УДК 004.89 DOI: 10.26102/2310-6018/2020.31.4.007

Метод формирования онтологии предметной области «Патентное представление технических систем» для поиска инновационных технических решений

Г.А. Верещак, Д.М. Коробкин, С.А. Фоменков, М.А. Фоменкова, С.Г. Колесников
Волгоградский государственный технический университет, Волгоград, Российская Федерация

Резюме: В данной работе решалась одна из самых насущных проблем синтеза новых технических решений – автоматизированное формирование информационного обеспечения на основе анализа патентов USPTO. С развитием направления автоматизированного изобретательства в последнее время все больше используются CAI-системы (Computer-Aided Invention). Наполненность баз знаний и полнота онтологий предметных областей напрямую влияет на успешность работы CAI-систем. Цель работы заключалась в разработке метода автоматизированного формирования онтологии предметной области «Патентное представление технических систем» для поиска инновационных технических решений. В качестве концептов онтологии предметной области «Патентное представление технических систем» рассматривались элементы конструкции технического объекта (ТО) и связи между ними, а так же описания решаемых изобретением проблем. Первый пункт формулы изобретения патентного документа выступал в качестве основного источника информации. Единицей извлечения являлись семантические структуры SAO (Subject-Action-Object). Были определены основные лингвистические особенности патентных документов. Сформированы методы предварительной обработки патентного массива, извлечения SAO из формулы патента, экспорта извлеченных SAO в онтологию предметной области. Разработанные методы были апробированы на патентных документах США. Среднее время разбора одного патента автоматизированной системой составляет 1.72316 секунды, показатели точности извлечения информации из текста патента - выше 70%.

Ключевые слова: технические системы, патенты, онтология, извлечение информации, SAO, проблема, решение

Для цитирования: Верещак Г.А., Коробкин Д.М., Фоменков С.А., Фоменкова М.А., Колесников С.Г. Критерии многопараметрического ранжирования жестких дисков по риску отказа. *Моделирование, оптимизация и информационные технологии.* 2020;8(4). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=853 DOI: 10.26102/2310-6018/2020.31.4.007

Method for forming ontology “Patent representation of technical systems” for creating innovative technical systems

G.A. Vereshchak, D.M. Korobkin, S.A. Fomenkov, M.A. Fomenkova, S.G. Kolesnikov
Volgograd State Technical University, Volgograd, Russian Federation

Abstract: In this work, one of the most pressing problems of the synthesis of new technical solutions was solved - the automated generation of information support based on the analysis of USPTO patents. As concepts of the ontology of the subject area "Patent representation of technical systems", 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. Methods for preprocessing the patent array, extracting SAO from the patent formula, exporting extracted SAOs to

the domain ontology have been formed. The developed methods have been tested on US patent documents. The average time for parsing one patent by an automated system is 1.72316 seconds, the accuracy of extracting information from the text of a patent is over 70%.

Keywords: technical systems, patents, ontology, fact extraction, SAO, problem, solution

For citation: Vereshchak G.A., Korobkin D.M., Fomenkov S.A., Fomenkova M.A., Kolesnikov S.G. Method for forming ontology “Patent representation of technical systems” for creating innovative technical systems. *Modeling, optimization and information technology*. 2020;8(4). Available from: https://moitvivt.ru/ru/journal/pdf?id=853 DOI: 10.26102/2310-6018/2020.31.4.007 (In Russ).

Введение

С развитием направления автоматизированного изобретательства в последнее время все больше используются CAI-системы (Computer-Aided Invention). CAI-системы представляют собой автоматизированные системы поддержки и поиска новых технических решений [1]. Наполненность баз знаний [2] и полнота онтологий предметных областей напрямую влияет на успешность работы CAI-систем. Таким образом, одной из серьезных проблем синтеза новых технических систем является отсутствие решений для автоматизации пополнения базы знаний.

Существующий более чем 20-миллионый мировой патентный массив может выступать в качестве источника информации для начальных этапов проектирования новых технических решений. Такие большие объемы данных нуждаются в автоматизированной обработке.

Одним из удобных способов концептуализированного представления знаний о какой-либо предметной области является модель онтологии. Онтологии представляют собой удобную организацию хранимых знаний, благодаря которой можно выполнять поиск и анализ данных. Учитывая, что массив патентных документов содержит множество полезной для анализа информации (описание, формула изобретения, авторы и т.д.), то онтологии предоставляют возможность структуризации и связывания информации.

Цель работы – разработка метода автоматизированного формирования онтологии предметной области «Патентное представление технических систем» для поиска инновационных технических решений.

Материалы и методы

Анализ патентного массива

Патент представляет собой документ, выдающийся уполномоченным органом государственной власти, подтверждающий исключительное право патентообладателя на изобретение, полезную модель либо на промышленный образец. Одной из наиболее полезных для анализа является формула изобретения патента, представляющую собой одну из частей спецификации патентного документа. Международная патентная классификация (МПК) является средством для единообразного в международном масштабе классифицирования патентных документов. В данной работе рассматриваются патенты, принадлежащие к классам электричество (H) и машиностроение (F).

