







Natural Language Search and Associative-Ontology Matching Algorithms Based on Graph Representation of Texts

Sergey Kuleshov , Alexandra Zaytseva  ,
and Alexey Aksenov 

Saint-Petersburg Institute for Informatics and Automation of RAS,
Saint-Petersburg, Russia
cher@ias.spb.su

Abstract. The ability to freely publish any information content is causing rapid growth of unstructured, duplicated and unreliable information volumes with irregular dynamics. This significantly complicates timely access to actual reliable information especially in the tasks of the specific scientific topics monitoring or when it is necessary to get quick insight of adjacent scientific fields of interest. The paper contains the description of the technology of text representation as a semantic graph. The algorithmic implementation of proposed technology in the tasks of fuzzy and exploratory information search is developed. The problems of current search technologies are considered. The proposed ontology-associative graph matching approach to post-full-text search system development is capable of solving the problem of document search under conditions of insufficient initial data for correct query formation.

The proposed graph representation of texts allows restricting usable ontology, which in turn gives the benefit of thematic localization of the search region in the field of knowledge.

Keywords: Ontology matching · Associative ontology approach · Natural language processing · Search algorithms · Graph representation

1 Introduction

Enormous data volumes accumulated and being progressively generated by digital society lead to the necessity of intensive development in the field of search technology. The ability to freely publish any information content is causing rapid growth of unstructured, duplicated and unreliable information volumes with irregular dynamics. This significantly complicates timely access to actual reliable information especially in the tasks of the specific scientific topics monitoring or when it is necessary to get quick insight of adjacent scientific fields of interest.

The current search technologies are mainly based on full-text search algorithms. The obvious feature of such approach – to find exact match of search query, becomes a disadvantage when user is not familiar enough with topic terminology to formulate the

correct search queries and thus the results of such search algorithm would be of doubtful relevance [1–3].

In the proposed approach a prepared corpus of texts and search index in given subject area are presented using an associative-ontological approach in the form of oriented loaded graphs representing relations of basic concepts [4–8], including the relationships between key parameters expressed numerically through the use of statistical analysis. This form of presentation is representative for the expert, in contrast to the forms of representation of knowledge obtained by machine learning methods that are “hidden” inside neural networks and are not available for their direct study and modification [9, 10]. The employment of the proposed form of semantic representation allows correction of the automatically generated ontologies using the tool of visualization. Many researches all over the world emphasize the importance of data visualization in different areas of scientific research [11–15].

Recent years the interest in semantic data processing technologies in specific subject areas using ontological models has not weakened, new systems are being created, actively using ontology matching and ontology mapping technologies. For example in the work [16] queries are processed and clustering by extracting content from text description using of NLP-technology. In [17] the mechanism of multi-criteria spatial semantic queries based on elements of ontology and dynamic construction of GeoSPARQL queries are developed. In [18, 19], technological innovations are considered as the way of optimization of the use of vital resources in social biological and economic systems. In order to manage technological innovations, an ontological structure is created and tested, which reflects the current body of knowledge in the subject area and allows identifying links and patterns.

To overcome the described problems, the various search technologies based on semantic search principles are being developed which could be referred as post-full-text search forming a new paradigm of exploratory information search.

This paradigm could be also characterized as “indirect information search” which allows using broadly defined topic as a search query. For example, topic could be defined as a sample document or a collection of documents. But as a new technology it encounters number of problems to be addressed. For the time being there is no unified standards for functionality and interface decisions for a such systems. Another disadvantage is a frequently unobvious or unexpected reaction to users query. This happens mostly in two general cases: the found documents are relevant to topic but contain unfamiliar terminology or documents are irrelevant topic but were found due to homonymy. The first case can give a positive result of new knowledge acquirement but the second case can confuse the user and give a negative result.

This paper proposes a variant of post-full-text search system based on the technology of representing texts as a semantic graph and considering the search task as a selecting of the documents with graph of text matching the graph of search query as well as the associative ontology matching method. The graph of text can be also referred as associative ontology – the ontology reflecting the associative dependencies between the elements of text [4–6].

