Technologies for Enrich Knowledge Warehouse for Solving Problems of Technology Forecasting and Research of Critical Infrastructures

Elena P. Khayrullina [1,2]

[1] *Melentiev Energy Systems Institute of Siberian Branch of the Russian Academy of Sciences,*
Lermontov str., 130
Irkutsk, Russia
E-mail: Lena-Skoklenyova@yandex.ru

[2] *Irkutsk National Research Technical University,*
Lermontov str., 83
Irkutsk, Russia
E-mail: ksevgelena@istu.edu

Abstract

The article considers scientific and technological forecasting as part of the intelligent energy system, as well as research of critical infrastructures in the field of energy. The task of creating a data warehouse and knowledge for the system described above is discussed in more detail. The advantages of the Big Data technology for solving this problem are given.

Keywords: Big data, technology forecasting, knowledge and data warehouse, critical infrastructures, ontologies, Porter Stemmer.

1. Introduction

Currently, intelligent information technologies are being developed actively in the world. They support the innovative development of many industries and economic spheres.

The importance of using intelligent technologies in Energy is undeniable [1]. Speaking about foreign experience, we can say that such technologies are already becoming the foundation in the development of electricity. In particular, in the US this is the main direction of improving the economy, in China – a way of strategic development of the state. The countries of the European Union use innovative technologies as a basis for a new energy policy [2]. One of the components of intelligent energy systems will be technology forecasting [3].

Technology forecasting can identify and make a preliminary assessment of trends in the development of science and technology. Scientific and technological forecasting can foresee large scientific and technical solutions capable of making changes in a country's overall scientific, technical and production potential, in social relations and world politics.

Information for technology forecasting will be formed on the basis of open data and linked open data. Also, information will be collected from sources such as state information systems that combine data on scientific and technical projects and developments, as well as various commercial systems such as SCOPUS, Web of Science, RINC, Science Index [3,4].

Analysis of information which will be collected from the sources described above can show general trends in the direction of development of scientific and engineering thought [4, 5]. In addition, the technologies and tools developed can be used to analyze threats and assess the risks of cyber-security breaches of critical infrastructures, including energy facilities [6].

2. Features of the knowledge and data warehouse for the tasks of technology forecasting and research of critical infrastructures

To solve the problems of technology forecasting, it is necessary to accumulate and analyze a large amount of data. The more information is processed, the more accurate the results will be. By increasing the amount of data, the completeness of information is increased and it becomes possible to assess the reliability and consistency of the data obtained.

The complexity of creating of knowledge and data warehouse increases due to the exponential increase of various information in the modern world. Annually, the volume of information in the world increases by 30%.

Thus, the knowledge and data warehouse is designed primarily for technology forecasting, based on the analysis of heterogeneous, unstructured data sets of large volumes. For their effective processing and storage, it is necessary to use horizontally scale-out software and distributed computing.

The accumulated volume of information structured in the repository can be used to solve related tasks in energy research. Related knowledge of information technology used at energy facilities can be used to analyze the cybersecurity of these facilities.

3. Big Data technologies

By the term Big Data in this article will mean some technologies, tools, methods for processing structured and unstructured large data, allowing distributed processing of information are meant.

Principles of working with Big Data:

- Horizontal scalability – with the increase in storage volumes, the system must support the scalability of the number of nodes;
- Fault tolerance;
- Data localization – data processing must take place on the same machine on which the data is stored, otherwise the data transmission costs may exceed the processing costs.

To solve the problem of data localization, Google employees proposed the MapReduce concept. MapReduce is a model of distributed computing, used for parallel processing of large amounts of information [7]. MapReduce assumes that the data is organized in the form of some records. Data processing takes place in 3 steps. An illustration of the work of this model can be seen in Figure 1.

Step “Map”. The input data of the problem being solved represent a large list of values, and the preliminary processing is performed on the Map-step. To do this, the main node of the cluster receives this list, divides it into parts and sends it to the work nodes. Next, each of the working nodes converts the elements of the resulting collection to zero or several intermediate key-value pairs.

Step “Shuffle”. It passes unnoticed for the user. At this step, the intermediate results are grouped.