В настоящем исследовании в качестве морфологических признаков [3], являющимися концептами онтологий предметных областей «Технические функции» [4] и «Реализации технических объектов» [5], определены 1) техническая реализация устройства и 2) структура «Проблема-Решение». Техническая реализация определяет конструктивный состав изобретения, а структура «Проблема-Решение» выражает решаемую технической реализацией проблему. Источником данных для первого

признака рассматривается формула изобретения устройства, а для второго – пункт технического результата в разделе описания изобретения.

С помощью модели SAO (Subject-Action-Object) [6] можно представить технические реализации объектов и структуру «Проблема-Решение» (S – решение, AO – проблема). Морфологические признаки технических объектов из патентных документов можно представить определенными синтаксическими конструкциями, которые могут быть использованы для автоматизированного построения онтологий.

Основными методами для извлечения концептов и отношений между концептами для построения онтологий предметных областей являются парсинг зависимостей и тегирование частями речи. Для представления реализаций технических объектов и технических функций используется модель SAO. Для извлечения концептов (пример представлен на Рисунке 1) из формулы изобретения патентного документа используются последняя версия Stanford NLP с названием Stanza [7].

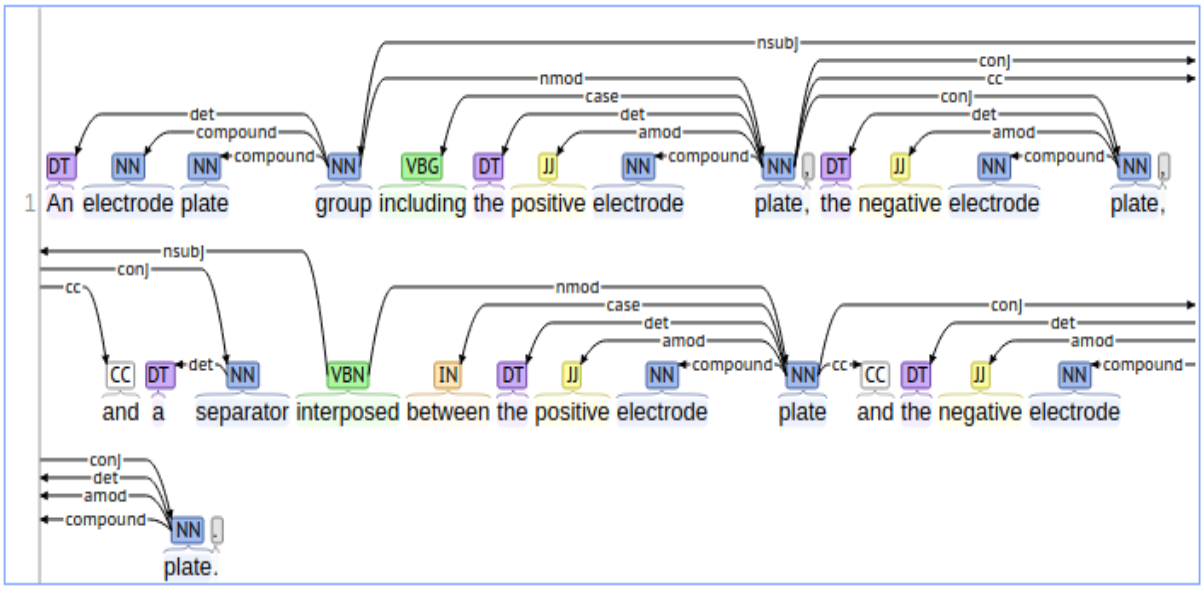

Рисунок 1 - Пример извлечения концептов из патентного документа
Figure 1 - Example of the extraction of concepts from the patent document

1. Разработанные методы извлечения информации из патентного массива

Особенности патентного представления технических систем: Описания реализаций технических объектов содержатся в формуле изобретения; решаемая устройством (устройство из названия патента) техническая проблема содержится в первом абзаце краткого содержания патента.

Прежде чем приступить к парсингу патентных документов, содержащих описания реализаций технических устройств, необходимо произвести предварительную обработку патентного массива [8], представляющего собой XML-файл. Фильтрация патентов осуществляется по классам H и F, которые соответствуют электричеству и машиностроению.

Для поиска и извлечения реализаций технических объектов производится анализ формулы изобретения. Первый пункт формулы изобретения наиболее обобщенный и содержит наиболее полное описание устройства, и именно он подвергается анализу.

Алгоритм предварительной сегментации. Основная идея подготовки сегментов первого пункта формулы состоит в «восстановлении» предложений для корректного анализа парсером Stanza. В примере 1 можно увидеть фрагмент первого

пункта изобретения в исходном виде.

Для левой части формулы изобретения выполняется поиск основного устройства. Формула изобретения начинается с основного устройства, после которого следует «comprising:». Для восстановления сегментов берется левая часть до символа «:», а правая часть, содержащая перечисление, разбивается по символу «;». В начало каждого сегмента, представляющего собой перечисляемый элемент, добавляется подстрока, содержащая главное устройство формулы патента.

