УДК 004.89 DOI: 10.26102/2310-6018/2020.31.4.006

Система формирования матрицы выполняемых физическими эффектами технических функций на основе анализа патентного массива

Д.М. Коробкин, Д.В. Шабанов, С.А. Фоменков, А.М. Дворянкин
Волгоградский государственный технический университет, Волгоград, Российская Федерация

Резюме: В статье представлена программная реализация метода формирования матрицы выполняемых физическими эффектами технических функций на основе анализа патентного массива. Для синтеза физического принципа действия новых технических систем могут использоваться физические эффекты из базы знаний, разработанной на кафедре САПРиПК ВогГТУ. Физические эффекты реализуют технические функции, которые в свою очередь составляют конструктивную функциональную структуру технической системы. На основе разработанного авторами метода извлечения описаний физических эффектов и технических функций из патентных документов США (USPTO) и Роспатента был сформирован метод автоматического построения таблицы технических функций, выполняемых физическими эффектами, основанный на выявлении латентных зависимостей в терм-документных матрицах «Физические эффекты-Патенты» и «Технические функции-Патенты». Автоматизированная система реализована на языке Python версии 3.7.2. Для морфологической разметки используется TreeTagger, для синтаксической разметки – UDPipe. Корректность работы алгоритмов была оценена на тестовой выборке, подготовленной вручную и состоящей из 60 патентных документов, описывающих 480 технических функций и 20 физических эффектов. Полученные результаты: метод извлечения ТФ показал на тестовой выборке точность – 0.87, полноту – 0.77 и F -меру – 0.82, поиск описания ФЭ функционирует с точностью - 0,92.

Ключевые слова: технические функции, физические эффекты, патенты, извлечение информации, SAO, CRUD

Для цитирования: Коробкин Д.М., Шабанов Д.В., Фоменков С.А., Дворянкин А.М. Система формирования матрицы выполняемых физическими эффектами технических функций на основе анализа патентного массива *Моделирование, оптимизация и информационные технологии*. 2020;8(4). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=852 DOI: 10.26102/2310-6018/2020.31.4.006

The software for formation the matrix of technical functions performed by physical effects based on patent database analysis

D.M. Korobkin, D.V. Shabanov, S.A. Fomenkov, A.M. Dvoryankin
Volgograd State Technical University, Volgograd , Russian Federation

Abstract: The article presents a software for forming a matrix of technical functions performed by physical effects based on the analysis of a patent database. For the synthesis of the physical operation principle of new technical systems, physical effects from the knowledge base developed at the CAD Department of VSTU can be used. Physical effects implement technical functions, which in turn constitute the constructive functional structure of the technical system. The automated system is implemented in Python version 3.7.2. TreeTagger is used for morphological analysis, UDPipe is used

for syntactic analysis. The correctness of the algorithms was evaluated on a test sample prepared by hand and consisting of 60 patent documents describing 480 technical functions and 20 physical effects. The results obtained: the method for extracting TF showed an accuracy of 0.87 on the test sample, completeness - 0.77 and F-measure - 0.82, the search for the description of the FE functions with an accuracy of 0.92.

***Keywords**:* technical functions, physical effects, patents, fact extraction, SAO, CRUD

For citation: Korobkin D.M., Shabanov D.V., Fomenkov S.A., Dvoryankin A.M. Criteria for hard disk drive multiparametric ranking by failure risk. *Modeling, optimization and information technology*. 2020;8(4). Available from: https://moitvivt.ru/ru/journal/pdf?id=852 DOI: 10.26102/2310-6018/2020.31.4.006 (In Russ).

Введение

В статье [1] показан разработанный метод автоматизированного построения базы данных выполняемых физическими эффектами технических функций. Для синтеза физического принципа действия [2] новых технических систем в ряде научных подходов используются физические эффекты (ФЭ) [3,4]. ФЭ реализуют технические функции, которые в свою очередь составляют конструктивную функциональную структуру технической системы [5,6]. Авторы разработали метод [7] извлечения описаний физических эффектов и технических функций из патентных документов США (USPTO) и Роспатента. Метод автоматического построения таблицы технических функций, выполняемых физическими эффектами, основан на выявлении латентных зависимостей в терм-документных матрицах «Физические эффекты-Патенты» и «Технические функции-Патенты».