Step “Reduce”. On the Reduce-step, the master node receives intermediate responses from the work nodes and passes them to the free nodes for the next step. The system sorts and groups all key-value pairs by

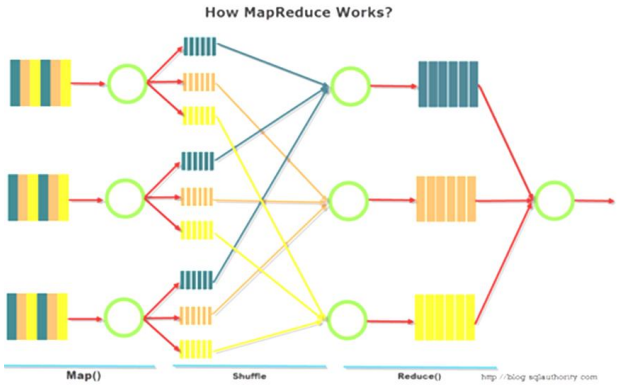

Fig. 1. MapReduce.

key and then, for each key-group of values, collapses values, often into one or an empty list. The result is the solution of the problem, which was originally formulated [7,8].

One of the possible solutions for organizing distributed storage and processing of large data is the NoSQL class database. They can be of 4 types:

- “Key-value” warehouse. “Key-value” warehouse is the simplest. Every item in the database is stored as an attribute name (or "key") together with its value. For example: Berkeley DB, MemcacheDB, Redis, Riak, Amazon DynamoDB.
- Wide-column stores. In this repository, the data is stored as a matrix, the rows and columns of which are used as keys. For example: Apache HBase, Apache Cassandra, Hypertable, SimpleDB.
- Document databases. Pair each key with a complex data structure known as a document. Documents can contain many different key-value pairs, or key-array pairs, or even nested documents. For example: CouchDB, Couchbase, MarkLogic, MongoDB, eXist, Berkeley DB XML.
- Graph databases. Graph databases are used to store information about networks, such as social connections. For example: Neo4j, OrientDB, AllegroGraph, Blazegraph [7].

4. The suggested architecture of the system

To implement the system of scientific and technological forecasting, it is supposed to adhere to the architecture presented in Figure 2. From the previously installed resources, such as elibrary, Scopus, etc., pages are searched that satisfy the search query. Further, the document is extracted (articles, monographs, etc.) and metadata. After that, the received document is converted to a text and it is saved with its metadata to the primary storage. After that, using a predetermined ontology of terms in the field of energy, a semantic

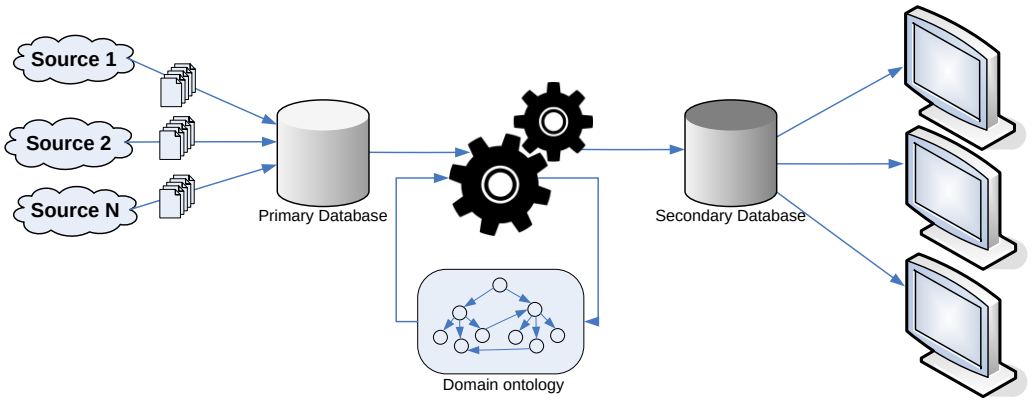

Fig. 2. The proposed architecture of the system.

analysis takes place and the results obtained are transferred to a graph database. After this step, the results will be available to an end user [9].

5. Extracting data for the formation of primary storage

To form a repository of primary documents on the basis of which scientific and technological forecasting will be carried out, it is necessary to be able to extract data from different sources. As mentioned above, such sources can be systems integrating data on scientific projects, such as SCOPUS, elibrary, WebOfScience, etc.