Каждый предпоследний перечисляемый элемент имеет после знака «;» союз «and», который может усложнить разбор предложения. Поэтому в первом пункте формулы изобретения комбинация символов «; and» заменяется на «;», затем выполняется поиск первого упоминания ключевого слова «wherein». Формула изобретения разбивается на две части – до данного ключевого слова и после. Если «wherein» отсутствует, то берется формула целиком. Поскольку «wherein» может быть несколько, то часть формулы после первого упоминания данного ключевого слова разбивается и для каждого полученного сегмента удаляются пробельные символы из начала и конца сегмента.

Пример 1. Фрагмент первого пункта формулы изобретения:

<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> wherein the low pressure compressor and the low pressure turbine are rotatable about a first axis …</claim-text>

После предварительной сегментации первый пункт формулы изобретения будет иметь вид, представленный в примере 2.

Пример 2. Вид первого пункта формулы изобретения после предварительной сегментации:

A decoupled gas turbine engine comprising a low pressure compressor.

A decoupled gas turbine engine comprising a high pressure compressor.

…

the low pressure compressor and the low pressure turbine are rotatable about a first axis...

Алгоритм извлечения SAO. На Рисунке 2 можно увидеть алгоритм извлечения реализаций технических объектов из формулы изобретения.

Для хранения и записи извлеченных компонентов устройства в форме модели SAO используется глобальный список извлеченных SAO. Для каждого полученного в результате предварительной сегментации сегмента выполняется извлечение всех SAO. Входной сегмент разбивается при помощи парсера на последовательность токенов. Обработке подлежат только те сегменты, среди токенов которых есть ключевые глаголы, характерные для извлечения реализации технических объектов. К ключевым глаголам относятся: comprise, consist, connect, include, attach, have. Извлечение технических реализаций необходимо продолжать до тех пор, пока в сегменте не останется необработанных ключевых глаголов.

Для непосредственного извлечения технической реализации используется парсинг зависимостей и определение частей речи. Алгоритм извлечения технической реализации предполагает наличие потенциального ключевого глагола, для которого необходимо найти субъект и объект.

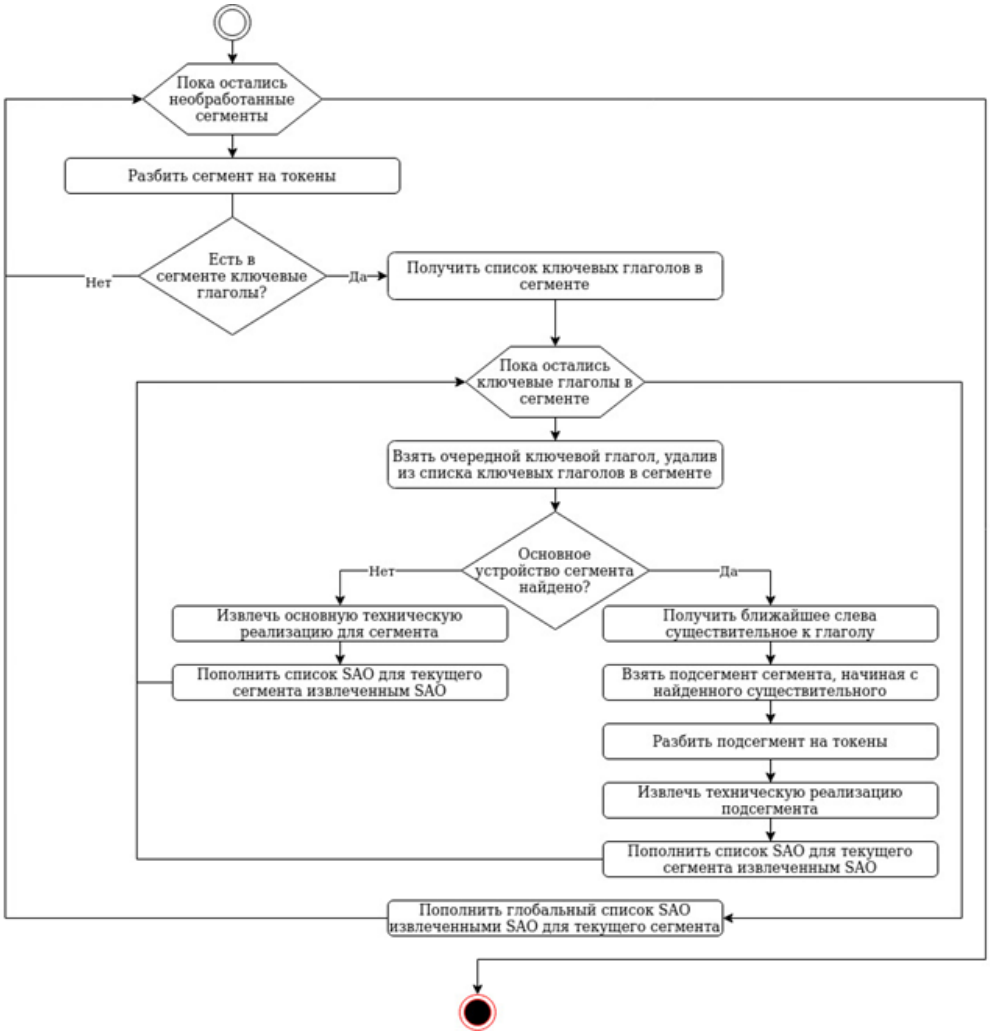

Рисунок 2 - Алгоритм извлечения реализаций технических объектов из формулы изобретения
Figure 2 - Algorithm for extracting technical object implementations from the claims

