The Formation of Morphological Matrix Based on an Ontology “Patent Representation of Technical Systems” for the Search of Innovative Technical Solutions

Dmitriy Korobkin, Sergey Fomenkov, Grigoriy Vereschak, Sergey Kolesnikov, Dmitriy Tolokin, and Alla G. Kravets

Abstract Today, the active development of technology leads to a huge increase in the amount of information in patent databases. In this regard, it is necessary to process this information and extract the most relevant data from the array of patents. This article discusses component extraction methods, technical problems, and solutions. Structured information from patents is stored in a structured form in an ontology. Information about system components is retrieved by querying the ontology. The paper also describes the grammar for extracting technical functions in the form of a “solution-problem”. The structure and components of the subsystem for allocating the functions of technical objects are described.

Keywords Technical systems · Patent · Ontology · Fact extraction

1 Introduction

With the development of the direction of the automated invention [1–5], CAI systems are being used more and more recently. Computer-Aided Invention is the search for innovative solutions using the computer. CAI systems are automated support systems and search for new technical solutions. The completeness of various knowledge bases and the completeness of ontologies of subject areas directly affect the success of support systems and the search for new technical solutions that find their application in the synthesis of new technical solutions. But one of the serious problems of CAI systems is the problem of updating the knowledge base since this process is rather difficult [6].

Scientific documents, patent documents, reference books can be the main sources of technical information to supplement existing knowledge bases. Patent documents

D. Korobkin (✉) · S. Fomenkov · G. Vereschak · S. Kolesnikov · D. Tolokin · A. G. Kravets
Volgograd State Technical University, Lenina Avenue, 28, Volgograd, Russia

A. G. Kravets
Dubna State University, Dubna, Moscow Region, Russia

A. G. Kravets et al. (eds.), *Cyber-Physical Systems*, Studies in Systems, Decision and Control 350, https://doi.org/10.1007/978-3-030-67892-0_13

can be considered one of the main sources of technical information since the number of patents in patent databases is quite large.

The existing more than 20 million worldwide patent database can act as a source of information for the initial stages of designing new technical solutions. Such volumes of data require automated processing.

One of the convenient ways of conceptualized knowledge representation about any subject area is the ontology model. Ontologies are a convenient organization of stored knowledge, thanks to which you can search and analyze data. Considering that the array of patent documents contains a lot of information useful for extraction and analysis, such as claims, classifications, country of origin, organization; ontologies provide the ability to structure and link information.

2 Analysis of the Patent Array

A patent document is a document issued by an authorized public authority confirming the exclusive right of the patent holder to an invention, utility model, or industrial design. One of the most useful for analysis is the patent claims, which are part of the specification of the patent document. The International Patent Classification (IPC) is a vehicle for internationally uniform classification of patent documents. This paper deals with patents belonging to the classes of electricity and mechanical engineering, that is, classes H and F, respectively.

In this study, as morphological features, which are concepts of ontologies of the subject areas “Technical functions” and “Implementation of technical objects”, the technical implementation and the structure “problem-solution” are highlighted. The technical implementation determines the constructive composition of the invention, and the problem-solution structure expresses the problem solved by the technical implementation. The source of data for the first feature is the claims of the device, and for the second—the item of the technical result in the description section of the invention.

Using the SAO (Subject-Action-Object) [7–9] model, technical implementations of objects can be represented, and the problem-solution structure is an incomplete part of the model. Morphological features of technical objects from patent documents can be represented by certain syntactic constructions that can be used for the automated construction of ontologies.

The main methods for extracting concepts and relationships between concepts for building domain ontologies are dependency parsing and part-of-speech tagging [10–15]. The SAO model is used to represent the implementations of technical objects and technical functions. To extract concepts from the claims of a patent document, the latest version of the Stanford NLP called Stanza [16] is used. An example of concept extraction is shown in Fig. 1

