

Научная статья

Статья в открытом доступе

УДК 004.89

doi: 10.30987/2658-6436-2023-2-14-22

РАЗРАБОТКА ПРОГРАММНОГО МОДУЛЯ СЕМАНТИЧЕСКОГО АНАЛИЗА ПАТЕНТНОГО МАССИВА

Дмитрий Михайлович Коробкин^{1✉}, Арсен Ваганович Манукян^{2✉}, Сергей Алексеевич Фоменков^{3✉}, Светлана Александровна Козина^{4✉}

^{1, 2, 3, 4} Волгоградский государственный технический университет, г. Волгоград, Россия

¹ dkorobkin80@mail.ru, <http://orcid.org/0000-0002-4684-1011>

² rasen13@mail.ru, <http://orcid.org/0000-0000-0000-0000>

³ saf@vstu.ru, <http://orcid.org/0000-0001-9907-4488>

⁴ ksvetlan54@gmail.com, <http://orcid.org/0000-0003-4049-620X>

Аннотация. В ходе проектирования программного модуля был разработан алгоритм парсинга текстов патентов и патентных заявок, алгоритм поиска патентов-аналогов на основе полнотекстового поиска с использованием технологии Amazon Twinword, алгоритм поиска патентов-аналогов на основе ключевых фраз, выявленных с использованием технологии Amazon Comprehend, кластеризации патентного массива. Было разработано программное обеспечение для кластеризации патентного массива и ускорения работы эксперта патентного ведомства за счет поиска ключевых фраз и патентов-аналогов.

Ключевые слова: патент, семантический анализ USPTO, AWS, Amazon Comprehend, Twinword, DynamoDB

Финансирование: Исследование выполнено за счет гранта Российского научного фонда № 23-21-00464, <https://rscf.ru/project/23-21-00464/>.

Для цитирования: Коробкин Д.М., Манукян А.В., Фоменков С.А., Козина С.А. Разработка программного модуля семантического анализа патентного массива // Автоматизация и моделирование в проектировании и управлении. 2023. №2 (20). С. 14-22. doi: 10.30987/2658-6436-2023-2-14-22

Original article

Open Access Article

DEVELOPING A SOFTWARE MODULE FOR THE SEMANTIC ANALYSIS OF THE PATENT ARRAY

Dmitry M. Korobkin^{1✉}, Arsen V. Manukyan^{2✉}, Sergey A. Fomenkov³, Svetlana A. Kozina⁴

^{1,2,3,4} Volgograd State Technical University, Volgograd, Russia

¹ dkorobkin80@mail.ru, <http://orcid.org/0000-0002-4684-1011>

² rasen13@mail.ru, <http://orcid.org/0000-0000-0000-0000>

³ saf@vstu.ru, <http://orcid.org/0000-0001-9907-4488>

⁴ ksvetlan54@gmail.com, <http://orcid.org/0000-0003-4049-620X>

Abstract. While designing the software module, the authors developed an algorithm for parsing the texts of patents and patent applications, an algorithm for searching patents-analogues based on full-text search using Amazon Twinword technology, an algorithm for searching patents-analogues based on key phrases identified applying Amazon Comprehend technology, clustering the patent array. The software was developed to cluster the patent array and speed up the patent office examiner's work by searching for key phrases and patent analogues.

Keywords: patent, USPTO semantic analysis, AWS, Amazon Comprehend, Twinword, DynamoDB

Funding: The research was carried out at the expense of the grant of the Russian Science Foundation No. 23-21-00464, <https://rscf.ru/project/23-21-00464/>

For citation: Korobkin D.M., Manukyan A.V., Fomenkov S.A., Kozina S.A. Developing a software module for the semantic analysis of the patent array. Automation and modeling in design and management, 2023, no. 2 (20). pp. 14-22. doi: 10.30987/2658-6436-2023-2-14-22.

Введение

Одна из самых больших проблем XXI века, связанная с патентным правом, это загруженность патентного ведомства. Только USPTO (ведомство по патентам и товарным знакам США) [1] в неделю рассматривает порядка 10 тысяч патентов. И каждому патенту нужно

уделить особое внимание, понять: не затронуты ли патентные права другого патента, не является ли патент псевдонаучным, «очевидным», в конечном итоге произвести полную классификацию патента со всеми описаниями и ссылками.

Семантический анализ патентного массива позволяет решить ряд современных проблем:

1. Кластеризация патентного массива (моделирование тем) позволяет выявлять группы связанных (не на основе патентной классификации, а на базе извлеченных из текстов ключевых терминов/фраз) патентов. Данная информация может быть полезной для выявления патентных трендов, ключевых современных технологий и прогноза востребованности технологий в будущем временном периоде.