Цель работы – программно реализовать систему формирования матрицы выполняемых физическими эффектами технических функций на основе анализа патентного массива

Материалы и методы

Автоматизированная система (АС) должна обеспечивать выполнение следующих функций (Рисунок 1):

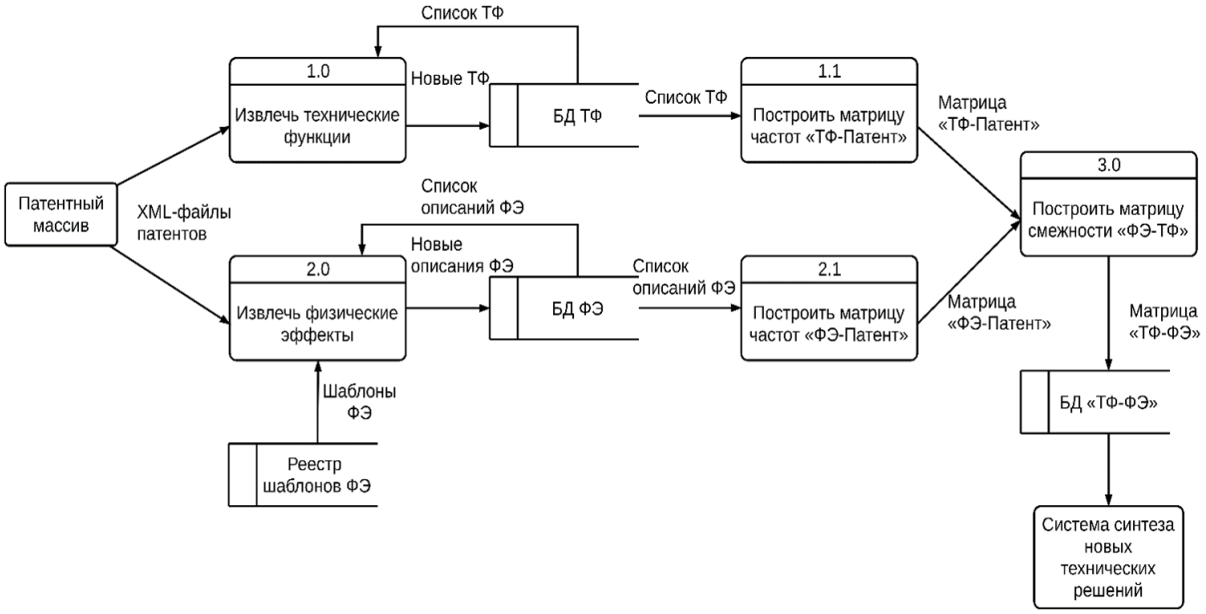

Рисунок 1- Диаграмма потоков данных метода построения матрицы выполняемых физическими эффектами технических функций

Figure 1- Method of matrix construction of technical functions performed by physical effects

- извлечение технических функций в формате «Субъект – Действие – Объект» («Subject – Action – Object», SAO) из текстов патентных документов;
- поиск описания физического эффекта в тексте патентных документов;
- построение терм-документной матрицы «Патент – Техническая функция», элементами которой являются значения частотной характеристики TF-IDF [8] для соответствующей технической функции и патентного документа;
- построение терм-документной матрицы «Патент – Физический эффект», элементами которой являются значения частотной характеристики TF-IDF для соответствующего физического эффекта и патентного документа;
- сокращение пространства технических функций для терм-документной матрицы «Патент – Техническая функция» и пространства физико-технических эффектов для терм-документной матрицы «Патент – Физический эффект»;
- построение из редуцированных терм-документных матриц «Патент – Физический эффект» и «Патент – Техническая функция» матрицы «Физический эффект – Техническая функция» на основе метода косинусов как характеристики близости представлений физического эффекта и технической функции в пространстве патентных документов.

АС реализована на языке python версии 3.7.2. Для морфологической разметки используется программа TreeTagger [9], для синтаксической разметки – UDPipe [10]. Информационная структура входных XML-документов должна соответствовать объявлению типа документа (DTD) «us-patent-application», версия патентного документа должна соответствовать «v4.4 2014-04-03», отвечающему общему описанию стандарта «ST.36».

АС состоит из двух основных и независимых частей (Рисунок 2): хранилище патентного массива и семантическое ядро.