2 The Algorithm for Text Search Based on Associative Ontology

Let's consider the case when associative ontology was obtained from corpus of texts. To bind corpus of texts to associative ontology the following structure will be used:

$$\left(\bigcup_i G_{i, D, K} \right), \tag{1}$$

where $D = \{d_1, d_2, d_3, \dots, d_n\}$ – the set of documents which are belong to corpus of texts, $K = \{k_1, k_2, k_3, \dots, k_n\}$ – the values calculated with use of ξ_{R2} for every document d_i . All the documents in D are then presented as united graph with set of edges $E_G = \bigcup_i E_{2i}$ and vertices $V_G = \bigcup_i V_{2i}$ consisting of graphs for every document d_i (Fig. 1).

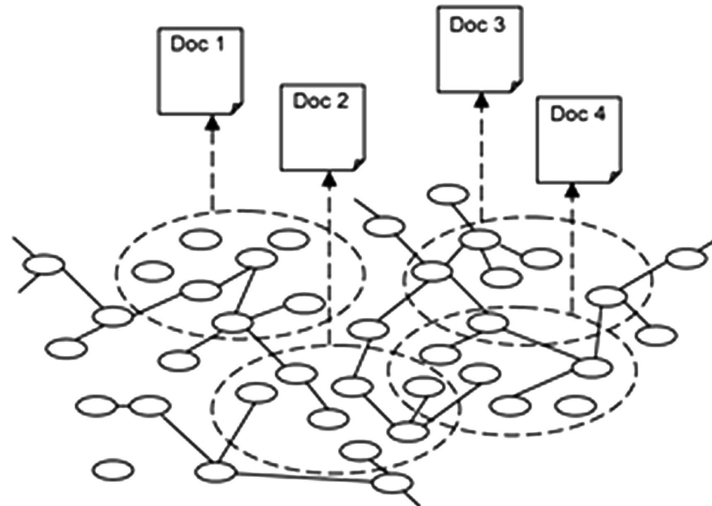


Fig. 1. The illustration of the document relation to semantic ontology region principle

We define the associative search as a process of determination which documents are containing the same semantic relations as search query [4, 7].

Each search query passes through the processes of stop-word removal and lemmatization $w_i \xrightarrow{m} \bar{w}_i$, then the graph with set of edges $E_Q \subset E_G$ is being created based on the methods that are represented in previous section. For the words w_i from query Q :

$$\forall \bar{w}_i \in Q \ \& \ \forall \bar{w}_j \in Q \Rightarrow e_Q(\bar{w}_i, \bar{w}_j) \in E_Q. \tag{2}$$

Formally the process of associative search can be defined as the process of getting document subset from corpus of texts: $\forall d, e_Q \in E_Q \vee e_Q \in E_d$. Wherein the number of matching elements e_Q can be considered as indicator of relevance.

The following cases are possible in search operation $E_R = E_Q \cap E_G$ on query Q:

- $|E_R| = |E_Q|$ – the documents containing all relations from query are exist in corpus;
- $|E_R| < |E_Q|$ – there is no document in corpus of texts containing all relations from query (the created associative ontology incomplete or the search query is incorrect;
- $|E_R| \equiv \emptyset$ – nothing is found on search query (associative ontology contains no relations satisfying search query).

The result of search query processing is a set of documents $\{d_i\}$, $E_{d_i} \cap E_Q \neq \emptyset$, which is being transformed into ordered list (Search Engine Results Page – SERP) by the application of ranking function. The search process is illustrated on Fig. 2.