2. Автоматизация работы эксперта патентного ведомства. На основе полнотекстового запроса (текста патентной заявки) может осуществляться поиск патентов-аналогов. Кроме того, может быть автоматизирован процесс выявления ключевых фраз как в тексте патентной заявки, так и в тексте патента.

Были проведены предпроектные исследования: изучена патентная классификация, структура патента, патентные поисковые системы (Google Patents [2], USPTO [1], Espacenet [3], ФИПС Роспатента [4]).

В работе было проведено сравнение существующих патентных поисковых систем по следующим критериям: поиск по ключевым словам; поиск по метаданным; выделение ключевых слов.

Результаты проведенного сравнительного анализа представлены в табл. 1.

Таблица 1

Результаты сравнения существующих решений

Table 1

Results of comparison of existing solutions

Система	Поиск по ключевым словам	Поиск по метаданным (патентной классификации)	Выделение ключевых слов
Espacenet	+	+	+
USPTO	+	+	-
Google Patents	+	-	+
ФИПС	+	+	-

Несмотря на то, что поиск по ключевым словам присутствует в функционале всех систем, одной из насущных проблем патентного поиска является отсутствие автоматизации труда эксперта патентного ведомства в части поиска ключевых слов/фраз.

Кластеризация патентного массива и определение патентов-аналогов внутри определенного кластера также позволяют автоматизировать труд эксперта патентного ведомства (рис. 1).

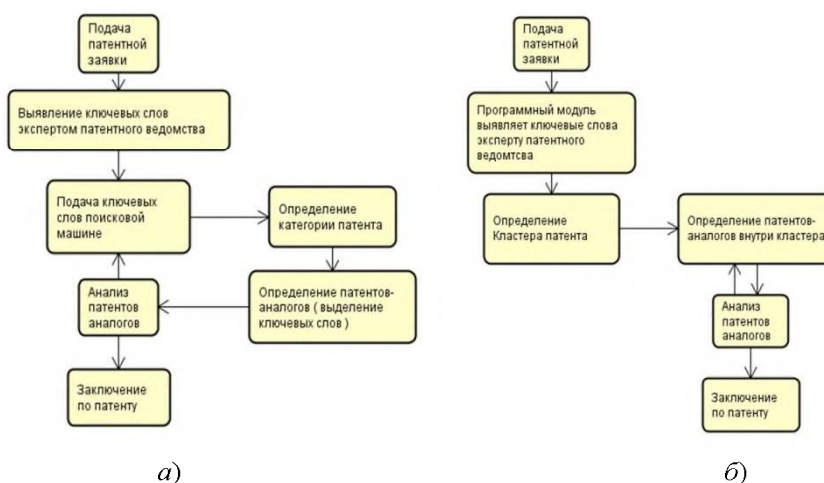


Рис. 1. Процесс работы эксперта патентного ведомства:

a – AS-IS; б – TO-BE

Fig. 1. The process of work of the patent office expert:

a – AS-IS; б – TO-BE

Была поставлена задача исследования – разработка методологии и технологии автоматизации работы эксперта патентного ведомства за счет выявления ключевых фраз в тексте патентов и поиска патентов-аналогов.

Установлены следующие требования к программному обеспечению:

- использование технологии AWS [5];
- применение технологий семантического анализа текста Amazon Comprehend [6], Twinword [7];
- исходные файлы патентов USPTO необходимо хранить в AWS S3 [8];
- извлечение элементов описания патентов должно осуществляться из патентных баз USPTO;
- для хранения элементов описания патентов необходимо использовать систему управления базами данных (СУБД) DynamoDB [9].

Парсинг патентных документов

На первом этапе происходит парсинг патента – извлечение метаданных (даты публикации, названия, классификации, имен авторов, кода и т.д.) для наполнения баз данных (БД) DynamoDB. Извлеченные поля патента (с которыми в дальнейшем и будет производиться обработка): поля рефератов, описаний и формул изобретений, помещаются в хранилище S3 для дальнейших преобразований технологиями Amazon.

Патентный архив представляет собой xml-файл, содержащий в себе патенты за определенный период времени (неделю). В процессе парсинга патентного архива извлекаются и помещаются в корневые директории патентные тексты в формате xml для дальнейшей обработки.