Since each source has its own warehouse structure

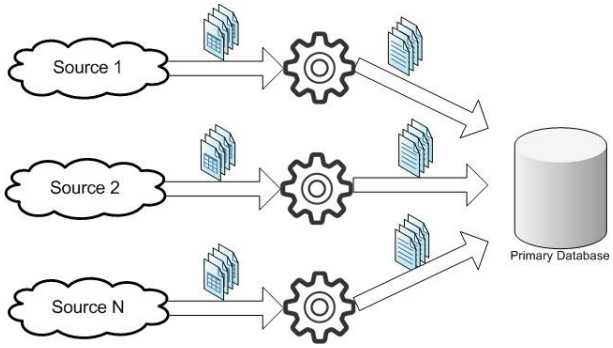

Fig. 3. Extract data.

and document format (PDF, HTML, etc.), then for each designated source you need to develop a separate program module. It will extract the necessary documents and corresponding metadata, be sure to convert each document into a text format and save the received text and metadata to the primary database [9]. The described algorithm is shown in Figure 3.

It is supposed together with the document to extract the following metadata:

- Article title
- The authors of the article
- Date of publication
- DOI
- URL
- Keywords
- Geographical position

6. Data analysis

6.1 Normalization of words

To analyze the resulting array of a text, a preliminary normalization is necessary. For this, stemming will be used. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form. The stem is not necessarily identical to the morphological root of the word; it is usually sufficient for the same reason.

Russian and English are parts of the group of inflectional synthetic languages, that is, languages in which word formation predominates using affixes that combine several grammatical meanings, so this language allows the use of stamping algorithms. The Russian language has a complex morphological variability of words, which is the source of errors when using stamping.

Porter Stemmer is a stemming algorithm that does not use the base of words, but consistently applies a set of rules, cuts off endings and suffixes, based on the characteristics of a language (Figure 4) [10].

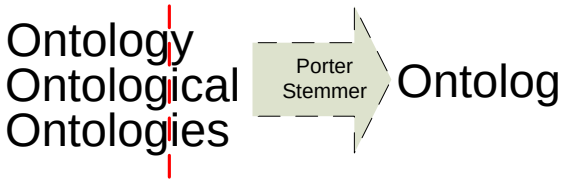

Fig. 4. Porter Stemmer.

6.2 Statistical indicator TF-IDF

In information retrieval, TF–IDF, short for term frequency–inverse document frequency, is a numerical statistics that is intended to reflect how important a

word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The TF–IDF value increases proportionally to the number of times a word appears in a document and is offset by the frequency of the word in the corpus, which helps to a adjust for the fact that some words appear more frequently in general [9, 11].

6.2.1 Statistical indicator TF

In the case of the term frequency **tf(t,d)**, the simplest choice is to use the raw count of a term in a document, i.e. the number of times that the term **t** occurs in the document **d**.

$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k} \quad (1)$$

TF term **t** = Number of times the term **t** was found in the document **d** / number of all words in the document **d** (1) [10].

6.2.2 Statistical indicator IDF

The inverse document frequency is a measure of how much information a word provides, that is, whether a term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|} \quad (2)$$

IDF term **t** = logarithm (The total number of documents **D** / The number of documents **d** in which the term **t** occurs) [10].

6.3 Calculation of TF-IDF

The TF-IDF calculation model is shown in Figure 5. From the primary repository, documents are selected for processing, and TF is calculated for each word in each document. And the IDF indicator is calculated for each word from the entire collection of documents and with each new document the indicator is recalculated.

Once these data have been prepared, the process of searching and classifying documents will be accelerated greatly.

Figure 6 shows an example of calculated data for the search query "Energy in Russia". In the "Term Frequency" table, it is seen that in Document 1 the

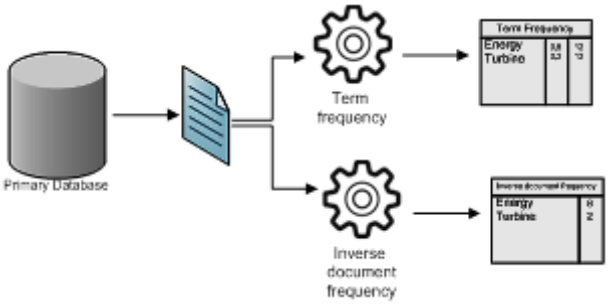

Fig. 5. Calculation TF-IDF.

normalized word "Energ" occurs with a frequency of 0.12 (12 of 100 words), and the word "Russ" once every 100 words. Also there are given the data of their 3 other documents.