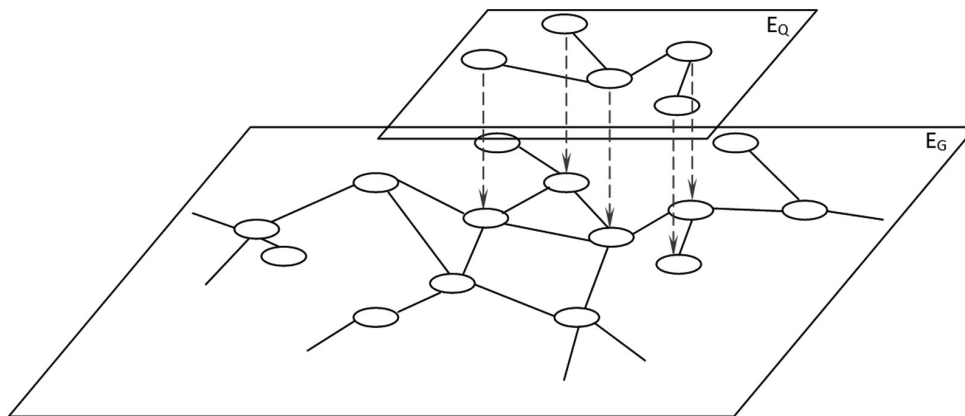


Fig. 2. The illustration of the search process based on associative ontology

3 Implementation

As an illustration of the search index structure representation described in previous section, the implementation based on relational model is shown. The corpus of texts can be represented within the framework of relational algebra by ER-diagram (Fig. 3). The set D corresponds to the document entity, the set corresponds to the link entity [8]. The set corresponds to the word entity.

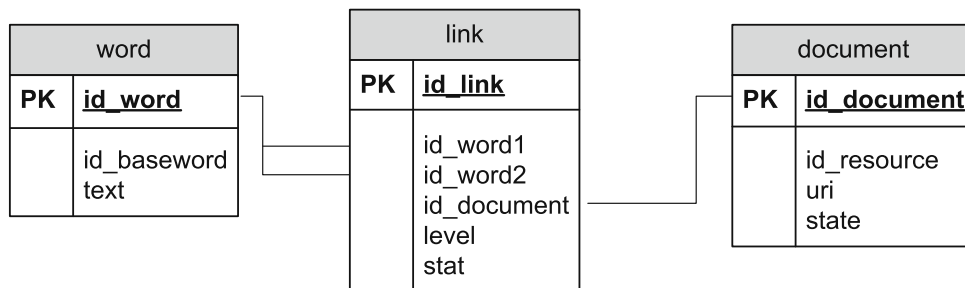


Fig. 3. The illustration of the search process based on associative ontology

The realization of search mechanism using database-management system (DBMS) is implemented as SQL-query. Generalized SQL-query for search phrase consisting of n words is as follows [8]:

```
select id_document from (
select id_document, count(*) as cnt from test where
W_CONDITION group by id_document
) where cnt>N;
where N=(n-1)!,
W_CONDITION="(((id_w1=id_word1)and(id_w2=id_word2))or((id
_w1=id_word2)and(id_w2=id_word1)))or(((id_w1=id_word1)and
(id_w2=id_word3))or((id_w1=id_word3)and(id_w2=id_word1)))
or..."
```

Thus the n -word query needs $4(n-1)!$ comparisons given the $W_CONDITION$ expression.

The speed of such query is limited only by the efficiency of DBMS which is dependent on the speed of hard drives containing DBMS data structures.

To speed up search operations, a number of approaches can be used, including NOSQL, “key-value” base, etc. for storing link table indexes (LINK entity).

The proposed realization gives the ability to execute search operations without the need to store copies of text documents and thus minimizes the demands for data storage.

4 Associative Ontology Matching

Ontology matching is the basis for fuzzy search algorithms based on the methods of entering sub-ontologies into ontology, for example for solving tasks of long-term monitoring in the field of interest.

In this paper we define E_A as the ontology built from analysis of texts for the period of time $t_1 + dt$, and E_B – built from analysis of texts for the period of time $t_2 + dt$, where $t_1 < t_2$. Then we can define, that

$$E_{\text{new}} = \frac{E_B}{E_A}$$

is the ontological graph structure, which corresponds to the new trends have appeared during the time period $t_2 - t_1$.

$$E_{\text{back}} = \frac{E_B}{E_A}$$

is the ontological graph structure, which corresponds to the concepts and phenomena obsolete during the same time period. Figure 4 illustrates the process of ontology matching.

