Irkutsk National Research Technical University, Irkutsk, Russia Federation

Abstract. The article deals with the application of natural language processing methods to support research and forecasting the innovative development of energy infrastructure. The main methods of NLP, which are used to build an intelligent system to support scientific research, are considered. Methods of building infrastructure for processing Open Linked Data and Big Data are described. Semantic analysis and knowledge integration are based on ontology system. Applying suggested methods allow increasing quality of scientific research in this area and make it more effectively

Keywords. natural language processing, knowledge management, ontology, forecasting innovative development.

1 Introduction

In recent years, methods of natural language processing and semantic analysis have been actively developing, which makes it possible to use them to solve problems of scientific and technological forecasting. These methods are also applied in the energy technology research area. The use of Data Science methods in high-tech industries can significantly increase the efficiency of management decisions. Such an approach to the management of individual organizations at various levels are called "data-driven management". When implementing this approach, specialized information and analytical departments are created in organizations under the leadership of CDO (Chief Digital Officer / Chief Data Officer). The opinion of CDO is key in matters of company development, identification of new business opportunities, ensuring entry into new market segments, bringing fundamentally new products, services and services to the market, etc. The main source for making management decisions is the results of the analysis of information and data collected from various sources. These are analysed to identify existing and emerging trends and shorten the response time to them, which reduces material losses, increases profits and ensures the sustainable development of the company. Ignoring technical and economic trends can be fatal or cause significant financial damage, examples of such companies are Polaroid, Kodak, Motorola, 3Com, Sun Microsystems.

The application of the described approach is relevant when solving forecasting problems and organizing monitoring of innovative technological solutions in any sector of the national economy, and especially for the development of the energy infrastructure of Russia. The energy infrastructure is the basis for the functioning of other industries, the results of which are ultimately aimed at improving the quality of life of the population. Predictive methods for the development of the energy industry, based on traditional mathematical models and software systems, are not always effective in conditions of uncertainty and lack of the necessary reliable information for the available models. The involvement of intelligent methods of semantic analysis, machine learning and Big Data technology to create tools that facilitate the work of experts and tools that perform preliminary processing of information analysed by experts is relevant.

2 Related Works

Forecasting as a method of research is used in the domain of Energy Infrastructure to study the development and functioning. In [1], authors considered the technological prospects of various directions of decisions of the problem of resource restrictions of the development of wind and solar energy. From that were drawn conclusions on the prospects of development of the Russian high-tech sectors of the Energy Industry and Economy. In addition, there is the problem of comparatively low installed capacity utilization of wind turbines in Russia [2]. Unfortunately, in view of the peculiarities of Russia's energy sector, the use of biofuel does not have any significant distribution. However, the use of biofuels is being actively studied in other countries of Europe and Asia [3, 4].

In [5] authors propose a language-independent semantic method for implementing an extractive multi-document summarizer system by using a combination of statistical, machine learning based, and graph-based methods. The author tool set learns the semantic

* Corresponding author: kopaygorodsky@isem.irk.ru

representation of words from a set of given documents via word2vec method.

Also, Artificial Intelligence techniques successfully for forecasting the conduct of separate energy technologies. In [6], authors use patent indicators to predict the technological advances in Hydrogen Storage Materials (HSM). The patent analysis was carried out using bibliometrics and Text Mining approaches in order to forecast the future trend of development. Authors evaluated the technological life cycle stage, HSM class prominence, and the role of different countries in HSM patenting. They suggested that the life cycle stage of HSM is near market deployment.

The importance of knowledge sharing and the impact of this process on innovation are reviewed in [7]. Authors note that dialogue is the main tool for turning knowledge into innovation. This knowledge is also shared to stimulate innovation. Authors determined from the key topic analysis that the most established topics are open innovation, knowledge transfer, and absorptive capacity. This last concept facilitates that organizations identify and interiorize external knowledge that contributes to the achievement of institutional goals.