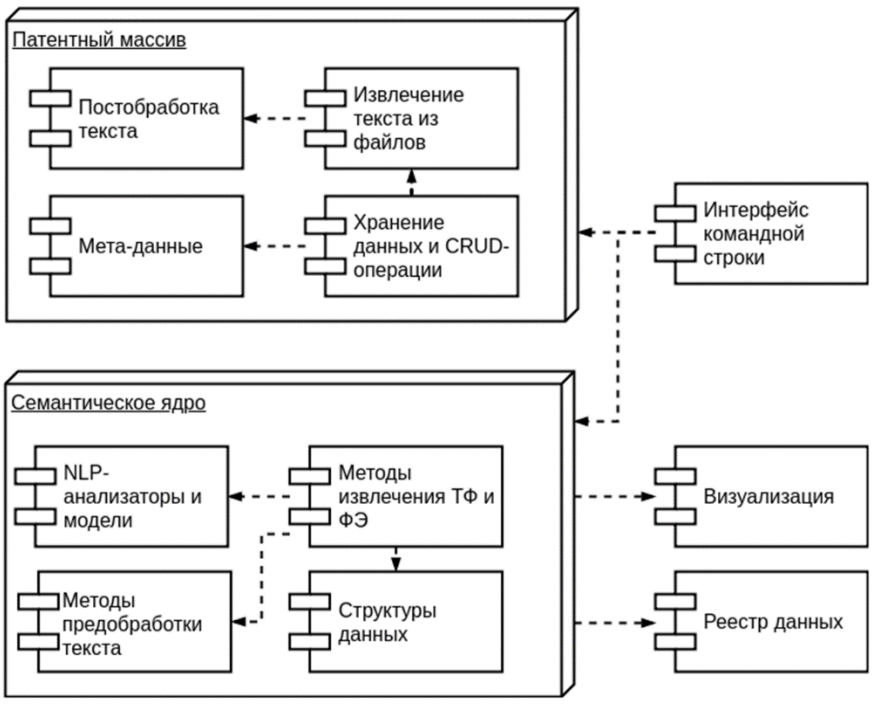

Рисунок 2 - Диаграмма компонентов
Figure 2 - Component Diagram

Хранилище патентного массива реализует функционал хранения и выдачи пользователю текстов патентных документов и их мета-данных, таких как номер, класс по классификации IPC и прочее. Хранилище реализует стандартный интерфейс CRUD-операций по созданию, чтению (получению), обновлению и удалению документов, а также функционал написания запросов к извлекаемым данным, позволяющий фильтровать документы по значению определенных полей. Такая реализация позволяет при необходимости заменить данный модуль на любое NoSQL хранилище в будущем. Реализация хранилища документов хранит данные патентов и все необходимые мета-данные локально, в указанной директории. При необходимости кластерных вычислений, папка с данными может быть помещена в любую сетевую файловую систему, поддерживающую технологию FUSE, и примонтирована в нужную папку. Со стороны программного интерфейса данное хранилище представляет собой интерфейс предоставляющий коллекцию документов в виде генератора, что позволяет не хранить всю запрошенную выборку в памяти, а извлекать данные только при обращении к ним.

Вся работа по извлечению данных из файлов патентных документов делегирована отдельной библиотеке, реализующей функционал чтения и извлечения данных для конкретных форматов патентных документов. Вся коммуникация между хранилищем и библиотекой извлечения данных осуществляется через набор классов данных. Архитектура блока работы с хранилищем данных представлена в виде диаграмм классов (Рисунок 3) и объектов (Рисунок 4).