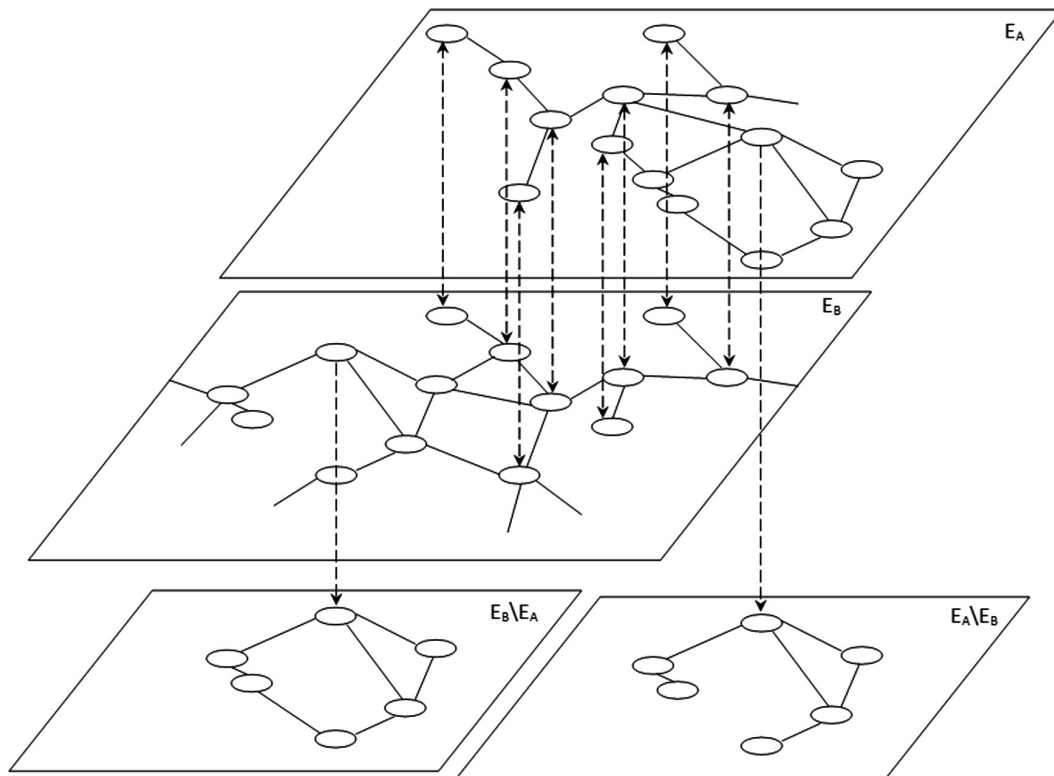


Fig. 4. The process of ontology matching when monitoring the temporary changes in the field of interest

The union of a collection of sets from available sources (anthology) forms the associative ontology E_G

The union of all the associative semantic environments of texts (associative ontologies) in the given field of interest give as the thematic area E_T :

$$\bigcup_i E_i = E_T,$$

where E_i is the associative environment of the text T_i (Fig. 5).