Патентные тексты обрабатываются по следующему алгоритму:

1. Патентный текст проверяется на валидность формата. Если патент валиден, алгоритм продолжается, иначе патент обработке не подлежит.
2. В процессе парсинга проверяется наличие патентных метаданных (даты публикации, названия, классификации, имен авторов, кода и т.д.), если хотя бы одна из метаданных отсутствует, патент признается невалидным (проверка происходит не после парсинга всех метаданных, а после проверки каждого поля), иначе алгоритм продолжается.
3. Патент, прошедший все проверки (валидный), сохраняется по следующей логике: метаданные патента – в DynamoDB, поля патента (рефераты, описания и формулы изобретений) – в корневые директории и далее в хранилище Amazon S3.

Выявление ключевых фраз

Выявление ключевых фраз в тексте патентной заявки и/или патента происходит при помощи технологии Amazon Comprehend – Detect Key Phrases (ACDKP) [10]. В выбранном патенте находятся ключевые слова с их значимостью. Уже на данном этапе это намного облегчает работу эксперта патентного ведомства, так как обычно поиск подобных ключевых фраз их использования в поисковых патентных системах (например, Google patents) экспертом производится вручную.

Кластеризация патентного массива

Кластеризация – задача группировки множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию.

Для кластеризации патентного массива используется технология Amazon Comprehend – Topic Modeling (ACTM) [11] (тематическое моделирование). Тематическое моделирование – способ построения модели коллекции текстовых документов, которая определяет к каким темам относится каждый из документов. Тематическая модель коллекции текстовых документов также определяет какие слова (термины) образуют каждую тему.

Переход из пространства терминов в пространство найденных тематик помогает раз-

решать синонимию и полисемию терминов, а также эффективнее решать такие задачи как: тематический поиск, классификация и т.п.

В нашем исследовании база (более чем 50 тысяч патентов) неявно кластеризуется на 10 топиков (кластеров) с помощью технологии Amazon Comprehend – Topic Modeling. Каждый патент, представленный в виде обработанного txt-файла, соотносится с выбранным в ходе моделирования тем топиком.

Поиск патентов-аналогов на основе полнотекстового поиска

Поиск патентов-аналогов на основе полнотекстового запроса осуществляется при помощи технологии Amazon Twinword TextSimilarity [7] (ATTS). Патентный массив после кластеризации обрабатывается с использованием различных подходов. Например, может быть выбрана группа патентов, принадлежащая к определенному топикю, и при этом патенты должны иметь принадлежность к топикю, превышающую установленное пороговое значение, например, 0,8 (не может превышать 1). Далее из этой группы выбирается один патент и сравнивается со всеми остальными для проверки «сильной» схожести текстов. Также может быть сделана обратная проверка, например, могут быть выбраны патенты, принадлежащие к различным топикам и имеющие принадлежность к этим топикам более чем 0,8 для проверки «слабой» схожести текстов.

Результаты

Для реализации разрабатываемой программы был выбран язык программирования Python версии 3.7. Для реализации парсинга патентных документов и заявок была выбрана библиотека xml.dom. Для получения доступа к технологиям AWS использовалась библиотека boto3, с помощью которой было реализовано подключение к Amazon сервисам: DynamoDB, Amazon S3, ACDKP, АСТМ, ATTS.

Диаграмма вариантов использования приведена на рис. 2.