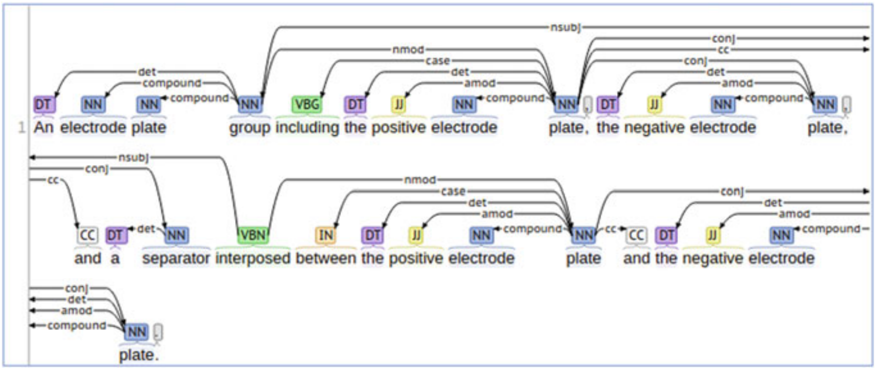
nsubj
nmod
conj
cc
case
det
compound
amod
DT
NN
VBG
JJ
1
An electrode plate
group including the positive electrode
plate, the negative electrode
plate,
CC
VBN
IN
and a
separator interposed between the positive electrode
plate
and the negative electrode
plate.

Fig. 1 Example of fact extraction

3 Developed Methods for Extracting Information from the Patent Array

Features of patent presentation of technical systems:

- Descriptions of realizations of technical objects are contained in the invention formula;
- The technical problem solved by the device (device from the name of the patent) is contained in the first paragraph of the summary of the patent.

Before proceeding to parse patent documents containing descriptions of implementations of technical devices, it is necessary to perform preliminary processing of the patent array, which is an XML file. The filtering of patents is carried out by classes H and F, which correspond to electricity, mechanical engineering, etc.

To search for and retrieve realizations of technical objects [17–20], the claims are analyzed. The first claim is the most generalized and contains the most complete description of the device, and it is he who is being analyzed.

3.1 Pre-segmentation Algorithm

The main idea of preparing the segments of the first paragraph of the formula is to "restore" sentences for correct analysis by the stanza parser. In Example 1, you can see a fragment of the first claim of the invention in its original form.

For the left side of the claims, the main device is searched. The claims begin with the main device, followed by the sequence of characters "comprising:". To restore the segments the left part is taken up to the ":" character and the right part containing the enumeration is split by the ";" character. At the beginning of each segment representing an enumerated element, a substring is added containing the main unit of the patent claims.

Each penultimate enumerated element has after the “;” the conjunction “and”, which can complicate the parsing of the sentence. Therefore, in the first claim, the combination of symbols “; and” is replaced with“; ”, after which the first occurrence of the word “ where ”is searched for. The claims are divided into two parts—before and after. If “where” is absent, then the whole formula is taken. Since there can be several “where”, then the part of the formula after the first mention of “where” is broken down by “where”, and for each resulting segment, whitespace characters are removed from the beginning and end of the segment.

Example 1. A fragment of the first claim of the invention

```
<claim-text> 1. A decoupled gas turbine engine
comprising:
  <claim-text> a low pressure compressor; </claim-text>
  <claim-text> a high pressure compressor;
</claim-text>
  …
  <claim-text> a second turning duct in fluid
communication between the combustor and the high
pressure turbine; </claim-text>
  <claim-text> where the low pressure compressor and
the low pressure turbine …
```

After preliminary segmentation, the first claim will have the form shown in Example 2.

Example 2. View of the first claim after preliminary segmentation

```
A decoupled gas turbine engine comprising a low
pressure compressor.
  A decoupled gas turbine engine comprising a high
pressure compressor.
```

3.2 SAO Extraction Algorithm

A global list of extracted SAOs is used to store and write retrieved device components in the form of an SAO model. For each pre-segmented segment, all SAOs are retrieved. The input segment is split into a sequence of tokens using a parser. Only those segments with tokens that contain key verbs typical for extracting the implementation of technical objects are subject to processing. Key verbs include the following: comprise, consist, connect, include, attach, have. The extraction of technical realizations should be continued until there are no unprocessed key verbs in the segment. Figure 2 shows an algorithm for extracting realizations of technical objects from the claims.