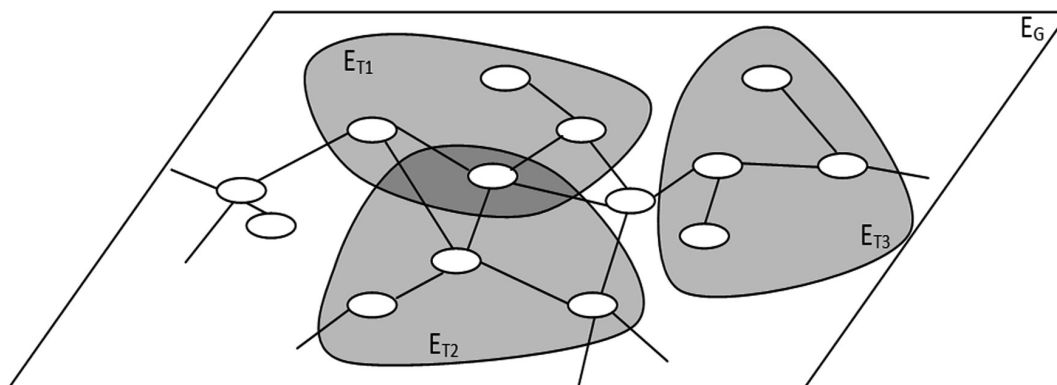


Fig. 5. The thematic areas illustration

We have to note that

$$E_T \subseteq E_G.$$

Figure 6 shows the common structure of method of ontology constructing, comparison and visualization.

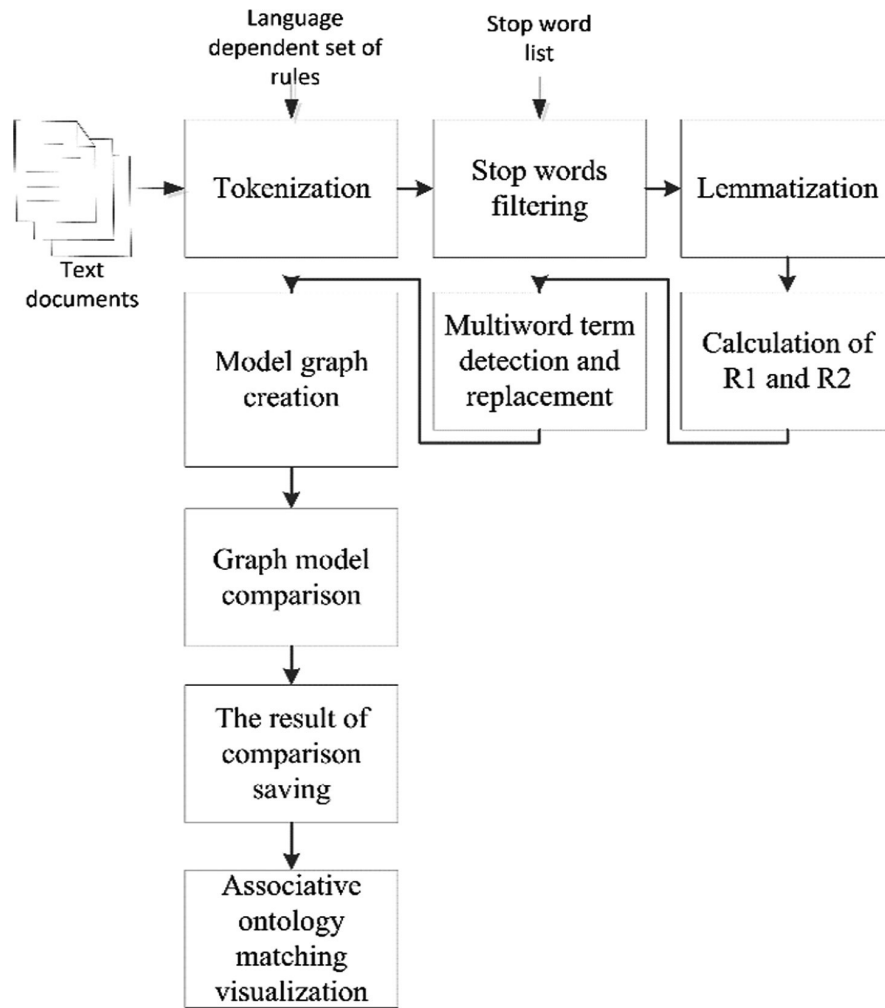


Fig. 6. The common structure of method of ontology constructing, comparison and visualization

The result of the developed algorithm of ontology matching and visualization of the resulting ontology is the structure formed by concatenation of three ontology graphs marked by three colors (red, green, white). The red graph nodes are the out-of-use terms (from E_{back} ontology), the green graph nodes are the new terms, which appeared in the result ontology because of diachrony, white nodes are the terms that present in both ontologies (the stable terms). The example of two ontologies associative matching is shown on Fig. 7.