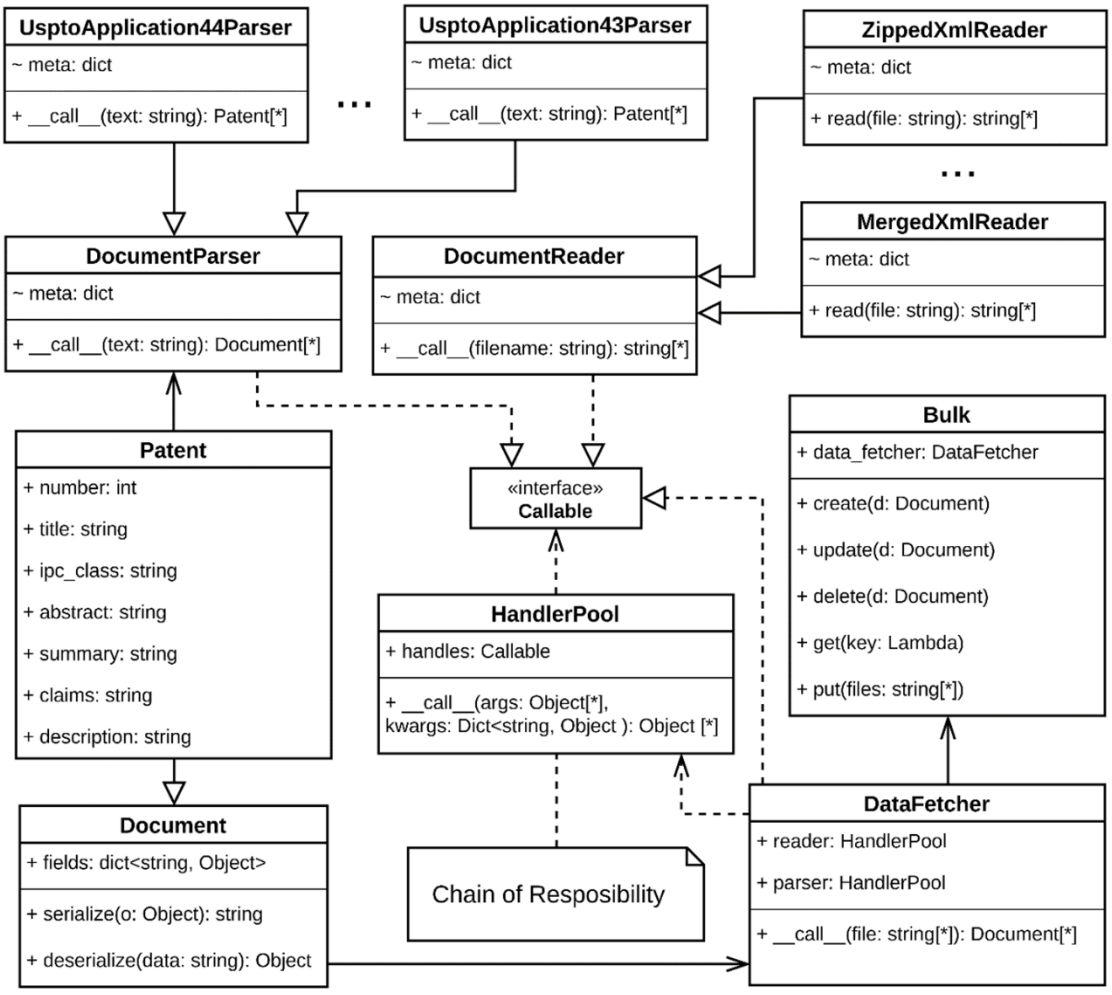

Рисунок 3 - Диаграмма классов хранилища документов
Figure 3 - Document Storage Class Diagram

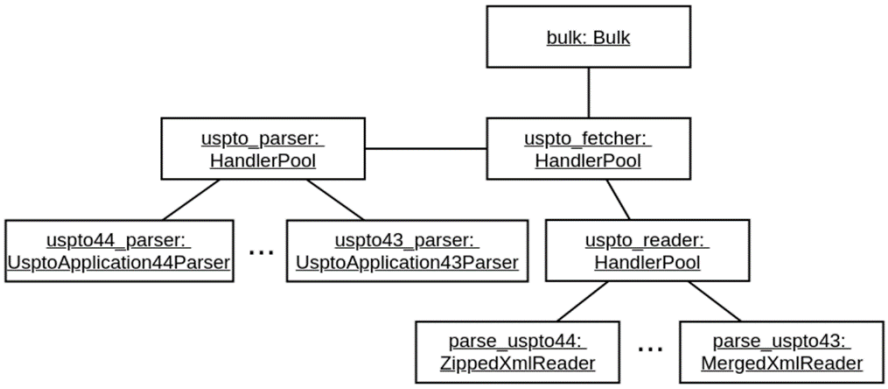

Рисунок 4 - Диаграмма объектов хранилища патентного массива
Figure 4 - Diagram of Patents Storage Objects

Функционал CRUD-операций к патентному массиву реализован в классе «Bulk», работающим с объектами, реализующими интерфейс «Document» – в данном случае «Patent». За извлечение данных и создание объектов документов данных отвечает «DataFetcher», для работы которого нужны два объекта классов, реализующих интерфейсы «DocumentReader» и «DocumentParser», осуществляющих извлечение текста патентов из файлов патентных документов в их исходных форматах (файлы из базы USPTO содержат в себе сразу несколько патентов, архивы и другие данные) и созданию объектов документов (класс «Patent») соответственно. Так как форматов входных файлов и форматов самих документов несколько десятков, то для упрощения на схеме большинство из них опущено и заменено знаком «…», символизирующем множество различных реализаций для каждого формата. По тем же причинам введен класс «HandlerPool», представляющий собой стандартный шаблон проектирования «Цепочка обязанностей». «HandlerPool» хранит список всех зарегистрированных обработчиков и при поступлении запроса на обработку делегирует его одному из них, что позволяет легко добавлять функционал обработки новых форматов входных данных малыми изменениями кода. Связи объектов классов продемонстрированы на рисунке Рисунок 4.

Вторая часть (семантическое ядро) – библиотека, реализующая весь функционал по обработке текста:

- графематический и лексический анализ, реализованный в виде текстовых обработчиков, а также адаптеров к сторонним библиотекам обработки текста, таким как NLTK;
- морфологический и синтаксический анализ, реализованный в виде адаптеров к сторонним библиотекам TreeTagger, MaltParser, UdPipe и прочим;
- синтаксический и семантический анализ — пакеты по работе с ФЭ и ТФ, программно реализующие методы и структуры данных, описанные в данной работе.

Архитектура блока семантического ядра представлена в виде диаграмм классов (Рисунок 5, Рисунок 6) и объектов (Рисунок 7).

При подходе, когда данные являются первопричиной, основной упор делается на: структуры хранения и работы с различными форматами и нотациями (классы SAO, Conll, ConllTree и прочее) и обработчики этих данных, принимающие на вход данные в одном формате и производя морфологический (класс TreeTagger), синтаксический (класс

UdPipe), семантический анализы, например, извлечение технических функций (класс SaoExtractor), продуцирующие данные в другом.