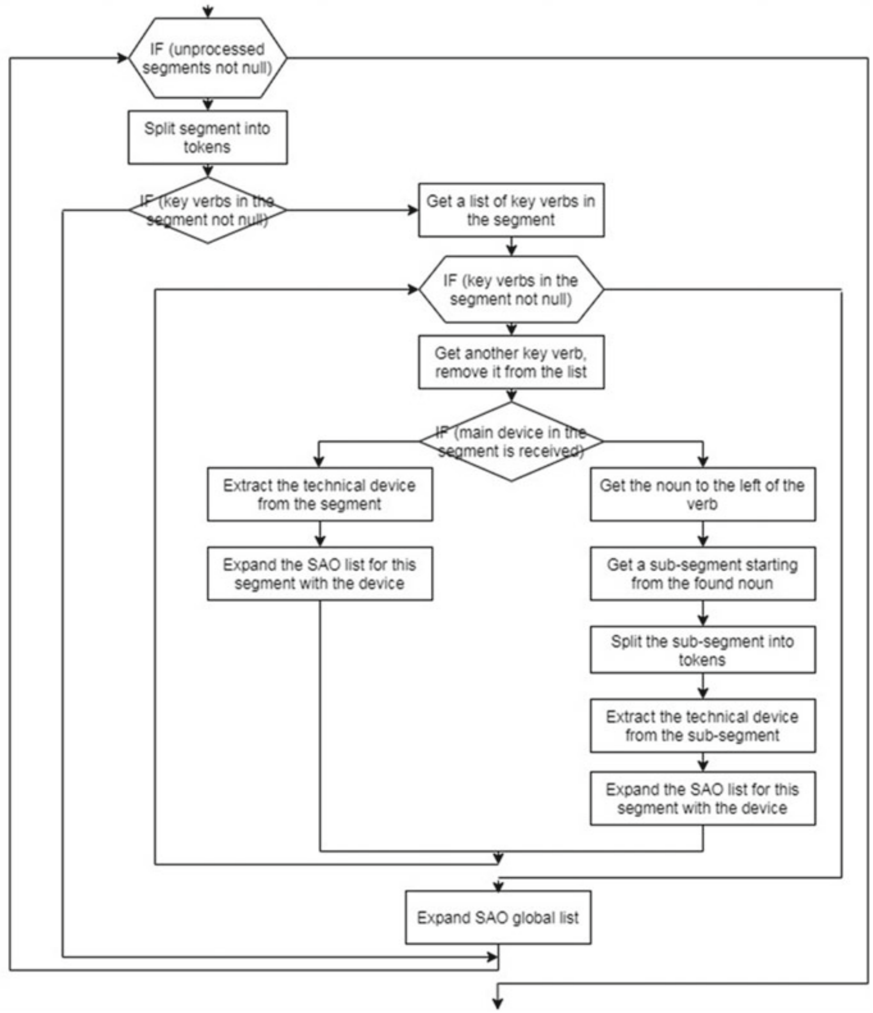
IF (unprocessed segments not null)
Split segment into tokens
IF (key verbs in the segment not null)
Get a list of key verbs in the segment
IF (key verbs in the segment not null)
Get another key verb, remove it from the list
IF (main device in the segment is received)
Extract the technical device from the segment
Expand the SAO list for this segment with the device
Get the noun to the left of the verb
Get a sub-segment starting from the found noun
Split the sub-segment into tokens
Extract the technical device from the sub-segment
Expand the SAO list for this segment with the device
Expand SAO global list

Fig. 2 Algorithm for extracting realizations of technical objects from the claims

Dependency parsing and parts of speech detection are used to directly extract the technical implementation. The algorithm for extracting technical implementation assumes the presence of a potential key vowel, for which it is necessary to find a subject and an object.

Figure 3 shows a detailed algorithm for extracting a specific implementation of a technical object.

An example of extracting the implementation of technical objects is shown in Fig. 4.