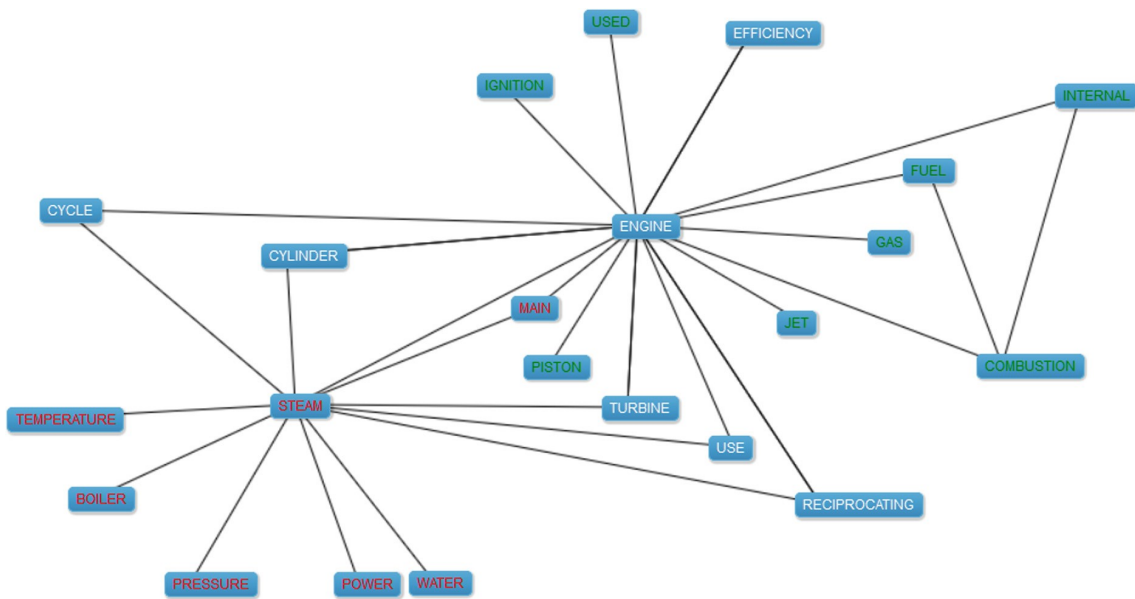


Fig. 7. Visualization of ontology matching result in the some field of interest

5 Conclusion

The proposed ontology-associative graph matching approach to post-full-text search system development is capable of solving the problem of document search under conditions of insufficient initial data for correct query creation. This is paradigm of exploratory information search. In this case, the minimal set of the search keywords allows formation of query by iterative adding of new keywords from query associative environment.

The proposed graph representation of texts allows restricting usable ontology, which in turn gives the benefit of thematic localization of the search region in the field of knowledge.

The valuable feature of graph matching approach is an ability to evaluate similarity of texts and overall text cohesion allowing for most significant parts of text detection and giving the user brief annotation of documents in the search results.

Another possible application of the associative search algorithm is the analysis of ontology changes over a given period. Within the framework of the subject domain under consideration, this makes it possible to create a prediction method based on the results obtained from ontology changes for the next time interval. Developing a technology to visualize the dynamics of changing the significance of concepts over time, studying causal relationships based on the significance of concepts and tools for automating them will further create an empirical model for predicting changes in a particular field of interest.

Acknowledgements. The research is partly supported by the RFBR, project N 16-29-12965\18 and by the budget 0073-2019-0005.