Employees of the Melentiev Energy Systems Institute of Siberian Branch of the Russian Academy of Sciences (ESI SB RAS) are engaged in the development of intelligent methods and tools to support decision-making in the field of energy research as well. Researchers use methods based on intelligent semantic technologies for searching, extracting and analysing heterogeneous data from electronic sources of information in accordance with the Big Data concept. [8-10]

3 Approaches for Natural language processing

There are many natural language processing methods used in information systems. All methods can be roughly divided into two large groups: statistical and linguistic methods. [11]

Statistical methods are based on the analysis of the frequency of occurrence of words: counting the number of occurrences of words in various fragments, the distribution of frequency across documents, etc. Linguistic analysis, on the other hand, is based on identifying individual words, analysing their morphological features, syntactic and semantic analysis of text fragments.

Stemming technology plays an important role in improving the analysis results and reducing the search area. Stemming allows you to identify the basis of a word, to associate many forms of the same word with each other, which makes it much easier to process text arrays. An important feature of stemming is its dependence on the language, since the word formation rules for each natural language are usually specific. The best results are obtained with stemming based on pre-calculated tables and dictionaries. However, this approach also has some difficulties: it is impossible to determine the stem of a new word that has not been previously processed.

One of the most common statistical measures is TF-IDF. TF is Term Frequency. IDF stands for Inverse Document Frequency. The TF-IDF measure allows you to assess the importance of a word in the context of a document, which is included in a collection of documents (corpus). TF-IDF of a word is proportional to the frequency of the word in the document and inversely proportional to the frequency of the word in all documents in the corpus. The TF-IDF measure makes it quite easy to separate words from general vocabulary from specific terms. However, an improperly selected corpus greatly affects the TF-IDF gauge. One of the main uses of TF-IDF is the vector representation of documents in a collection.

Another statistical analysis method is word2vec. This algorithm allows you to represent words as a multicomponent vector. Word2vec was developed by Google in 2013 and has been reflected in commercial projects of many companies. Word2vec is based on a collection of artificial neural network models designed to generate vector representations of words in natural language. Word2vec uses a large collection of text documents as input and maps each word to an N-vector. The resulting vectors allow calculating the semantic proximity of words.

The most common in the field of data analysis are two languages - Python and R.

Python's popularity is driven by a large number of libraries. One of the more interesting libraries for linguistic analysis of Python texts is the Natural Language Toolkit (NLTK). The NLTK library includes many tools for corpus management and analysis.

To solve the problem of supporting the forecasting of the innovative development of the energy infrastructure, both statistical and linguistic methods of text processing were used. After the construction of the terminological dictionary of the subject area, vector representations were calculated for its elements. When processing arrays of textual data, at the first stage, filtering and separation from frequently used words and other similar language elements were performed. Then the classification and semantic comparison with the elements of ontologies was performed. Further, the obtained characteristics of each document were processed statistically in order to identify general trends and patterns.

4 Semantic analysis of Big Data

Linked Open Data from state information systems, as well as from some commercial systems, are used as sources of information to predict the innovative development of energy infrastructure. All of these sources can contain potentially interesting information, but there can also be information noise. Examples of such sources are databases of scientific publications, conducted research, the results of intellectual activity, etc. Such sources, as a rule, adhere to a certain structure of the published data, and therefore can be processed using software adapters. In addition, Internet search

engines can be used to search for unstructured but potentially interesting information for researchers. After the automatic acquisition of information, the results are analysed, classified and subsequently evaluated.

Scanning information sources regularly allows not only filling the knowledge store but also tracking the dynamics of changes in qualitative and quantitative indicators based on cognitive models. Analysis of the results of the issuance of Internet search engines allows you to assess information interest and track trends in technology development since the order and search results depend on the interest of many users. The technology for organisation monitoring of open and big data is based on the use of a pool of crawlers (software robots) specific to certain types of information resources. The main task of search robots is to extract and unify information about resources. It is necessary to perform some stages to analyse information posted on various sites. The first step is to get links to such sites. Today there are well-known information retrieval systems that are widespread on the Internet and have indexed information resources. The work of crawlers is based on extracting primary information from the databases of such search engines: at this stage, you can get the resource address, title and description. Then a detailed analysis of the content of the resource obtained from the found address is performed.