Главной концепцией, которая взята за основу, является конвейер данных (Data Pipeline). Большинство существующих систем пакетной обработки данных (англ. Batch Processing) построены на этом принципе. В данной работе используется несколько сторонних библиотек, построенных на конвейере данных, которые имеют свою реализацию, специфичную для используемого ими формата данных (матрицы, последовательности токенов и другое). Было решено не использовать одну из существующих систем конвейерной обработки в виду их громоздкости и специфики, а реализовать более простую и гибкую реализацию на основе шаблона проектирования «Итератор» и элементов функционального программирования. Основная идея заключается в построении цепочки задач, реализованных на основе шаблона проектирования «Команда», позволяющего реализовать отсроченное выполнения некоторого набора функций. Иллюстрацией применения данного подхода являются текстовые обработчики (Рисунок 6), и этот принцип был применен ко всей системе.

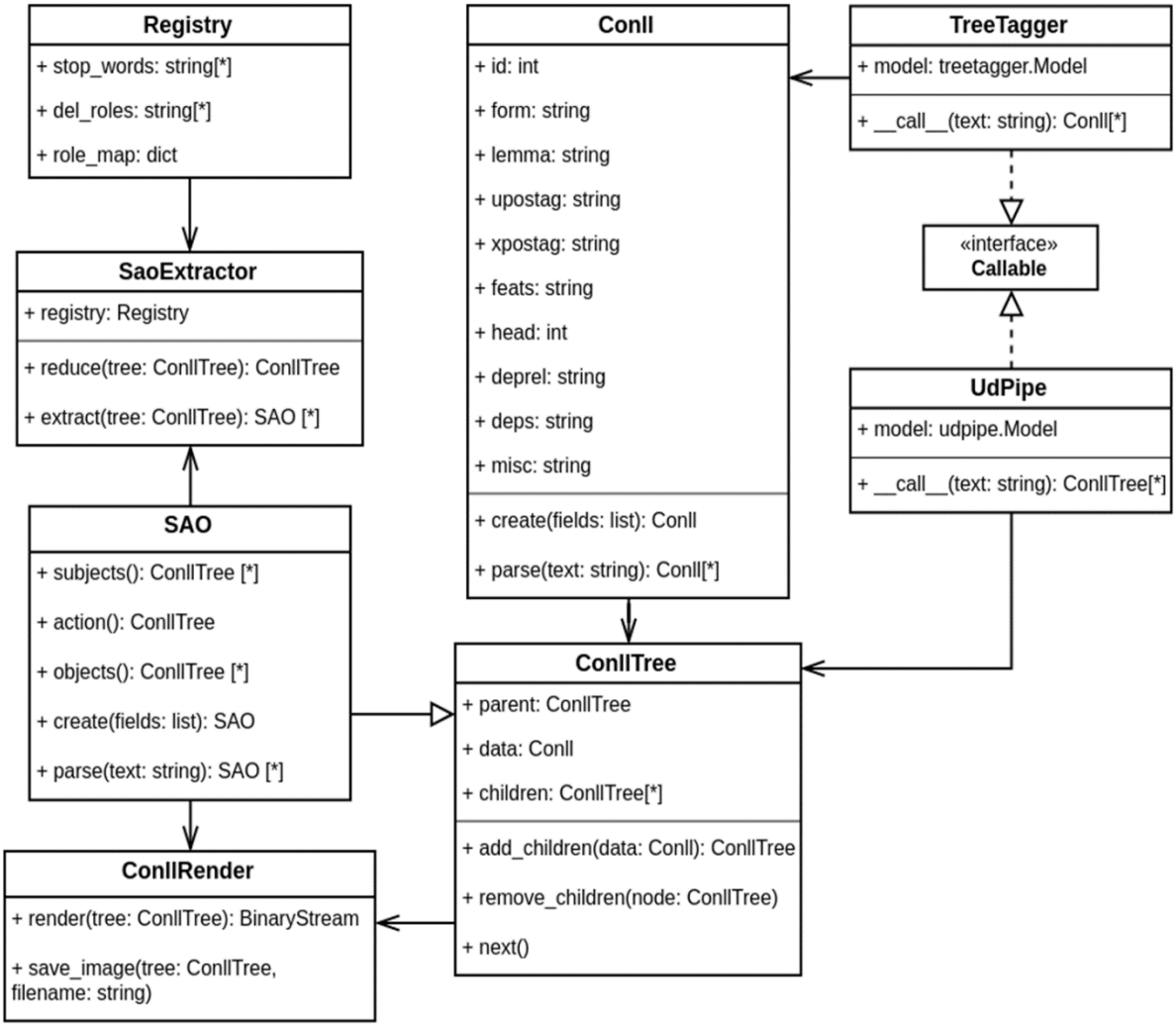

Рисунок 5 – Диаграмма классов семантического ядра
Figure 5 - Semantic Core Class Diagram