In the table "Inverse Document Frequency" the reverse frequency of the word "Energ" - 2,5 and "Russ"

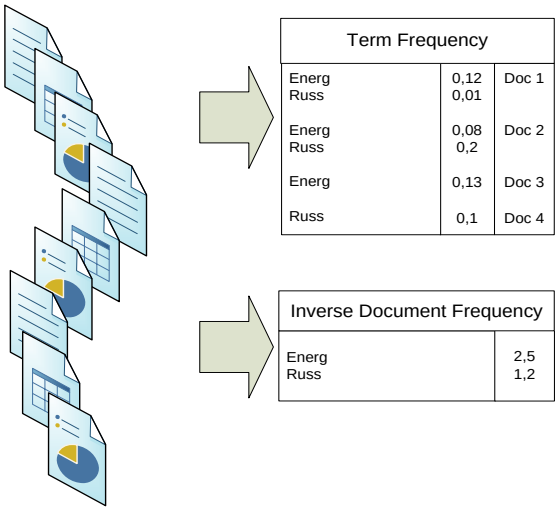

Fig. 6 An example of using the indicator TF-IDF.

- 1,2 is given. Since the indicator is inversely proportional, it means that the word "Energetic" is less common than the word "Russ", which means that it is more important when selecting relevant documents.

After calculating the final TF-IDF (by multiplying TF and IDF), we get the following result:

- Doc 1: TF-IDF = 0.12 * 2.5 + 0.01 * 1.2 =0.31
- Doc 2: TF-IDF = 0.44
- Doc 3: TF-IDF = 0.25
- Doc 4: TF-IDF = 0.156

According to the results it is clear that the most suitable document from the Doc 2 collection, in which the word "Energ" and the word "Russ" are found with a frequency of 0.08.

The next step is to map the ontology classes of the subject domain to the calculated TF-IDF indicators, which will allow the classification of documents [4].

7. Conclusion

The use of Big Data technologies, at the moment, allows storing and processing information distributed on many servers and computers. This solution will support the increase of the storage size in accordance with the amount of required information for the tasks of technology forecasting in the field of energy and the study of critical infrastructures.

TF-IDF indicators will significantly improve the process of classifying and searching relevant documents in the collection.

Acknowledgements

The results were obtained during the implementation of the basis scientific project of the fundamental research programs of SB RAS III.17.2.1, reg. № AAAAA17-117030310444-2, III.17.1.4, reg. № AAAA-A17-117030310436-7, with partial financial support of the RFBR grants № 17-07-01341, № 18-37-00271.

References

1. N. I. Voropay, *Rationale for the development of electric power systems: Methodology, models, methods, their use* .(Novosibirsk: "The science". 2015), p.448.
2. Cagnin C., *Future-Oriented Technology Analysis. Strategic Intelligence for an Innovative Economy*. (Springer, 2008), p.170.
3. Alexey V. Mikheev, Feasibility of tech mining for forecasting innovation pathways in energy sector, in *Information and mathematical technologies in science and management*, vol. 4(8), (2017) 166–176 ISSN 2513-0133). (In Russian).
4. Kopaygorodsky A.N., Searching for information and development of expert assessment of new technological solution in technical infrastructure, in *Proceedings of the 19th International Workshop on Computer Science and Information Technologies*. *Germany, Baden-Baden.* (USATU, Vol. 1. 2017, ISBN 978-5-1030-8, ISBN 978-4-4221-1031-5), pp. 71–76.
5. Khayrullina E. P., Kopaygorodsky A. N, Application of ontologies in the design and implementation of information systems, in *Informatization and visualization of economic and social life*, (2015) pp. 200-204.
6. Gaskova D. A., Analysis of cybersecurity violations in the energy sector, in *System research in the energy* (Irkutsk: ESI SB RAS, №. 47, 2017), 101–107.
7. Manoj K. S., Dileep K. G., *Effective big data management and opportunities for implementation* (Information Science Reference, America, 2016), p.440
8. Dumbill E. *Planning for Big Dat*a (O'Really Radar Team, America, 2012), p.78.
9. Woon, Wei Lee, Andreas Henschel, Stuart Madnick., *A Framework for Technology Forecasting and Visualization* (IEEE, 2009), pp. 155–159.
10. P. Willett., The Porter stemming algorithm: then and now in *Program: Electronic Library and Information Systems*, 2006, pp. 219–223.
11. Huilong Fan, Yongbin Qin., Research on Text Classification Based on Improved TF-IDF Algorithm in *The International Conference on Network, Communication, Computer Engineering*, 2018, pp. 501–506.