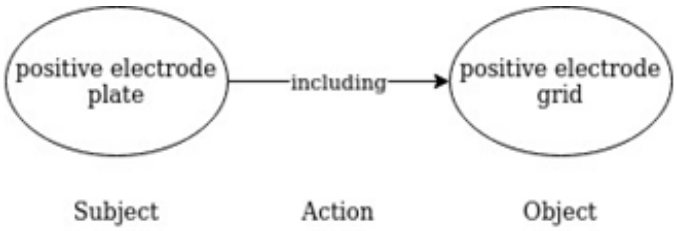

Рисунок 3 - Пример реализации метода
Figure 3 - Example of the method implementation

Пример извлечения реализации технических объектов приведен на Рисунке 3.

На Рисунке 4 приведен детальный алгоритм извлечения конкретной реализации технического объекта.

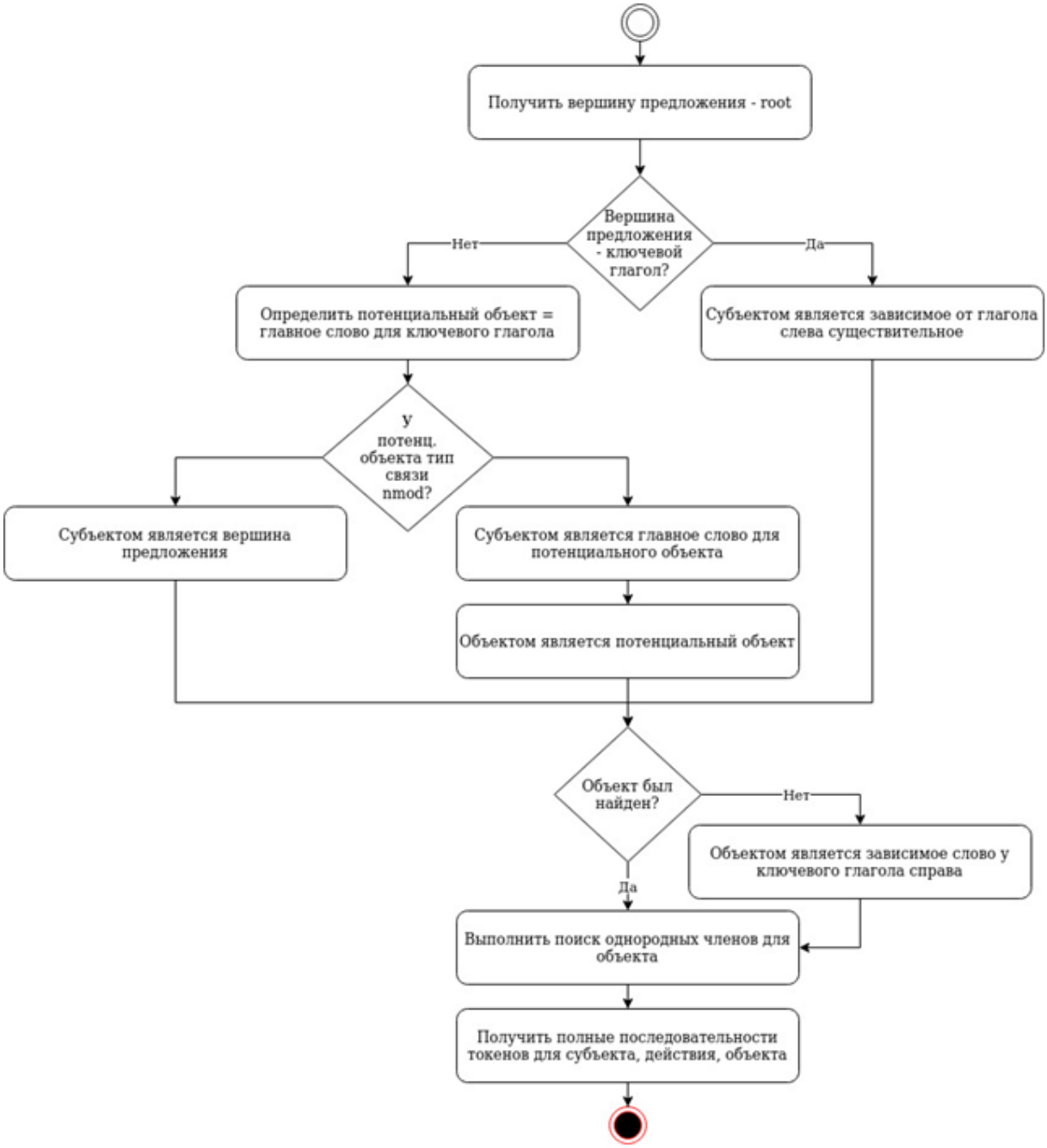

Рисунок 4 - Алгоритм извлечения реализации технического объекта
Figure 4 - Algorithm for extraction of technical object implementation

Для извлечения решаемой устройством технической проблемы (технических функций) анализируется не патентная формула, а раздел патента с заголовком «Technical Problem». На Рисунке 5 представлен алгоритм извлечения решаемой проблемы устройства и технических функций.