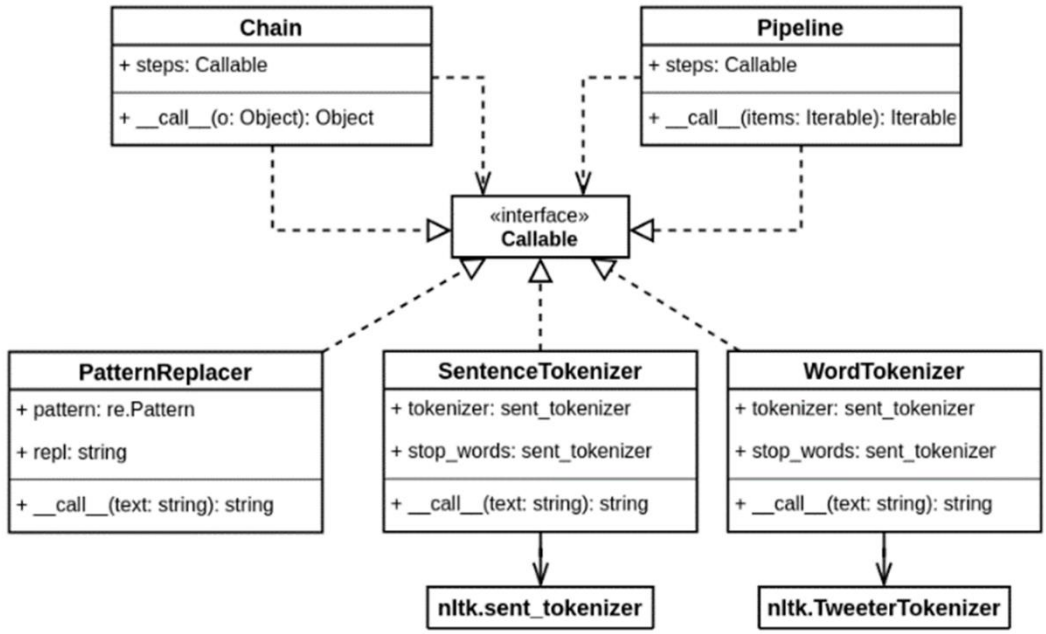

Рисунок 6 – Диаграмма классов конвейерной обработки данных
Figure 6 - Pipeline Data Processing Class Diagram

На диаграмме представлены текстовые обработчики «PatternReplacer», «WordTokenizer» и «SentenceTokenizer», обрабатывающие текст на разном уровне: как последовательность символов, слов и предложений соответственно. «PatternReplacer» предназначен для замены или удаления блоков текста по шаблону, экземпляры данного класса реализую свой специфичный функционал по, например, удалению формул и нумерации параграфов, что проиллюстрировано на Рисунке 7. Классы «WordTokenizer» и «SentenceTokenizer» добавляют функционал к аналогичным классам библиотеки NLTK, «WordTokenizer» объединяет несколько слов в одно для именованных сущностей по переданным правилам, «SentenceTokenizer» сегментирует предложения. Данные классы предназначены для устранения из текста конструкций, не несущих семантической значимости в рамка решаемой задачи, но негативно влияющих на корректность работы морфологических и синтаксических анализаторов. Описанные выше классы реализуют стандартный шаблон проектирования «Команда», что позволяет создавать объекты отложенного исполнения и составлять цепочки задач (класс «Chain») или конвейеры (класс «Pipeline») обработки данных. Класс «Chain» представляет собой цепочку выполнения, хранит список обработчиков, реализующих интерфейс «Callable», передает входные данные — единичный объект — первому, его результат следующему и так далее по цепочке. Класс «Pipeline» реализует схожий функционал, но работает с коллекцией объектов.

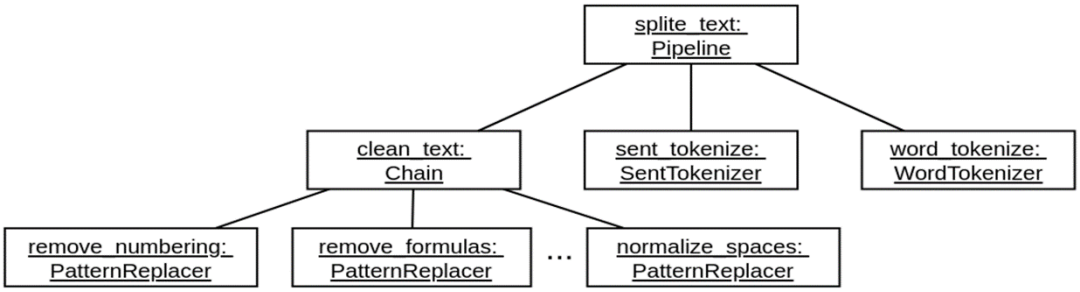

Рисунок 7 – Диаграмма объектов конвейерной обработки
Figure 7 - Pipeline Object Diagram