References

1. Kuznecova, Ju.M., Osipov, G.S., Chudova, N.V.: Intellectual analysis of scientific publications and the current state of science. *J. Large-Scale Syst. Control* **44**, 106–138 (2013). (in Russian)
2. Smirnov, A.V., Pashkin, M., Chilov, N., Levashova, T.: Agent-based support of mass customization for corporate knowledge management. *J. Eng. Appl. Artif. Intell.* **16**(4), 349–364 (2003)
3. Smirnov, A., Levashova, T., Shilov, N.: Patterns for context-based knowledge fusion in decision support systems. *J. Inf. Fusion* **21**, 114–129 (2015)
4. Kuleshov, S.V., Zaytseva, A.A., Markov, S.V.: Associative-ontological approach to natural language texts processing. *J. Intellect. Technol. Transp.* **4**, 40–45 (2015). (In Russian)
5. Zaytseva, A.A., Kuleshov, S.V., Mikhailov, S.N.: The method for the text quality estimation in the task of analytical monitoring of information resources. *J. SPIIRAS Proc.* **37**(6), 144–155 (2014). <https://doi.org/10.15622/sp.37.9>. (In Russian)
6. Mikhailov, S.N., Malashenko, O.I., Zaytseva, A.A.: The method for the infology analysis of patients complaints semantic content in order to organize the electronic appointments. *J. SPIIRAS Proc.* **42**(5), 140–154 (2015). <https://doi.org/10.15622/sp.42.7>. (In Russian)
7. Kuleshov, S., Zaytseva, A., Aksenov, A.: The tool for the innovation activity ontology creation and visualization. *Adv. Intell. Syst. Comput.* **763**, 292–301 (2019)
8. Kuleshov, S.V.: The development of automatic semantic analysis system and visual dynamic glossaries. Ph.D. (Tech) theses, Saint-Petersburg (2005). (in Russian)
9. Malagrino, L.S., Roman, N.T., Monteiro, A.M.: Forecasting stock market index daily direction: a bayesian network approach. *J. Expert Syst. Appl.* (2018). <https://doi.org/10.1016/j.eswa.2018.03.039>
10. Todd, A., Beling, P., Scherer, W., Yang, S.Y.: Agent-based financial markets: a review of the methodology and domain. In: *Proceedings of 2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (2016). <https://doi.org/10.1109/SSCI.2016.7850016>
11. Zakharova, A., Vekhter, E., Shklyar, A., Pak, A.: Visual modeling of multidimensional data. *J. Dyn. Syst. Mech. Mach.* **5**(1), 125–128 (2017). (in Russian)
12. Roshchina, M.K., Il'yashenko, O.Yu.: Data visualization as a management decision-making tool for retailers. In: *Materials of SPbPU Science Week Scientific Conference with International Participation*, pp. 112–114 (2016). (in Russian)
13. Wang, C., Ma, X., Chen, J.: Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. *J. Comput. Geosci.* **115**, 12–19 (2018). <https://doi.org/10.1016/j.cageo.2018.03.004>
14. Dew, R., Ansari, A.: Bayesian nonparametric customer base analysis with model-based visualizations. *J. Mark. Sci.* **37**(2), 216–235 (2018). <https://doi.org/10.1287/mksc.2017.1050>
15. Keim, D., Andrienko, G., Fekete, J., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: definition, process, and challenges. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4950, LNCS, pp. 154–175 (2008)
16. Zhang, N., Wang, J., Ma, Y., He, K., Li, Z., Liu, X.F.: Web service discovery based on goal-oriented query expansion. *J. Syst. Softw.* **142**, 73–91 (2018)

17. Abburu, S.: Ontology driven cross-linked domain data integration and spatial semantic multi criteria query system for geospatial public health. *Int. J. Semantic Web Inf. Syst.* **14**(3), 1–30 (2018)
18. Cancino, C.A., La Paz, A.I., Ramaprasad, A., Syn, T.: Technological innovation for sustainable growth: an ontological perspective. *J. Cleaner Prod.* **179**, 31–41 (2018)
19. Kondratyev, A.S., Aksyonov, K.A., Buravova, N.A., Aksyonova, O.P.: Cloud-based microservices to decision support. In: *International Conference on Ubiquitous and Future Networks, ICUFN*, July 2018, pp. 389–394 (2018). <https://doi.org/10.1109/ICUFN.2018.8437015>