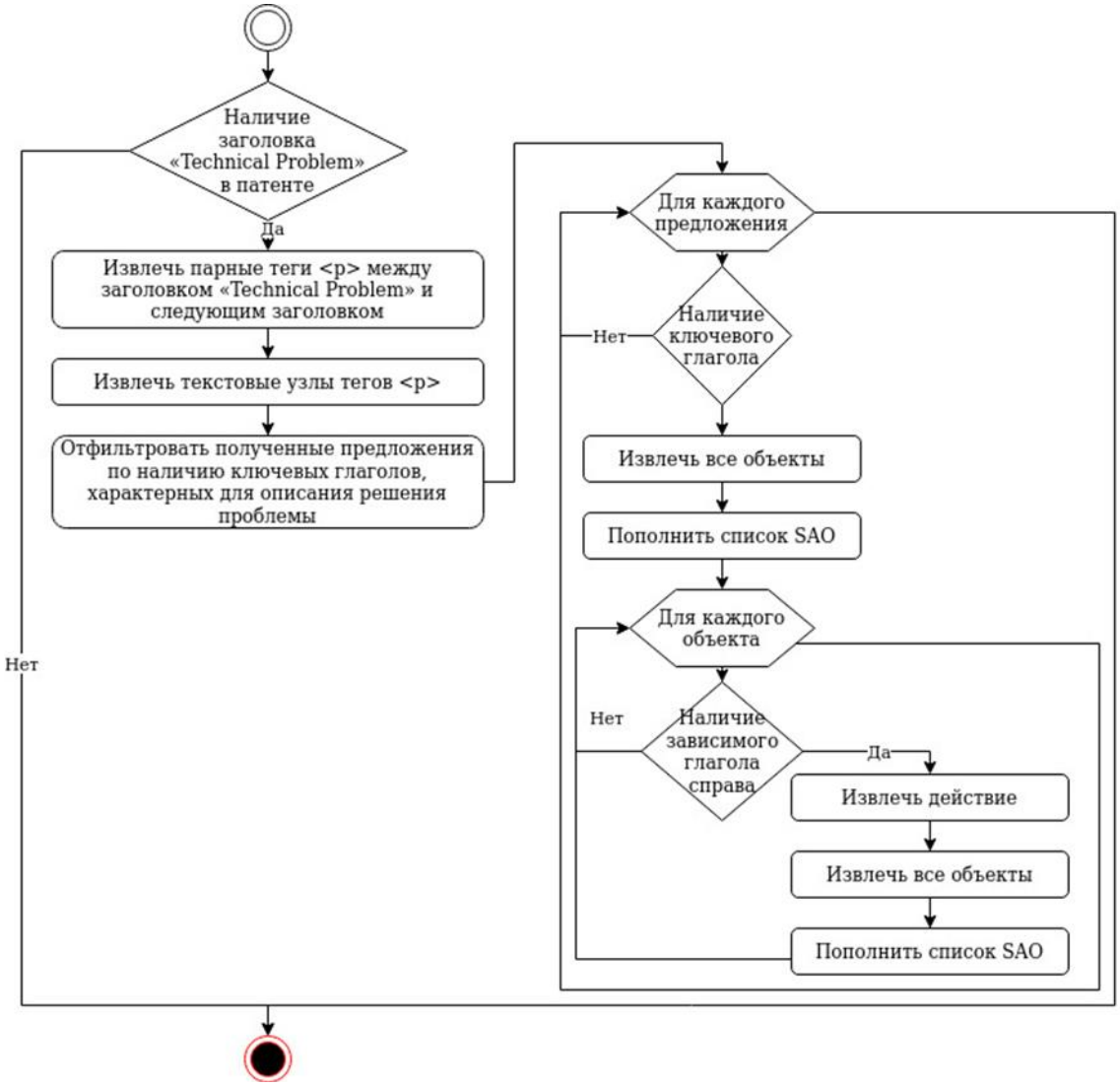

Рисунок 5 - Алгоритм извлечения решаемой устройства проблемы и технических функций
Figure 5 - Algorithm for removing the problem and technical functions to be solved

2. Формирование онтологии

Триплеты являются основным способом выражения информации в онтологиях. Триплет состоит из трех компонент – субъекта, предиката и объекта. Такая модель идеально подходит для хранения извлеченных реализаций технических объектов в виде SAO. Так, триплет будет состоять из трех компонент – субъекта, действия, объекта.