Результаты

Реализованы интерфейсы командной строки для работы с хранилищем данных (Рисунок 8) и семантическим ядром (Рисунок 9).

```
root@notebook:~# ./bulk.py -h

Bulk command line tool.

Usage:
  bulk put <file>... [--override]
  bulk list [--verbose|--count=<kn>]
  bulk stat
  bulk -h | --help
  bulk --version

Options:
  -h --help     Show this screen.
  --version     Show version.
  --override    Override existing documents.
  --verbose     Verbose mode.
```

Рисунок 8 – Интерфейс командной строки хранилища данных
Figure 8 - Data Warehouse Command Line Interface

```
root@notebook:~# ./semcore.py -h

SemCore command line tool.

Usage:
  semcore morph <file>...
  semcore synt <file>... [--output=(conll | picture)]
  semcore sao <file>... [--output=(conll | picture)]
  semcore effect <file>... [--output=(plain | table)]

  semcore -h | --help
  semcore --version

Options:
  -h --help     Show this screen.
  --version     Show version.
```

Рисунок 9 – Интерфейс командной строки семантического ядра
Figure 9 - Semantic Core Command Line Interface

В соответствие с моделью физического эффекта, каждая его компонента описывается регулярным выражение, пример описания приведен в Таблице 1, для наглядности шаблон представлен в следующем виде: опциональные части шаблона представлены в прямоугольных скобках, альтернативы перечислены, разделенные символом «|». Примеры найденных описаний физического эффекта приведены в Таблице 2.

Таблица 1 – Описание физического эффекта «Закон Ома»
Table 1 - Description of the physical effect of "Om Law"

Компонент		Описание	Шаблон
Вход	Воздействие	электрическое поле	[weak] electric[al] field
	Характеристика воздействия	слабое	
	Физическая величина	напряженность электрического поля (В/м)	(electric[al] field) density | pressure) | voltage
Выход	Воздействие	электрический ток	
	Характеристика воздействия	Постоянный, переменный,	([alternating | direct | ionic | mixed]

Компонент		Описание	Шаблон
		электронный, ионный, смешанный	[electric[al]] current) | AC | DC
	Физическая величина	плотность тока (А/м**2)	[electric[al]] current density
Объект		проводник, полупроводник	[semi]conductor | resistance | resistor

Таблица 2 – Примеры поиска описаний технических функций в тексте патента
Table 2 - Examples of searching technical function descriptions in the patent text

Номер	Найденный ФЭ	Входные данные
1	ФЭ № 303 «Термо-фотоэлектрический эффект»	Патент US6380534B1 The amplitude of the Brillouin peaks and the frequency shift of the Brillioun peaks compared with the Rayleigh peak is a measure of the voltage and temperature of the optical fiber at the point from which the light was backscattered.
2	ФЭ № 37 «Закон Ома»	Патент US2965301A Conductors, as indicated, are connected to the resistors for application thereto of factor- representing voltages and/or currents and for deriving therefrom an output, all as more fully explained hereinafter.

Примеры найденных описаний технических функций в формате SAO приведены в Таблице 3.

Таблица 3 – Примеры поиска технических функций в тексте патента
Table 3 - Examples of searching for technical functions in the patent text

Номер	Найденный SAO	Входные данные
1	S: method and apparatus A: measure O: temperature and strain within a structure	Патент US6380534B1 …A method and apparatus for measuring the temperature and strain within a structure consists in having optical fibres incorporated in the structure, passing pulses of light down the fibre and detecting the backscattered light…
	S: optical fibres A: pass O: pulses of light down the fibre	
	S: optical fibres A: detect O: backscattered light	
2	S: object of the invention A: provide O: simple and reliable multiplier-divider computer unit of increased capacity.	Патент US2965301A … object of the invention is the provision of a simple and reliable multiplier-divider computer unit of increased capacity. Another object of the invention is the provision of a simple and reliable D.C. multiplier-divider computer unit…
	S: object of the invention A: provide O: simple and reliable D.C. multiplier-divider computer unit	

Обсуждение

Корректность работы алгоритмов была оценена на тестовой выборке, подготовленной вручную, технические функции извлекались из поля «Краткое

изложение изобретения» (англ. Summary of Invention) документа, а физические эффекты искались в поле полного описания изобретения (англ. Description). Тестовая выборка была составлена из 60 патентных документов и насчитывает 480 технических функций и описание 20 физических эффектов, причем один документ содержит описание только одного физического эффекта.

Метод извлечения ТФ: точность – 0.87 , полнота – 0.77 и F -мера – 0.82.

Поиск описания ФЭ: точность - 0,92.

Для тестирования метода построения базы данных выполняемых физическими эффектами технических функций было произведено объединение тестовой и проектировочной выборок, размерами 60 и 10 тысяч патентных документов соответственно.