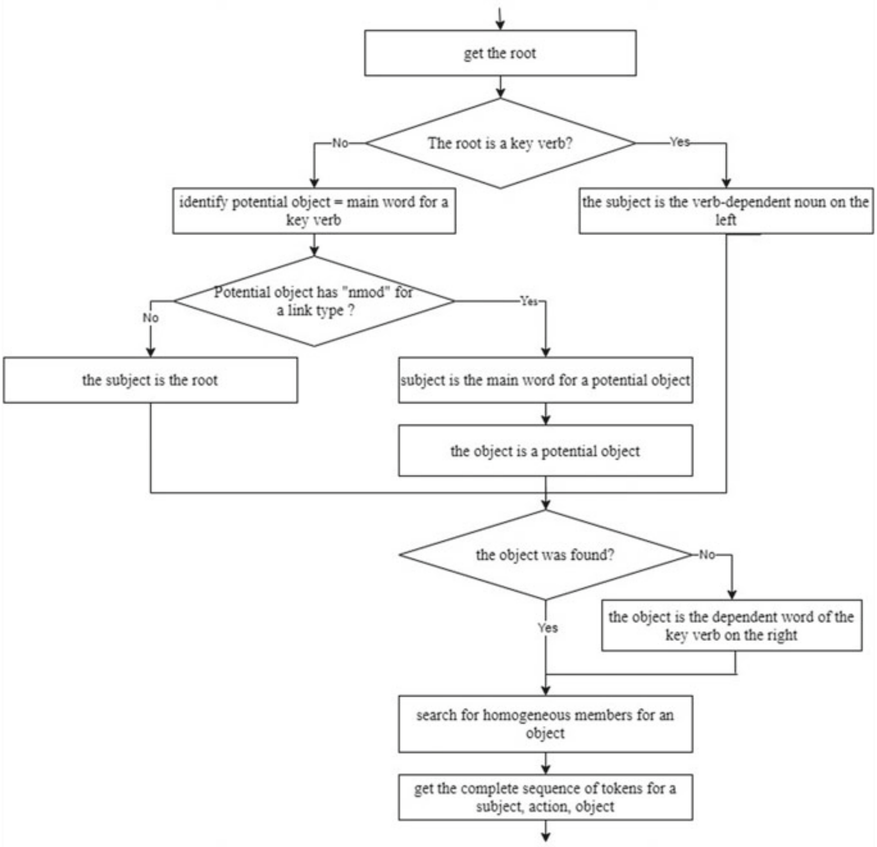
get the root
The root is a key verb?
No
Yes
identify potential object = main word for a key verb
the subject is the verb-dependent noun on the left
Potential object has "nmod" for a link type ?
No
Yes
the subject is the root
subject is the main word for a potential object
the object is a potential object
the object was found?
No
the object is the dependent word of the key verb on the right
Yes
search for homogeneous members for an object
get the complete sequence of tokens for a subject, action, object

Fig. 3 Algorithm for extracting the implementation of a technical object

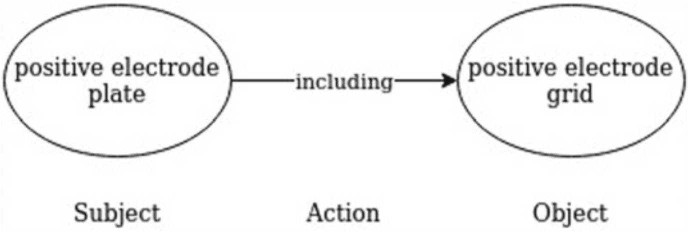

Fig. 4 Example of the method implementation

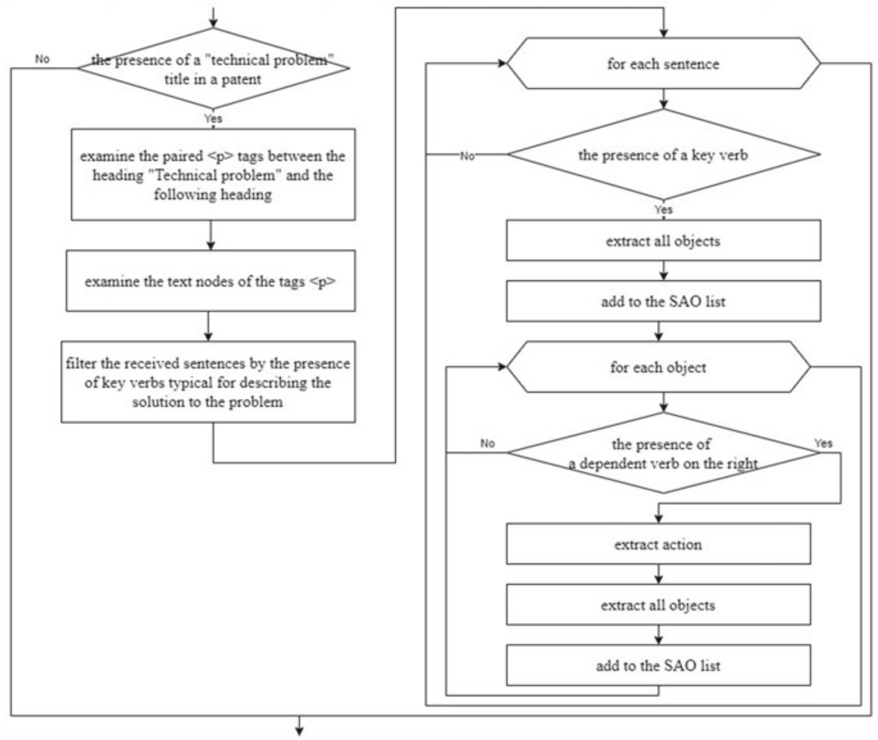
the presence of a "technical problem" title in a patent
No
Yes
examine the paired <p> tags between the heading "Technical problem" and the following heading
examine the text nodes of the tags <p>
filter the received sentences by the presence of key verbs typical for describing the solution to the problem
for each sentence
the presence of a key verb
No
Yes
extract all objects
add to the SAO list
for each object
the presence of a dependent verb on the right
No
Yes
extract action
extract all objects
add to the SAO list