На Рисунке 6 можно увидеть схему классов онтологии предметных областей «Технические функции» и «Реализации технических объектов»

Были выделены следующие свойства объектов:

- hasFunction – свойство для связи технической функции и компонента;
- comprises – свойство для связи между компонентами устройства (глагол «comprise»);
- connectedTo – свойство для связи между компонентами устройства (глаголы «connect», «attach»);
- consists – свойство для связи между компонентами устройства (глаголы «consists», «include»);
- parentFor – указание наличия родительского отношения между элементами (глагол «have»);

- partOf – указание принадлежности компоненту к устройству патентного документа;
- solutionFor – свойство для связи проблемы и решаемой ею устройством;
- connected_to – связь между элементами (глаголы «устанавливать», «подключать», «соединять» и т.д.).

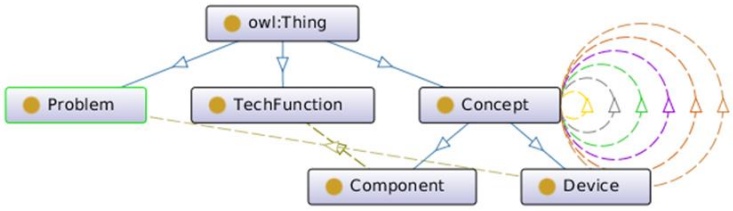

Рисунок 6 - Схема классов онтологии предметных областей «Технические функции» и «Реализации технических объектов»

Figure 6 - The diagram of ontology classes of subject areas "Technical functions" and "Technical objects implementation"

На Рисунке 7 можно увидеть алгоритм пополнения онтологии.

Полученная онтология экспортируется в файл в формате OWL, который потом может быть открыт для дальнейшей работы в Protege.

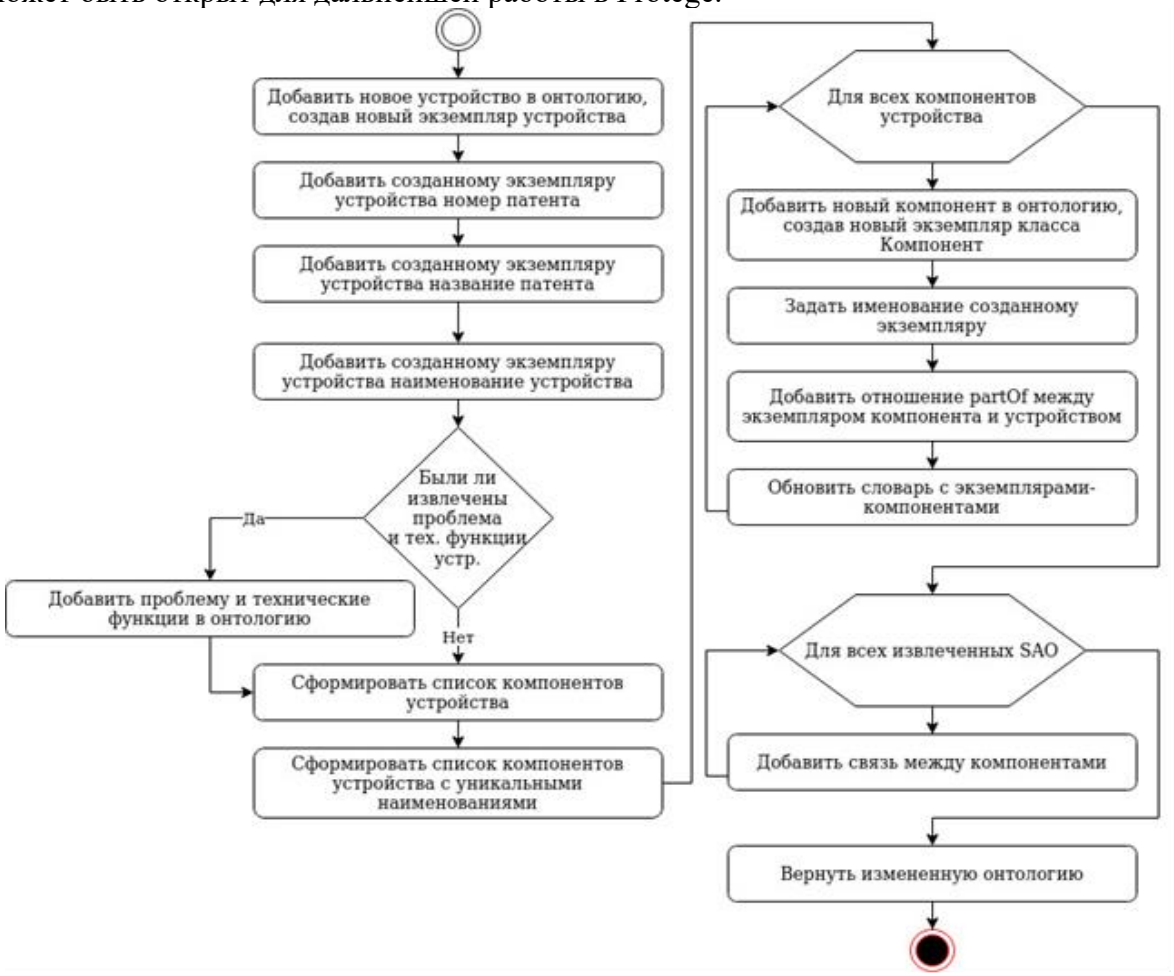

Рисунок 7 - Алгоритм пополнения онтологии предметных областей «Технические функции» и «Реализации технических объектов»