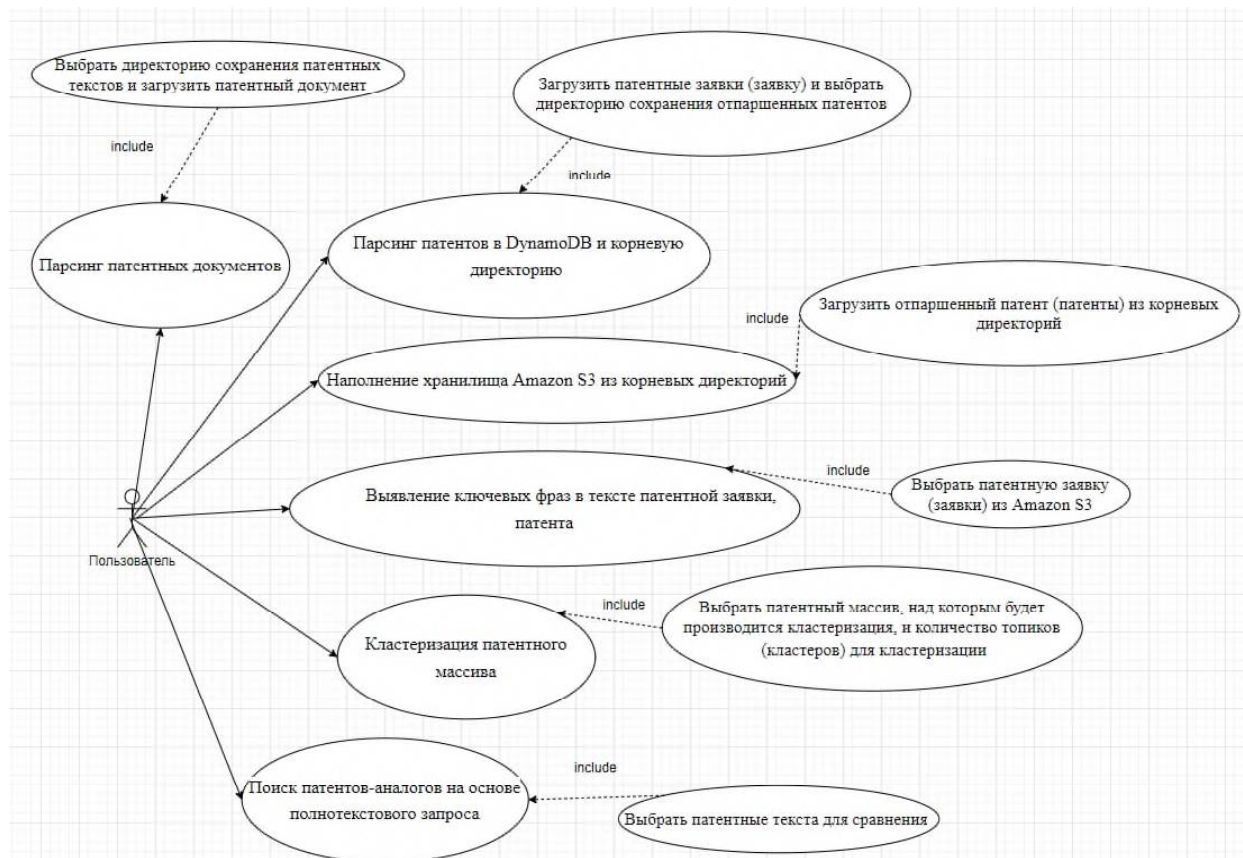


Рис. 2. Диаграмма вариантов использования программы
Fig. 2. Diagram of the use cases of the program

Архитектура программы представлена на рис. 3, где 1 – запись распарсенных патентов в директорию; 2 – чтение патентных документов; 3 – запись метаданных распарсенных патентов в DynamoDB; 4 – подключения и проверки DynamoDB; 5 – заполнение хранилища Amazon S3 распарсенными патентами из корневых директорий; 6 – подключение к технологиям Amazon; 7 – подключение к технологии Amazon Comprehend; 8 – подключение к технологии Amazon Twinword; 9 – кластеризация патентного массива с помощью технологии Topic Modeling; 10 – выявление ключевых фраз с помощью технологии DetectKeyPhrases; 11 – запись результата работы Topic Modeling в Amazon S3; 12 – запись результата работы DetectKeyPhrases в Amazon S3; 13 – поиск патентов-аналогов с помощью технологии Text Similarity; 14 – запись результата работы Text Similarity в корневые директории.

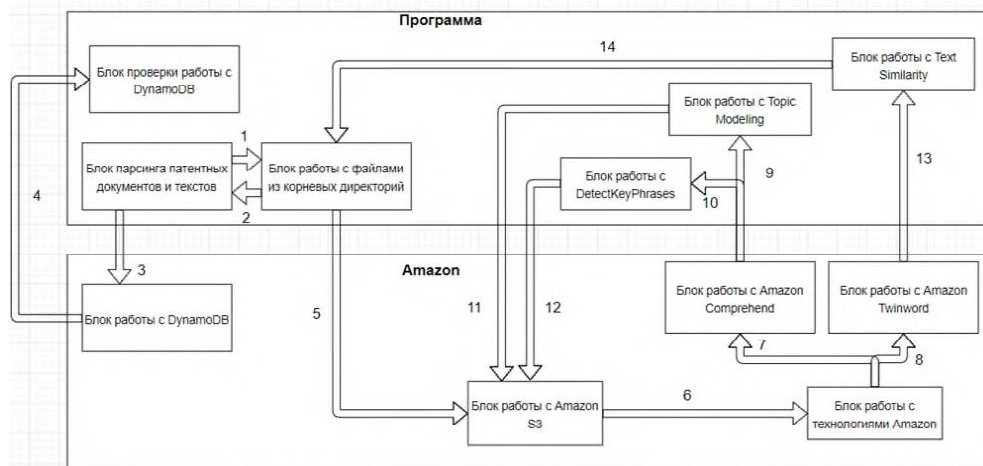


Рис. 3. Архитектура программного модуля
Fig. 3. Architecture of the software module

Для хранения основных характеристик патентов: дата публикации, название, классификация, имена авторов, код, ссылка на патент в хранилище Amazon S3, используется СУБД класса NoSQL в формате ключ-значение – DynamoDB.