Fig. 5 Algorithm for extracting the problem of the device and technical functions to be solved

To extract technical functions and the problem to be solved, the device does not analyze the patent formula, but the section of the patent with the title “Technical Problem”. Figure 5 shows an algorithm for extracting the device problem to be solved and technical functions.

4 The Ontology

Triplets are the main way of expressing information in ontologies. A triplet consists of three components—subject, predicate, and object. This model is ideal for storing retrieved realizations of technical objects as SAO. So, a triplet will consist of three components—subject, action, object.

In Fig. 6 you can see the class diagram of the ontology of the subject areas “Technical functions” and “implementations of technical objects”.

The following properties of objects were selected:

- hasFunction—a property for linking a technical function and a component;

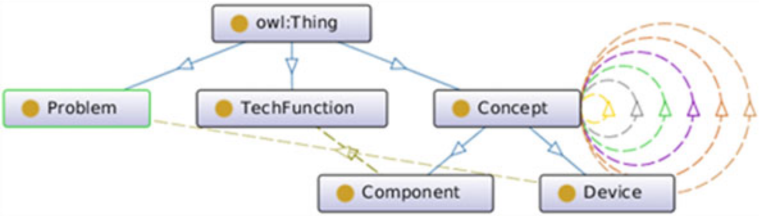

Fig. 6 Scheme of classes of the ontology of the subject areas "Technical functions" and "implementation of technical objects"

- comprises—a property for communication between the components of a device (the verb "comprise");
- connectedTo—property for communication between device components (verbs "connect", "attach");
- consists—a property for communication between device components (the verbs "consists", "include");
- parentFor—indication of the presence of a parent relationship between elements (the verb "have");
- partOf—indication of the belonging of the component to the device of the patent document;
- solutionFor—property for linking the problem and the device being solved by it;
- connected_to—connection between elements (verbs "install", "connect", "connect", etc.).

Figure 7 shows the ontology replenishment algorithm.

The resulting ontology is exported to an OWL file, which can then be opened for further work in Protege.

5 The Software

The automated system is implemented as a desktop application for Linux operating systems. Development was carried out on the Ubuntu 18.04.4 operating system. The system is implemented in the Python 3.6.9 programming language. The PyQt5 library was used to create the user interface. For the analysis of natural language texts, the latest version of Stanford NLP called Stanza was used. The MySQL DBMS was used to store the extracted SAOs, and the Python PyMySQL library was used for development. XML files were parsed using the lxml library. The Owlready2 library was used to work with ontologies.

The automated system allows you to download patent documents, extract technical functions and implementations of technical objects, display the extracted implementations of technical objects in a form, build ontologies for a user-selected patent, as well as for all uploaded patents for which technical functions and technical object