In the structure of the subsystem for information retrieval and analysis of open sources, two types of nodes can be distinguished: operational and control. The main task of operational nodes is to ensure the work of crawlers and data adapters. Control nodes are responsible for setting tasks for operational nodes, overall coordination and agreement of results. Due to its specifics, the task of information retrieval and analysis can be divided into many parallel processes, while the number of operational nodes can be large enough, which will increase the power of the information retrieval and analysis subsystem. Thus, it is advisable to locate some components for search in data centres (as close as possible to the backbone communication channels). In data centres on rented computing resources, clusters of data adapters that process structured sources of open data on demand and crawler farms that scan the Internet space are deployed. When testing data collection, the control node locate in Eastern Siberia, where connection to large trunk communication channels is difficult, operational nodes were located at traffic exchange points and were geographically locate in Central Europe.

5 Knowledge integration based on ontologies

When justifying decision-making on the paths of innovative development, due to the complexity and multifactorial nature of the estimates and options obtained, a rational approach is a combination of formalized methods and the informal experience of experts. The results obtained at each stage should be presented in a comfortable and easily perceived form both for experts and, as a result, for a decision-maker. This feature is very important. Using the visual analytics apparatus allows you to get such an idea. Visual analytics tools make it possible to evaluate in detail the main data sets for substantiating the innovative development of the energy sector: by research and development in the field of energy technologies, by the composition and topology of connections, by intensity and significance, by activity centres and other parameters presented in various analytical sections. Also, researchers perform visualization of solutions to multi-criteria problems on a variety of predefined scenarios. The scenario approach reduces the uncertainty factors of external and internal conditions at the stages of choosing a solution.

The integration of information and knowledge obtained from various sources is carried out based on a system of ontologies. Semantic integration uses a common conceptual framework for mapping (linking) individual elements based on it. An ontological space that includes a collection of ontologies does this. The system of ontologies consists of the ontologies of the fuel and energy complex, energy industries and individual energy technologies, energy research. The ontology system is the main component of the search engine, based on which the domain setting and allow to make a knowledge integration.

The ontology of energy technologies describes a hierarchical end-to-end conceptual model of the subject area that describes the object of research, specifies the structure of energy technologies, levels and slices of their consideration. The ontology of energy technology indicators describes the classification of technical, economic, environmental and social indicators of energy technologies, contains a description of the models of aggregation, generalization, comparison and use of technology indicators.

The main indicators include the level of technology maturity, the intensity of R&D in the main and related areas, the growth of application potential, key countries and organizations, technical barriers, and trends. The ontology defines the current forecast and limit values of indicators. Indicators are used in the selection of promising technologies for subsequent systemic consideration.

Ontologies describe the hierarchy of energy technologies with sufficient detailing of their components and interrelationships at different levels (resources, functions, types of energy conversion, consumer services, infrastructure, management, etc.); specifications of technologies or their characteristics for technical and economic efficiency; specifications of the full life cycle of energy technology (Life Cycle Assessment); specifications of socio-economic factors; specifications of indicators of innovative technology development; define a conceptual basis with bilingual reference and synonyms for concepts of all levels.

A fragment of a bilingual ontology is shown in Figure 1.

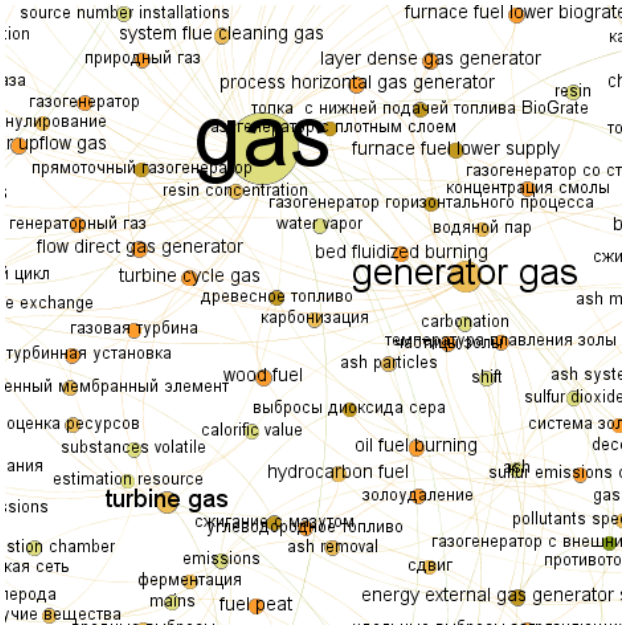

Fig. 1. Fragment of the bilinguistical ontology of energy technologies.