На рис. 4 представлен вид таблицы patent в DynamoDB. Работа с DynamoDB осуществляется из публичного облака Amazon как часть пакета AWS.

id	author_name(s)	date_of_publicati	ipc_class	patent_L
251	{ "Christophe Alepee", "Thierry Gulton" }	2020.12.22	{ "A61B 17/00", "A61B 17/0023", "A61B 17/00477", "A61B 17/00482", "A61...	US1069
1838	{ "Jochen Hiemeyer", "Martin Kerstner", "Michael Freitag" }	2020.12.22	{ "F25B 1/00", "F25B 2339/046", "F25B 2600/02", "F25B 2600/11", "F25B 2...	US1067
3389	{ "Stefan Müller" }	2020.12.22	{ "C23C 14/08", "G11C 11/22", "G11C 11/223", "G11C 11/2273", "H01L 21/...	US1067
3416	{ "Li Wang", "Ling Shi", "Song Zhang" }	2020.12.22	{ "H01L 2251/5338", "H01L 27/1244", "H01L 27/3227", "H01L 27/3244", "H...	US1067
4119	{ "Hao Chen", "Maowei Yang" }	2020.12.22	{ "G06F 21/6254", "G06Q 10/10", "G06Q 50/01", "H04L 12/4625", "H04L 4...	US1067
4666	{ "Matthew Lucius Coletreglio", "Michael T. Stanhope" }	2019.11.26	{ "A41D 1/06", "A41D 13/00", "A41D 13/01", "A41D 13/012", "A41D 13/015...	US1049
4847	{ "Yoshihiro Kojima" }	2019.11.26	{ "A61B 5/0077", "A61B 5/024", "A61B 5/11", "A61B 5/18", "A61B 5/6893", ...	US1048
5628	{ "Fuyuki Sugiura", "Motohiro Nagaya" }	2019.11.26	{ "B23Q 17/0961", "G05B 13/026", "G05B 19/19", "G05B 2219/45244", "H0...	US1048
5749	{ "Kim L. Walton", "Yushan Hu" }	2019.11.26	{ "C08K 3/16", "C08L 23/10" }	US1048
6207	{ "Benoit Prouvost", "Delphine Destal", "Marie-Edith Quereau", "Philippe Mi..." }	2019.11.26	{ "B05D 7/14", "B32B 15/08", "B32B 2435/02", "B65D 41/023", "C08F 283/...	US1048
6633	{ "Gregory Patton", "Janine Graham", "Nicole Nichols" }	2019.11.26	{ "A61K 9/19", "C12N 9/1252", "C12Q 1/6848", "C12Q 1/686", "C12Y 207/0...	US1048
7304	{ "Chunming-Parker Zhang", "Lixian Liu", "Sylvain Yvon" }	2019.11.26	{ "B60Q 3/82", "F21K 9/00", "F21S 41/143", "F21S 41/192", "F21S 43/14", ...	US1048

Рис. 4. Структура БД DynamoDB
Fig. 4. DynamoDB database structure

Для хранения извлеченных полей патента (полей рефератов, описаний и формул изобретений) в txt-формате используется хранилище Amazon S3. Выбор данного хранилища в главной степени обусловлен тем, что обработка патентного массива на основе технологий AWS требует доступа к «родному» файловому хранилищу Amazon S3.

Высший слой файлового хранилища Amazon S3 представлен в виде Buckets. В свою очередь Buckets имеют внутри себя различные объекты. В данном случае реализация имеет следующий вид: каждая папка patent.storage внутри Bucket является аналогом патентного документа и включает в себя все обработанные патенты в формате txt. Хранилище обработанных патентов показано на рис. 5. Также хранилище Amazon S3 используется для хранения результатов работы АСДКР и АСТМ (Buckets «key phrases» и «topic.modeling», соответственно).

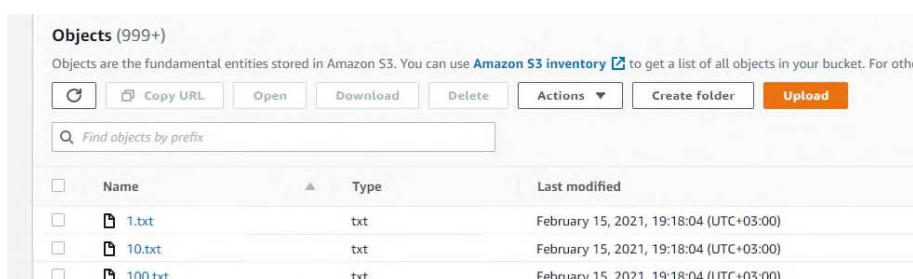


Рис. 5. Хранилище обработанных патентов в Amazon S3
Fig. 5. Storage of processed patents in Amazon S3

Ключевая фраза – это строка, содержащая словосочетание, описывающее определенный объект. Обычно она состоит из существительного и модификаторов. Например, слово «day» – это существительное, а «a beautiful day» – это словосочетание, включающее артикль «a» и прилагательное «beautiful».