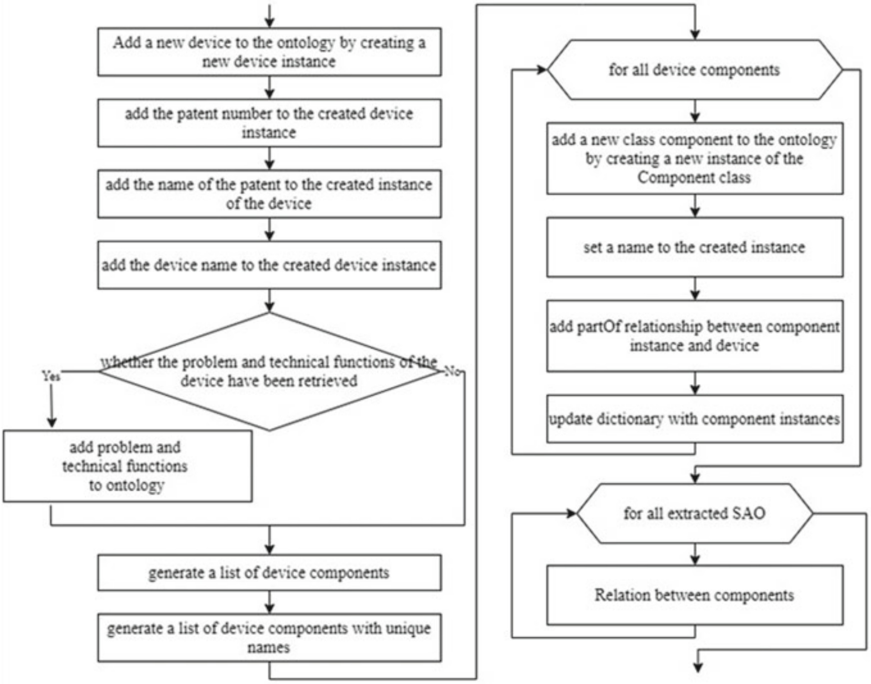
Add a new device to the ontology by creating a new device instance
add the patent number to the created device instance
add the name of the patent to the created instance of the device
add the device name to the created device instance
whether the problem and technical functions of the device have been retrieved
Yes
No
add problem and technical functions to ontology
generate a list of device components
generate a list of device components with unique names
for all device components
add a new class component to the ontology by creating a new instance of the Component class
set a name to the created instance
add partOf relationship between component instance and device
update dictionary with component instances
for all extracted SAO
Relation between components

Fig. 7 Algorithm for replenishing the ontology of the subject areas "Technical functions" and "implementation of technical objects"

implementations have been extracted. Figure 8 shows the constructed ontology for one patent document.

As a computational experiment, patent documents were manually sorted out, the number of SAO retrieved for each patent and the time taken to parse each patent document were recorded. The extraction accuracy (P) was calculated using the formula (1)

$$P = \frac{\mathrm{E}}{\mathrm{N}}, \tag{1}$$

where E is the number of correctly extracted by the SAO system, N is the number of SAO in the patent document.

In Table 1 you can see the results of the experiment.

The average time for parsing one patent by the system was 1.72316 s, the average time for parsing one patent by an expert was 46.6 s. Accuracy rates are above 70%.

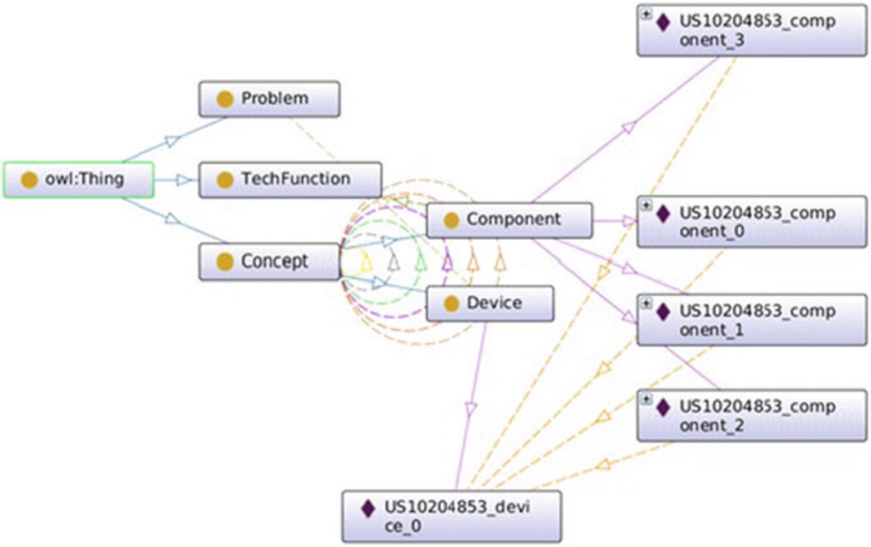

Fig. 8 The ontology for one patent document

Table 1 Results of the experiment

Experiment number	Time spent processing by the system, c	Time spent on processing by an expert, c	Extraction accuracy, %
1	1.324552	47.0	85.7
2	2.366441	54.0	100.0
3	2.608219	51.0	76.5
4	0.656219	38.0	100.0
5	1.660411	43.0	71.4

6 Discussion

This work solved the general problem of information support for the synthesis of new technical solutions based on the analysis of USPTO patents.