Будем считать, что для конкретного документа тестовой выборки найдено соответствие ФЭ и реализуемых им ТФ, если хотя бы 80% размеченных экспертами в документе и верно найденных на этапе тестирования метода извлечения ТФ технических функций были отмечены в матрице ТФ-ФЭ для данного ФЭ. Пороговое значение введено в связи с возможным исключение технических функций на этапе сокращения их пространства.

По результатам тестирования точность извлечения выполняемых физическими эффектами технических функций составила 0.78.

Заключение

Теоретическая ценность данной работы заключается в разработанной методике анализа графических представлений математических формул для расширения описаний научно-технических эффектов и созданной на ее 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. Korobkin D., Shabanov D., Fomenkov S., Golovanchikov A. Construction of a matrix "Physical effects – Technical functions" on the base of patent corpus analysis. *Communications in Computer and Information Science*. 2019;1084:52-68.
2. Korobkin D.M., Fomenkov S.A., Kolesnikov S.G., Kamaev V.A. Synthesis of the physical principle of operation of engineering systems in the software environment CPN Tools. *Research Journal of Applied Sciences*. 2014;9(11):749-752.
3. 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.
4. Korobkin D.M., Fomenkov S.A., Kolesnikov S.G., Voronin Y.F. System of physical effects extraction from natural language text in the Internet. *World Applied Sciences Journal*. 2013;24(24):55-61.
5. Коробкин Д.М., Фоменков С.А., Колесников С.Г. Метод верификации синтезированной функциональной структуры посредством построения физического принципа действия технической системы. *Моделирование, оптимизация и информационные технологии*. 2019;7(2):97-109.

6. 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.
7. Шабанов Д.В., Коробкин Д.М., Фоменков С.А., Колесников С.Г. Формирование матрицы "Физические эффекты - Технические функции" на основе данных анализа патентных массивов. *Математические методы в технике и технологиях - ММТТ*. 2019;7:94-99.
8. Uther, William & Mladenić, Dunja & Ciaramita, Massimiliano & Berendt, Bettina & Kołcz, Aleksander & Grobelnik, etc. TF–IDF. 2010. DOI: 10.1007/978-0-387-30164-8_832.
9. Schmid, & Helmut,. TreeTagger - a language independent part-of-speech tagger. 2009.
10. Straka, Milan & Strakova, Jana & Hajič, Jan. UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging. 2019.

REFERENCES

1. Korobkin D., Shabanov D., Fomenkov S., Golovanchikov A. Construction of a matrix "Physical effects – Technical functions" on the base of patent corpus analysis. *Communications in Computer and Information Science*. 2019;1084:52-68.
2. Korobkin D.M., Fomenkov S.A., Kolesnikov S.G., Kamaev V.A. Synthesis of the physical principle of operation of engineering systems in the software environment CPN Tools. *Research Journal of Applied Sciences*. 2014; 9(11):749-752.
3. 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.
4. Korobkin D.M., Fomenkov S.A., Kolesnikov S.G., Voronin Y.F. System of physical effects extraction from natural language text in the Internet. *World Applied Sciences Journal*. 2013;24(24):55-61.
5. Korobkin D.M., Fomenkov S.A., Kolesnikov S.G. Metod verifikacii sintezirovannoj funkcional'noj struktury posredstvom postroeniya fizicheskogo principa dejstviya tekhnicheskoj sistemy. *Modelirovanie, optimizaciya i informacionnye tekhnologii*. 2019;7(2-25):97-109
6. 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(2-25):135-148.
7. SHabanov D.V., Korobkin D.M., Fomenkov S.A., Kolesnikov S.G. Formirovanie matricy "Fizicheskie effekty - Tekhnicheskie funkcii" na osnove dannyh analiza patentnyh massivov. *Matematicheskie metody v tekhnike i tekhnologiyah - MMTT*. 2019;7:94-99.
8. Uther, William & Mladenić, Dunja & Ciaramita, Massimiliano & Berendt, Bettina & Kołcz, Aleksander & Grobelnik, etc. TF–IDF. 2010. DOI: 10.1007/978-0-387-30164-8_832.
9. Schmid, & Helmut,. TreeTagger - a language independent part-of-speech tagger. 2009.
10. Straka, Milan & Strakova, Jana & Hajič, Jan. UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging. 2019.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

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

Шабанов Дмитрий Владимирович,
аспирант кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: *shabanov.dmitry.v@gmail.com*

Dmitry V. Shabanov, graduate student 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*

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

Александр Михайлович Дворянкин, д-р техн. наук, профессор кафедры ПОАС, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: *dvam@vstu.ru*

Alexander M. Dvoryankin , Doctor of Tech.Sciences, Professor of the POAS Department Volgograd State Technical University, Volgograd, Russian Federation