Каждая ключевая фраза в Amazon Comprehend имеет оценку, которая указывает на уровень уверенности в том, что данная строка является словосочетанием, содержащим существительное (рис. 6). Данную оценку можно использовать для того, чтобы определить, достаточно ли высок уровень обнаружения того или иного объекта. Операции по обнаружению ключевых фраз могут выполняться с использованием любого из языков, поддерживаемых Amazon Comprehend. При этом необходимо учитывать, что все документы должны быть на одном языке.

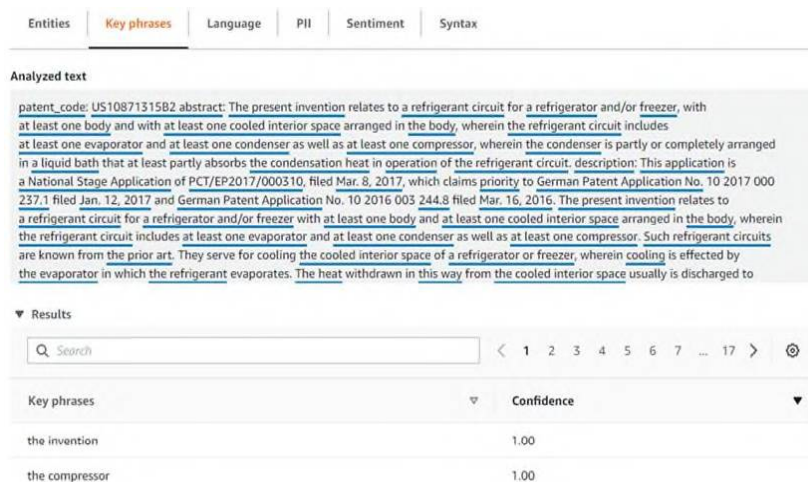


Рис. 6. Результаты Text Analysis (Key Phrases)
Fig. 6. The result of Text Analysis (Key Phrases)

Amazon Comprehend использует скрытую модель обучения на основе распределения Дирихле для определения тем в наборе документов. Он проверяет каждый документ, чтобы определить контекст и значение того или иного слова. Набор слов, которые часто принадлежат к одному и тому же контексту во всем наборе документов, составляет тему. Слово ассоциируется с темой в документе в зависимости от того, насколько распространена эта тема в документе и насколько близка тема к слову. Одно и то же слово может быть связано с раз-

ными темами в разных документах в зависимости от распределения тем в конкретном документе. Например, слово «глюкоза» в статье, в которой говорится преимущественно о спорте, может быть отнесено к теме «спорт», в то время как то же самое слово в статье о «медицине» будет отнесено к теме «медицина». Каждому слову, связанному с темой, присваивается вес, который указывает, насколько это слово помогает определить тему. Вес – это показатель того, сколько раз слово встречается в теме по сравнению с другими словами в теме во всем наборе документов. Результат моделирования темы Amazon Concept показан на рис. 7. По результату можно оценить сходство двух слов, предложений, абзацев или документов, а также получить оценку того, насколько похожи или отличаются два текста.

Например, Amazon говорит о реальном примере использования, данный API был применен при создании первого инструмента семантического исследования ключевых слов, который может быть отсортирован по релевантности. Исследование ключевых слов включает в себя просмотр длинных списков ключевых слов, чтобы найти наиболее релевантные из них. Результат сходства текста Amazon Twinword показан на рис. 8.