As concepts of the ontology of subject areas, the structural elements of a technical object (TO) and the relationship between them, as well as descriptions of the problems solved by the invention, were considered. The first claim of the patent document acted as the main source of information. The unit of extraction was the semantic structures SAO (Subject-Action-Object).

The main linguistic features of patent documents were identified. The method of preliminary processing of the patent mass has been formed. A separate auxiliary tool has been developed for the preliminary processing of the patent array. An algorithm for extracting SAO from the patent formula has been formed. A method has been developed for exporting extracted SAOs from English-language patents to a domain ontology.

The developed methods were tested on US patent documents

Acknowledgements The reported study was funded by RFBR (research projects 18-07-01086, 19-07-01200), RFBR and Administration of the Volgograd region (projects 19-47-340007, 19-41-340016).

References

1. Arel, E.: Goldfire Innovator. Volume II: Patents and Innovation Trend Analysis User Guide. Invention Machine Corporation, Boston, MA (2004)
2. Arel, E., Verbitsky, M., Devoino, I., Ikovenko, S.: TechOptimizer Fundamentals. Invention Machine Corporation, Boston, MA (2002)
3. Zlotin, B., Zusman, A.: Directed Evolution: Philosophy, Theory and Practice. Farmington Hills, Ideation International (2001)
4. Fey,V., Rivin, E.: Innovation on Demand: New Product Development Using TRIZ. Cambridge University Press, Cambridge (2005)
5. Souili, A., et al. Starting from patents to find inputs to the problem graph model of IDM-TRIZ. Procedia Eng. **131**, 150–161. https://doi.org/10.1016/j.proeng.2015.12.365 (2015)
6. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
7. Choi, S., et al. SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. Scientometrics, 863–883. https://doi.org/10.1007/s11192-011-0420-z (2011)
8. Guo, J., et al.: Subject–action–object-based morphology analysis for determining the direction of technological change. Technol. Forecasting Soc. Change **105**, 27–40 (2016)
9. Yufeng, D., Duo, J., Lixue, J.: Patent similarity measure based on SAO structure. Chin. Sentence Clause Text Inf. Process. **30**(1), 30–36 (2016)
10. Asiryan, A.K.: Morphological tagging tools comparison. Intellectual potential of the XXI century 2017, November. https://www.sworld.com.ua/konferu7-317/27.pdf (2017)
11. Mel'čuk, I.: Dependency Syntax Theory and Practice. SUNY, New York (1988)
12. Link Grammar Parser. http://www.abisource.com/projects/linkgrammar (2020)
13. MaltParser. http://maltparser.org/ (2018). Accessed 26 Oct 2020
14. UFAL UDPipe. http://ufal.mff.cuni.cz/udpipe (2020). Accessed 26 Oct 2020
15. CoNLL-U Format. https://universaldependencies.org/format.html (2020). Accessed 26 Oct 2020
16. Stanza. https://stanfordnlp.github.io/stanza/ (2020). Accessed 26 Oct 2020
17. Korobkin, D., Shabanov, D., Fomenkov, S., Golovanchikov, A: Construction of a matrix «Physical Effects – Technical Functions» on the base of patent corpus analysis. In: Creativity in Intelligent Technologies and Data Science (CIT&DS 2019), pp. 52–68 (Ser. Communications in Computer and Information Science (CCIS); Volume 1084) (2019)
18. Vasilyev, S., Korobkin, D., Kravets, A., Fomenkov, S., Kolesnikov, S.: Extraction of cyber-physical systems inventions' structural elements of Russian-language patents. In: Cyber-Physical Systems: Advances in Design & Modelling, pp. 55–68. https://link.springer.com/book/10.1007/978-3-030-32579-4#toc (Book ser. Studies in Systems, Decision and Control (SSDC); vol. 259) (2020)

19. Fomenkova, M., Korobkin, D., Kravets, A., Fomenkov, S.: Extraction of knowledge and processing of the patent array. In: Creativity in Intelligent Technologies and Data Science (CIT&DS 2019), pp. 3–14 (Ser. Communications in Computer and Information Science (CCIS); Volume 1084) (2019)
20. Vayngolts, I., Korobkin, D., Fomenkov, S., Kolesnikov, S.: The Software and Information Complex Which Uses Structured Physical Knowledge for Technical Systems Design / Creativity in Intelligent Technologies and Data Science (CIT&DS 2019), pp. 42–51 (Ser. Communications in Computer and Information Science (CCIS); Volume 1084) (2019)