During the data collection process, more than 2 million primary documents were obtained from various open sources, more than 450 thousand documents were identified and classified based on the built ontology system. The knowledge portal organises access to information on 135 thousand scientific articles on energy systems published in 2009-2019. And about 213 thousand valid patents issued in different countries in 2009-2019, the application of which is possible in the field of energy.

The author analysed information on advanced research in the field of energy. A sharp increase in China's work in the field of energy and related engineering and materials science over the past 5 years has been revealed. Such an increase in scientific interest and the number of patents indicates improvements in existing and the creation of new mobile sources of electricity (high-capacity batteries) and, possibly, superconductors. Practical results will allow China to make a significant contribution to the development of the global electric car market in the next 5-7 years.

6. Conclusion

Natural language processing and semantic analysis of Big Data can be successfully applied to predict the innovative development of energy infrastructure. Also, these methods can be applied in solving scientific and practical problems of a similar class in other subject areas. Application of these methods and tools will facilitate preparation and increase the validity of advanced recommendations and decisions in the field of strategic energy development. These methods provide the organization of monitoring of innovative scientific and technical solutions and technologies in the energy sector, including their assessment of their effectiveness and feasibility, taking into account the characteristics and needs of the economy.

The author is grateful to the Russian Foundation for Basic Research (RFBR) for financial support. The reported study was funded by RFBR, project number 20-07-00994.

References

1. R.M. Nizhegorodtsev, S.V. Ratner, *Trends in the Development of Industrially Assimilated Renewable Energy: The Problem of Resource Restrictions*, Thermal Engineering, **63**, 3, DOI:10.1134/S0040601516030083 (2016)
2. V. Iosifov, E. Khrustalev, S. Larin, O. Khrustalev, *Strategic Planning of Regional Energy System Based on Life Cycle Assessment Methodology*, International Journal of Energy Economics and Policy, **10**, DOI: 10.32479/ijeep.8791 (2020)
3. M. Banja, R. Sikkema, M. Jegard, V. Motola, J.-F. Dallemand, *Biomass for energy in the EU – The support framework*, Energy Policy, **131**, ISSN: 0301-4215, DOI: 10.1016/j.enpol.2019.04.038 (2019)
4. S. Kripal, *India's bioenergy policy*, Energy, Ecology and Environment, **4**, ISSN: 2363-8338, DOI: 10.1007/s40974-019-00125-6 (2019)
5. M. Bidoki, M.R. Moosavi, M. Fakhrahmad, *A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities*, Information Processing and Management, **57**, DOI: 10.1007/10.1016/j.ipm.2020.102341 (2020)
6. L.F. Chanchetti, S.M.O. Diaz, D.R. Leiva et al., *Technological Forecasting of Hydrogen Storage Materials using Patent Indicators*, International Journal of Hydrogen Energy, **41**, DOI:10.1016/j.ijhydene.2016.08.137 2016
7. D.I. Castaneda, S. Cuellar, *Knowledge sharing and innovation: A systematic review*, Knowledge and Process Management, ISSN: 1092-4604, DOI: 10.1002/kpm.1637 (2020)
8. A. Kopaygorodsky, *Technology of Application of Software Tools for Energy Technology Forecasting*, Atlantis Highlights in Computer Sciences, **3**, ISBN 978-94-6252-868-0 DOI: 10.2991/csit-19.2019.47 (2019)
9. A. Kopaygorodsky, I. Khairullin, *Support of Collective Decision-making for Forecasting of Energy Technology*, Advances in Intelligent Systems Research, **169**, ISSN: 1951-6851 DOI: 10.2991/iwci-19.2019.2 (2019)
10. A. Mikheev, *Ontology-based Data Access for Energy Technology Forecasting*, Advances in Intelligent Systems Research, **158**, ISSN: 1951-6851 DOI: 10.2991/iwci-18.2018.26 (2018)
11. V.V. Dikovickij, M.G. Shishaev, *Natural language text processing in search engine models*, Proceedings of the Kola Scientific Center of the Russian Academy of Sciences, **3** (2010, in Russian).