Figure 7 - Algorithm for the replenishment of ontology of subject areas "Technical functions" and "Implementation of technical objects"

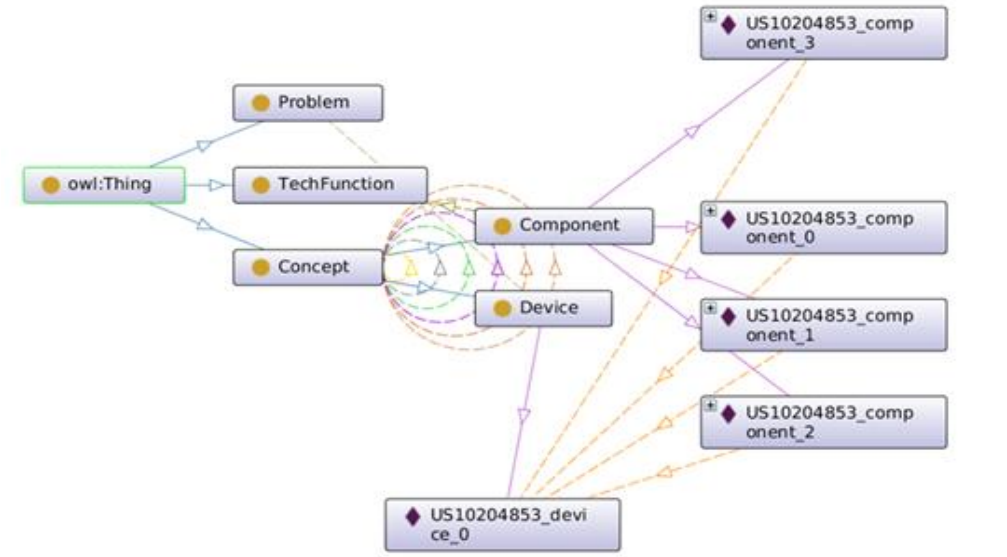

Рисунок 8 - Интерфейс системы
Figure 8 - The program interface

Результаты

Автоматизированная система реализована в виде десктопного приложения для операционных систем семейства Linux. Разработка велась на операционной системе Ubuntu 18.04.4. Система реализована на языке программирования Python 3.6.9. Для создания пользовательского интерфейса использовалась библиотека PyQt5. Для анализа текстов на естественном языке использовалась последняя версия Stanford NLP под названием Stanza. Для хранения извлеченных SAO использовалась СУБД MySQL, для разработки использовалась библиотека для Python PyMySQL. Анализ XML-файлов осуществлялся при помощи библиотеки lxml. Для работы с онтологиями использовалась библиотека Owlready2.

Автоматизированная система позволяет загружать патентные документы, извлекать технические функции и реализации технических объектов, выводить извлеченные реализации технических объектов в форму, строить онтологии для выбранного пользователем патента, а так же для всех загруженных патентов, для которых были извлечены технические функции и реализации технических объектов. На Рисунке 8 представлена построенная онтология для одного патентного документа.В качестве вычислительного эксперимента были вручную разобраны патентные документы, зафиксировано количество извлеченных SAO для каждого патента и время, потраченное на разбор каждого патентного документа. Точность извлечения (P) считалась по формуле (1)

$$P = \frac{\mathrm{E}}{\mathrm{N}}, \quad (1)$$

где E - количество верно извлеченных системой SAO, N - количество SAO в патентном документе.

В Таблице 1 можно увидеть результаты проведенного эксперимента.

Таблица 1 - Результаты проведенного эксперимента
Table 1 - Results of the experiment

Номер эксперимента	Время, затраченное на обработку системой, с	Время, затраченное на обработку экспертом, с	Точность извлечения, %
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%

Среднее время разбора одного патента системой оставило - 1.72316 секунды, среднее время разбора одного патента экспертом - 46.6 секунд. Показатели точности составляют выше 70%.

Обсуждение

В данном работе решалась общая проблема информационного обеспечения синтеза новых технических решений на основе анализа патентов USPTO.

В качестве концептов онтологии предметных областей рассматривались элементы конструкции технического объекта (ТО) и связи между ними, а так же описания решаемых изобретением проблем. Первый пункт формулы изобретения патентного документа выступал в качестве основного источника информации. Единицей извлечения являлись семантические структуры SAO (Subject-Action-Object).

Были определены основные лингвистические особенности патентных документов. Сформирован метод предварительной обработки патентного массива. Разработан алгоритм извлечения SAO из формулы патента. Сформирован метод экспорта извлеченных SAO из англоязычных патентов в онтологию предметных областей. Разработанные методы были апробированы на патентных документах США

Заключение

Теоретическая ценность данной работы заключается в разработанном методе автоматизированного формирования онтологии предметной области «Патентное представление технических систем» для поиска инновационных технических решений, а также построенной на его основе автоматизированной системе.