topic,term,weight	docname,topic,proportion
000,method,0.016401418	patent_id_dynamodb_5/31342.txt,001,0.531576
000,comprise,0.019146329	patent_id_dynamodb_5/31342.txt,005,0.468424
000,claim,0.021325674	patent_id_dynamodb_2/9956.txt,009,0.867848
000,material,0.007775651	patent_id_dynamodb_2/9956.txt,000,0.084929
000,composition,0.0047576763	patent_id_dynamodb_2/9956.txt,002,0.047223
000,group,0.005427863	patent_id_dynamodb_1/690.txt,000,0.457953
000,invention,0.007046096	patent_id_dynamodb_1/690.txt,002,0.2895
000,compound,0.004202903	patent_id_dynamodb_1/690.txt,004,0.23198
000,acid,0.0033280468	patent_id_dynamodb_1/690.txt,005,0.020567
000,cell,0.00490938	patent_id_dynamodb_2/6776.txt,004,0.408577
001,user,0.0239455	patent_id_dynamodb_2/6776.txt,002,0.352574
001,information,0.023247061	patent_id_dynamodb_2/6776.txt,000,0.206286
001,network,0.013106099	patent_id_dynamodb_2/6776.txt,003,0.032563
001,system,0.0153446505	patent_id_dynamodb_3/16650.txt,009,0.612836
001,base,0.012511535	patent_id_dynamodb_3/16650.txt,000,0.263136
001,method,0.014428071	patent_id_dynamodb_3/16650.txt,006,0.124028
001,communication,0.010289393	patent_id_dynamodb_5/26888.txt,003,0.443265
001,computer,0.009094545	
001,receive,0.010849599	
001,content,0.0075957146	
002,system,0.03701485	
002,vehicle,0.014393139	
002,sensor,0.008779968	

Рис. 7. Результаты Amazon Comprehend Topic Modeling:

a – термины кластеров; *б* – патенты, подвергшиеся кластеризации

Fig. 7. The result of Amazon Comprehend Topic Modeling:

a – cluster terms; *b* – patents that have undergone clustering

```
{'similarity': 0.7703936811034408,
{'similarity': 0.8912862283337145,
{'similarity': 0.5967785286591871,
{'similarity': 0.6410642220816088,
{'similarity': 0.3910770880741828,
{'similarity': 0.315626944456838,
{'similarity': 0.3482453255650077,
{'similarity': 0.46596123512583165,
```

Рис. 8. Результат работы Amazon Twinword Text Similarity

Fig. 8. The result of work Amazon Winword TextSimilarity

Заключение

Актуальность работы обусловлена тем, что семантический анализ патентного массива позволяет решить ряд современных проблем:

– автоматизация работы эксперта патентного ведомства. На основе полнотекстового запроса (текста патентной заявки) может осуществляться поиск патентов-аналогов. Кроме того, может быть автоматизирован процесс выявления ключевых фраз как в тексте патентной заявки, так и в тексте патента;

– кластеризация патентного массива (моделирование тем) позволяет выявлять группы связанных (не на основе патентной классификации, а на базе извлеченных из текстов ключевых терминов/фраз) патентов.

В результате данной работы был разработан программный модуль, обеспечивающий возможность проведения кластеризации патентного массива и позволяющий идентифицировать группы связанных патентов с использованием технологий АСТМ. Также на основе полнотекстового запроса (текст патентной заявки) был проведен поиск патентов-аналогов с использованием технологий АТТС. Процесс определения ключевых фраз как в тексте патентной заявки, так и в тексте патента был автоматизирован с использованием технологий АСДКР.

Теоретическая значимость работы заключается в разработанных алгоритмах парсинга текстов патентов и патентных заявок USPTO; кластеризации патентного массива; извлечения ключевых фраз из патентных текстов; полнотекстового поиска патентов-аналогов.

Практическая значимость работы заключается в разработанном программном модуле семантического анализа патентного массива для задач патентного поиска и кластеризации. В данной работе использовались технологии AWS: семантический анализ текста Amazon Comprehend, Twinword, хранилище AWS S3, СУБД DynamoDB.

Список источников:

1. Patents. *USPTO*. Available from: <https://www.uspto.gov/patents> (Accessed 07.11.2022).
2. Google Patents. Available from: <https://patents.google.com/> (Accessed 07.11.2022).
3. Espacenet – patent search. Available from: <https://worldwide.espacenet.com/> (Accessed 07.11.2022).
4. Федеральный институт промышленной собственности [Электронный ресурс]. URL: <https://www.fips.ru/> (дата обращения 07.11.2022).
5. Joe Baron, Hisham Baz, Tim Bixler, Biff Gaut, Kevin E. Kelly, Sean Senior, John Stamper. AWS Certified Solutions Architect Official Study Guide: Associate Exam. 2016. Available from: <https://www.pdfdrive.com/aws-certified-solutions-architect-official-study-guide-associate-exam-e38558089.html> (Accessed 7.11.2022).
6. Возможности Amazon Comprehend [Электронный ресурс] // aws. URL: <https://aws.amazon.com/ru/comprehend/features/> (дата обращения 07.11.2022).
7. Text Similarity API. *aws marketplace*. Available from: https://aws.amazon.com/marketplace/pp/B071G93T67?ref_=srh_res_product_title (Accessed 07.11.2022).
8. Amazon S3 [Электронный ресурс] // aws. URL: <https://aws.amazon.com/ru/s3/> (дата обращения 07.11.2022).
9. Amazon DynamoDB [Электронный ресурс] // aws. URL: <https://aws.amazon.com/ru/dynamodb/> (дата обращения 07.11.2022).
10. What is Amazon Comprehend? *aws*. Available from: <https://docs.aws.amazon.com/comprehend/latest/>

References:

1. Patents. *USPTO* [Internet] [cited 2022 Nov 07]. Available from: <https://www.uspto.gov/patents>
2. Google Patents [Internet] [cited 2022 Nov 07]. Available from: <https://patents.google.com/>
3. Espacenet – Patent Search [Internet] [cited 2022 Nov 07]. Available from: <https://worldwide.espacenet.com/>
4. Federal Institute of Industrial Property [Internet] [cited 2022 Nov 07]. Available from: <https://www.fips.ru/>
5. Baron J., Baz H., Bixler T., Gaut B., Kelly K.E., Senior S., Stamper J. AWS Certified Solutions Architect Official Study Guide: Associate Exam [Internet]. 2016 [cited 2022 Nov 07]. Available from: <https://www.pdfdrive.com/aws-certified-solutions-architect-official-study-guide-associate-exam-e38558089.html>
6. Opportunities of Amazon Comprehend [Internet]. AWS [cited 2022 Nov 07]. Available from: <https://aws.amazon.com/ru/comprehend/features/>
7. Text Similarity API [Internet]. AWS Marketplace [cited 2022 Nov 07]. Available from: https://aws.amazon.com/marketplace/pp/B071G93T67?ref_=srh_res_product_title
8. Amazon S3 [Internet]. AWS [cited 2022 Nov 07]. Available from: <https://aws.amazon.com/ru/s3/>
9. Amazon DynamoDB [Internet]. AWS [cited 2022 Nov 07]. Available from: <https://aws.amazon.com/ru/dynamodb/>
10. What is Amazon Comprehend? [Internet]. AWS [cited 2022 Nov 07]. Available from: <https://docs.aws.amazon.com/>

dg / get – started – api – key - phrases. html (Accessed 07.11.2022).

11. Topic modeling. *aws*. Available from: <https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html> (Accessed 07.11.2022).

Информация об авторах:

Коробкин Дмитрий Михайлович – кандидат технических наук, доцент кафедры «Системы автоматизированного проектирования и поискового конструирования» Волгоградского государственного технического университета, ORCID: 0000-0002-4684-1011

Манукян Арсен Ваганович – магистрант Волгоградского государственного технического университета, ORCID: 0000-0000-0000-0000

Фоменков Сергей Алексеевич – доктор технических наук, профессор кафедры «Системы автоматизированного проектирования и поискового конструирования» Волгоградского государственного технического университета, ORCID: 0000-0001-9907-4488

Козина Светлана Александровна – магистрант Волгоградского государственного технического университета, ORCID: 0000-0000-0000-0000.

[amazon. com/comprehend/latest/dg/get-started-api-key-phrases.html](https://docs.aws.amazon.com/comprehend/latest/dg/get-started-api-key-phrases.html)

11. Topic Modelling [Internet]. AWS [cited 2022 Nov 07]. Available from: <https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html>

Information about the authors:

Korobkin Dmitry Mikhailovich – Candidate of Technical Sciences, Associate Professor of the Department «Computer-Aided Design and Exploratory Design» of Volgograd State Technical University, ORCID: 0000-0002-4684-1011

Manukyan Arsen Vaganovich – undergraduate of Volgograd State Technical University, ORCID: 0000-0000-0000-0000

Fomenkov Sergey Alekseevich – Doctor of Technical Sciences, Professor of the Department «Computer-Aided Design and Exploratory Design» of Volgograd State Technical University, ORCID: 0000-0001-9907-4488

Kozina Svetlana Alexandrovna – undergraduate of Volgograd State Technical University, ORCID: 0000-0000-0000-0000.

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.

Contribution of the authors: the authors contributed equally to this article.

Авторы заявляют об отсутствии конфликта интересов.

The authors declare no conflicts of interests.

Статья поступила в редакцию 21.11.2022; одобрена после рецензирования 16.12.2022; принята к публикации 16.03.2023.

The article was submitted 21.11.2022; approved after reviewing 16.12.2022; accepted for publication 16.03.2023.

Рецензент – Малаханов А. А., кандидат технических наук, доцент, Брянский государственный технический университет.

Reviewer – Malakhanov A.A., Candidate of Technical Sciences, Associate Professor, Bryansk State Technical University.