Благодарности

Работа выполнена при финансовой поддержке РФФИ (грант № 18-07-01086 а), РФФИ и Администрации Волгоградской области (гранты №№ 19-47-340007 р_а, 19-41-340016 р_а).

ЛИТЕРАТУРА

1. Коробкин Д.М., Фоменков С.А., Колесников С.А. Метод синтеза функциональной структуры новых технических решений на основе данных патентных массивов. *Моделирование, оптимизация и информационные технологии*. 2019;7(2):135-148.
2. Коробкин Д.М., Фоменков С.А., Колесников С.Г. Автоматизация процесса формирования информационного обеспечения базы данных физических эффектов. *Вестник компьютерных и информационных технологий*. 2005;3(9):22-25.

3. Kharitonov A., Korobkin D., Fomenkov S., Kolesnikov S. Extraction of morphological features of technical systems from russian patent. В сборнике: CEUR Workshop Proceedings. *IS 2019 - Proceedings of the 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction*. 2019:205-213.
4. Korobkin D.M., Vasiliev S.S., Fomenkov S.A., Lobeyko V.I. Extraction of structural elements of inventions from russian-language patents. В сборнике: *Multi Conference on Computer Science and Information Systems, MCCSIS 2019 - Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019 and Theory and Practice in Modern Computing 2019. 4*. 2019:159-166.
5. Васильев С.С., Коробкин Д.М., Фоменков С.А. Метод извлечения элементов конструкции изобретений из русскоязычных патентов. *Математические методы в технике и технологиях - ММТТ*. 2019;7:105-110.
6. 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, 2011:863-883.. DOI: 10.1007/s11192-011-0420-z.
7. Stanza, 2020. URL: https://stanfordnlp.github.io/stanza/.
8. Kravets A.G., Korobkin D.M., Dykov M.A.E-patent examiner: two-steps approach for patents prior-art retrieval. В сборнике: IISA 2015 - 6th International Conference on Information, Intelligence, Systems and Applications. 2015. DOI: 10.1109/IISA.2015.7388074.

REFERENSES

1. Korobkin D.M., Fomenkov S.A., Kolesnikov S.A. Metod sinteza funkcional'noj struktury novyh tekhnicheskih reshenij na osnove dannyh patentnyh massivov. *Modelirovanie, optimizaciya i informacionnye tekhnologii*. 2019;7(2):135-148.
2. Korobkin D.M., Fomenkov S.A., Kolesnikov S.G. Avtomatizaciya processa formirovaniya informacionnogo obespecheniya bazy dannyh fizicheskih effektov. *Vestnik komp'yuternyh i informacionnyh tekhnologij*. 2005;3(9):22-25.
3. Kharitonov A., Korobkin D., Fomenkov S., Kolesnikov S. Extraction of morphological features of technical systems from russian patent. V sbornike*: CEUR Workshop Proceedings. IS 2019 - Proceedings of the 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction*. 2019:205-213.
4. Korobkin D.M., Vasiliev S.S., Fomenkov S.A., Lobeyko V.I. Extraction of structural elements of inventions from russian-language patents. V sbornike: *Multi Conference on Computer Science and Information Systems, MCCSIS 2019 - Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019 and Theory and Practice in Modern Computing 2019. 4*. 2019:159-166.
5. Vasil'ev S.S., Korobkin D.M., Fomenkov S.A. Metod izvlecheniya elementov konstrukcii izobretenij iz russkoyazychnyh patentov. *Matematicheskie metody v tekhnike i tekhnologiyah - MMTT*. 2019;7:105-110.
6. Choi, S. et al, 2011. SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. Scientometrics, 2011:863-883. DOI: 10.1007/s11192-011-0420-z.
7. Stanza, 2020. URL: https://stanfordnlp.github.io/stanza/.
8. Kravets A.G., Korobkin D.M., Dykov M.A.E-patent examiner: two-steps approach for patents prior-art retrieval. V sbornike: *IISA 2015 - 6th International Conference on Information, Intelligence, Systems and Applications.* 2015. DOI: 10.1109/IISA.2015.7388074.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Григорий Алексеевич Верещак, аспирант кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: *grigoryg37@gmail.com*

Grigoriy.A. Vereshchak, graduate student of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation

Дмитрий Михайлович Коробкин, канд. техн. наук, доцент кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: *dkorobkin80@mail.ru*

Dmitry M. Korobkin, PhD, Associate Professor of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation

Сергей Алексеевич Фоменков, д-р техн.наук, профессор кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: *saf550@yandex.ru*

Sergey A. Fomenkov, Doctor of Tech.Sciences, Professor of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation

Марина Александровна Фоменкова, аспирант кафедры САПР, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: *dvam@vstu.ru*

Marina A. Fomenkova, graduate student of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation

Сергей Григорьевич Колесников
старший научный сотрудник кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: *sk375@bk.ru*

Sergey Grigorievich Kolesnikov
Senior Researcher of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation