

Сравнение эффективности применения различных алгоритмов интеллектуального анализа текстовых данных для построения предметной OWL-онтологии

А.А. Дырnochкин¹, В.С. Мошкин¹✉, Н.Г. Ярушкина¹

¹ Ульяновский государственный технический университет, г. Ульяновск, 432027, Россия

Ссылка для цитирования

Дырnochкин А.А., Мошкин В.С., Ярушкина Н.Г. Сравнение эффективности применения различных алгоритмов интеллектуального анализа текстовых данных для построения предметной OWL-онтологии // Программные продукты и системы. 2025. Т. 38. № 2. С. 234–242. doi: 10.15827/0236-235X.150.234-242

Информация о статье

Группа специальностей ВАК: 2.3.1, 2.3.7

Поступила в редакцию: 14.02.2024

После доработки: 23.09.2024

Принята к публикации: 03.10.2024

Аннотация. В статье описывается программный сервис анализа текстовой информации с целью формирования OWL-онтологии при решении задачи анализа состояния сложных технических систем, например, предметной области нефтедобывающих систем, данные с которых поступают в реальном времени. Основным направлением работы сервиса является обработка текстовой информации для формирования OWL-онтологии – формальной описательной модели, позволяющей семантически структурировать и формализовать информацию. В статье представлены основные принципы функционирования данного сервиса. В процессе работы программный сервис выполняет предобработку текстовых данных, включающую этапы токенизации, удаления стоп-слов, лемматизации и выделения ключевых терминов. Эти операции необходимы для повышения качества и целостности данных перед последующим анализом. Главным этапом работы сервиса является извлечение ключевых слов и терминов из текстов с использованием алгоритмов машинного обучения и статистических методов. Сервис группирует тексты по сходству и формирует кластеры. Этот этап позволяет выявить скрытую структуру в текстовых данных и выделить общие тематики или концепции. Результаты экспериментов, описанные в статье, подтверждают возможность эффективного использования программного сервиса для формирования онтологических структур, демонстрируют способность сервиса классифицировать и группировать текстовые данные, что является важным шагом в процессе создания семантических моделей сложных технических систем. Также в статье проводится сравнительное исследование трех различных концепций извлечения ключевых слов и терминов: статистических методов извлечения ключевых слов, кластеризации и тематического моделирования. Оценивается эффективность каждого метода на основе точности извлечения ключевых терминов, их связанности и понятности. Результаты исследования позволяют сделать вывод о преимуществах и недостатках каждого подхода и определить наиболее продуктивный метод для конкретных задач построения онтологий в информационных системах.

Ключевые слова: интеллектуальный анализ текста, онтология, предобработка, кластеризация, векторизация, тематическое моделирование, извлечение ключевых слов, тематический кластер, семантическая модель, структурирование информации

Благодарности. Работа выполнена при поддержке РФФИ, проект № 23-71-01101, <https://rscf.ru/project/23-71-01101/>

Введение. Проблема извлечения онтологических структур из текстов с целью дальнейшего применения онтологий, полученных при анализе состояния сложных технических систем, в настоящее время является актуальной. Неструктурированные текстовые ресурсы сложны в обработке и нуждаются в структуризации для получения OWL-онтологии, реализующей дескрипционную логику, с возможностью реализации логического вывода с помощью ризоверов (Fact++, Pellet и пр.) [1].

Для решения подобной задачи необходимы эффективные алгоритмы извлечения терминов (объектов OWL-онтологии) и определения тематик (классов OWL-онтологии), а также категоризация текстов (классификация или кластеризация) [2].

Кроме того, важно учитывать контекст и семантику текста, чтобы точно определить связь между терминами и тематиками. Эффективная реализация подобных задач требует применения методов машинного обучения и статистических методов.

Для извлечения терминов можно использовать методы, основанные на анализе частотности и сочетаемости слов, а также на использовании грамматических и лингвистических правил. Для определения тематик необходимо обратиться к семантическим моделям, таким как Word2Vec или GloVe, которые позволяют представить слова в виде векторов и находить их семантические аналогии.

Категоризация текстов может быть выполнена методами машинного обучения, такими как классификация или кластеризация [3, 4].

Для извлечения ключевых слов из текста существует несколько подходов. Один из них – использование программных алгоритмов, анализирующих текст и выделяющих из него наиболее важные слова и фразы.

Эти алгоритмы могут быть основаны на различных подходах, таких как выделение самых часто встречающихся слов, анализ синонимов и антонимов, а также определение связей между словами в тексте.

Например, в работе [5] исследуется задача автоматического построения и обновления онтологии предметной области при добавлении новых знаний в вопросно-ответной системе. Метод, представленный в данной работе, основан на автоматическом построении онтологии из неструктурированного текста научных статей. Для этого применяются алгоритмы извлечения цепочек символов, их идентификации как объектов онтологии и последующей классификации для отнесения к различным семантическим категориям.

В рамках данного проекта была реализована программная система интеллектуального анализа текстов, позволяющая выбирать наиболее эффективные алгоритмы для каждого из видов анализа.

Научная новизна заключается в использовании различных методов обработки естественного языка для автоматического формирования онтологий, что позволяет системе автоматически адаптироваться к различным типам текстов и обновлять онтологию в реальном времени, значительно улучшая ее способность к самообучению и наращиванию объема знаний.

Описание используемых алгоритмов

Алгоритмы извлечения ключевых слов. Для извлечения ключевых слов использованы алгоритмы Rake, Yake, TextRank, биграммы, TF-IDF.

Основной идеей алгоритма Rake является то, что ключевые слова зачастую находятся в окружении стоп-слов и пунктуации.

Стоп-слова и пунктуация расцениваются как разделители фраз: текст разбивается по этим элементам на фразы-кандидаты. Далее фразы-кандидаты ранжируются по формуле и выбираются k кандидатов с наибольшим значением метрики:

$$\text{Rake} = \frac{\text{deg}(w)}{\text{freq}(w)},$$

где $\text{freq}(w)$ – частота слова в тексте (поощряет часто встречающиеся слова); $\text{deg}(w)$ – сумма

совместных появлений других слов с этим словом (совместным считается появление в одной фразе [6]).

Алгоритм Yake (*Yet Another Keyword Extractor*) использует комбинацию статистических подходов и обученных моделей машинного обучения для извлечения ключевых слов. Алгоритм анализирует свойства слов, такие как частота встречаемости, длина, частеречная информация и другие, для определения их важности. Затем применяется классификатор, обученный на размеченных данных, для прогнозирования ключевых слов.

Метод TextRank наиболее отличается от двух предыдущих. Он использует идею, что любой текст можно представить в виде графа, где слова являются вершинами, а связи между ними – ребрами. После перевода текста в графовое представление используется классическая метрика важности вершин графа PageRank.

Для построения графа берется контекст вокруг каждого слова, то есть все слова, которые находятся на расстоянии n слов от главного. Например, для контекста размера 2 берутся два слова слева и два слова справа от текущего. Все слова в контексте текущего связываются с ним ребрами графа.

Метод PageRank использует матрицу связей между словами в тексте, чтобы определить, какие слова наиболее важны для каждого текста. На основе этой матрицы алгоритм вычисляет рейтинг для каждого слова путем учета его семантической связи с другими словами и их частоты в тексте.

В отличие от других методов, использующих для определения важности слов только их частоту, метрика извлечения ключевых слов PageRank учитывает еще и семантическую связь этих слов с другими словами в тексте. Это позволяет выделить наиболее существенные слова для конкретного текста [7].

Алгоритм тематического моделирования. Тематическое моделирование – метод машинного обучения, который используется для анализа больших объемов текстовой информации. Он позволяет автоматически выявлять скрытые структуры в текстах, выделяя темы, наиболее подходящие для описания содержания текстов.

Основная идея тематического моделирования заключается в том, что каждый текст можно рассматривать как смесь нескольких тем. Каждая тема представляет собой распределение вероятностей на словах, которые часто встречаются в текстах, связанных с этой темой.

Таким образом, тематическое моделирование позволяет выделить совокупность слов, образующих конкретную тему, и определить, какие темы обнаруживаются в текстах [8].

Тематическое моделирование также позволяет выделить ключевые слова и фразы, наиболее характерные для каждой темы. Это может быть полезно при поиске информации и классификации текстов.

Одним из основных методов тематического моделирования является модель LDA (*Latent Dirichlet Allocation*). Она использует байесовский подход для оценки распределения слов в темах и распределения тем в текстах. Для обучения модели LDA используются методы максимального правдоподобия и вариационного вывода.

Алгоритм LDA работает следующим образом:

- задается количество тем k , которые необходимо выделить в коллекции документов;
- для каждого слова в документе выбирается тема из распределения тем в документе;
- для каждой темы выбирается распределение вероятностей слов, присутствующих в этой теме;
- алгоритм повторяется до достижения или заданного числа итераций, или стабильности распределений [9].

Одним из основных преимуществ тематического моделирования является возможность автоматического выделения скрытых структур в текстовых данных без необходимости ручной разметки. Это значительно упрощает и ускоряет процесс анализа текстов, позволяя исследователям сосредоточиться на интерпретации результатов [10].

Описание разработанного программного средства

Сервис интеллектуального анализа текстовых данных реализован на языке Python. Сервис выполняет предобработку исходных текстовых документов, после чего происходит определение доминирующих тематик в тексте.

Определение доминирующих тематик основано на трех различных подходах: кластеризация текста, тематическое моделирование, извлечение ключевых слов.

Разработанное приложение состоит из двух микросервисов. Первый, написанный на языке Java, выполняет функции по загрузке статей и сохраняет их в БД. Второй, на языке Python, отвечает за предобработку статей, извлечение ключевых слов и анализ тенденций, связанных с этими ключевыми словами. Он обрабатывает данные, полученные от первого сервиса, и предоставляет информацию пользователю о том, какие темы и тренды можно выделить на основе анализа ключевых слов. Диаграмма, представленная на рисунке 1, наглядно демонстрирует связи между компонентами разработанного приложения и позволяет лучше понять его структуру и функциональность.

Диаграмма последовательности отражает последовательность действий системы в рамках единой временной оси и показывает жизненный цикл объектов и взаимодействие модулей в рамках единой информационной системы. На рисунке 2 представлена диаграмма последовательности анализа тенденций ключевых слов, где прослеживается порядок обработки и анализа данных. Сначала осуществля-

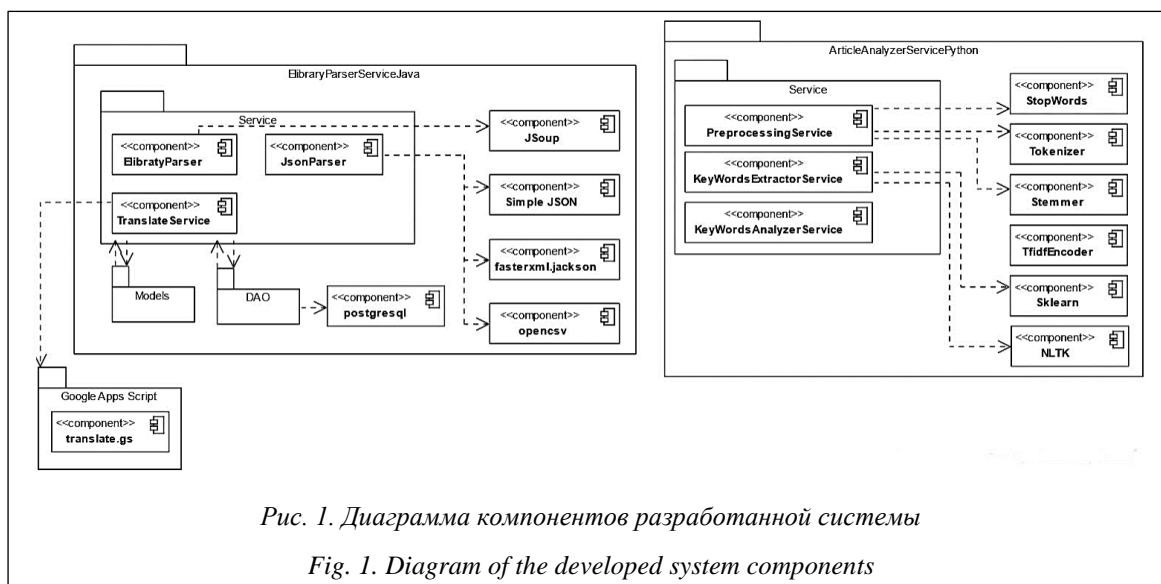


Рис. 1. Диаграмма компонентов разработанной системы

Fig. 1. Diagram of the developed system components

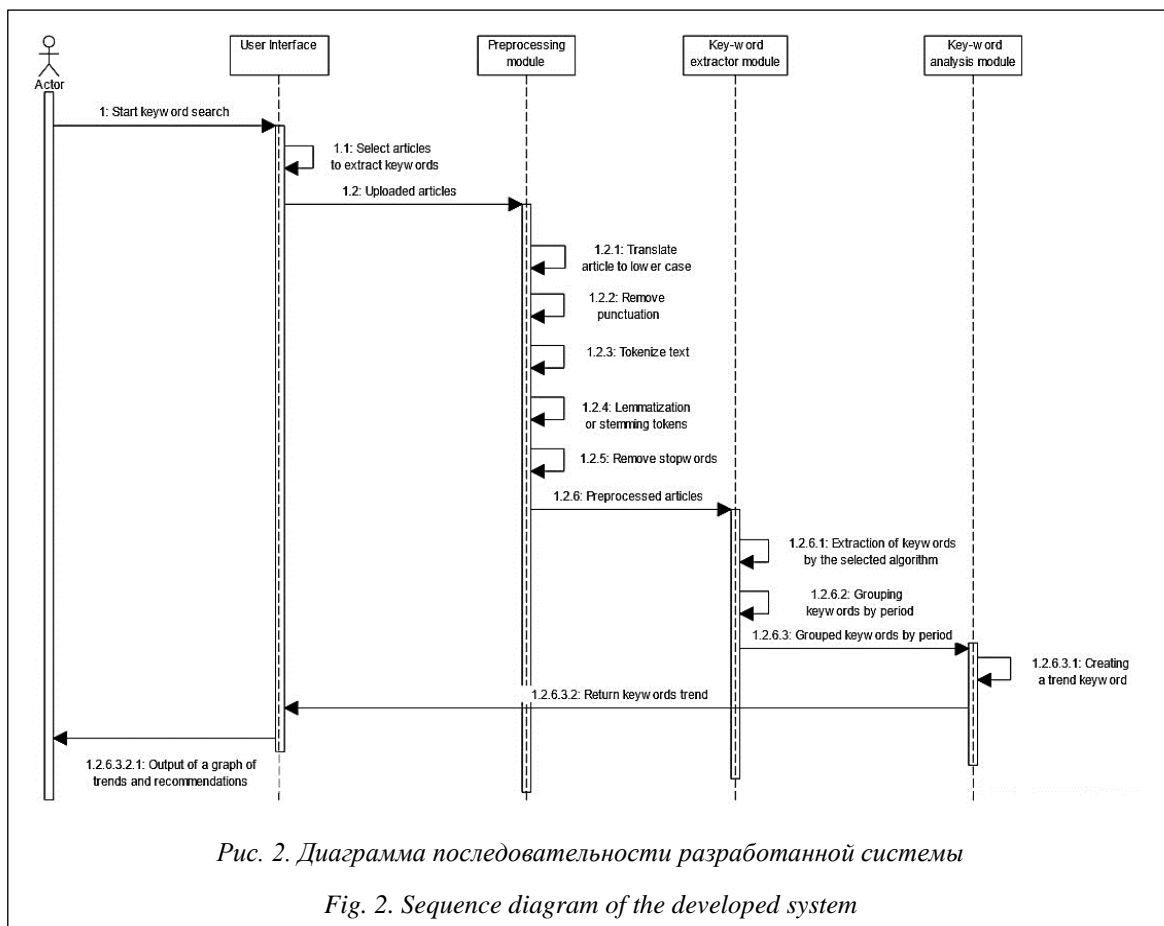


Рис. 2. Диаграмма последовательности разработанной системы

Fig. 2. Sequence diagram of the developed system

ется сбор исходных текстов, потом их предобработка – удаление стоп-слов, лемматизация и т.д. Затем выполняется векторизация текстов, где каждое слово преобразуется в числовой вектор, а после нее – кластеризация, где тексты группируются по их сходству. Далее проводится моделирование тематик, с помощью которого выделяются основные темы в текстах. В конце извлекаются ключевые слова, наиболее полно отражающие содержание текстов. Таким образом, диаграмма последовательности анализа тенденций ключевых слов позволяет наглядно представить этапы обработки и структурирования информации.

На главной форме сервиса пользователю представлены основные функции. Для извлечения ключевых слов необходимо загрузить статьи в формате CSV и выбрать соответствующий алгоритм извлечения ключевых слов. Кроме того, пользователь имеет возможность настраивать дополнительные параметры предобработки и извлечения. Все эти функции представлены на главной форме приложения, которая является удобным наглядным средством взаимодействия с сервисом (<http://www.swsys.ru/uploaded/image/2025-2/1.jpg>).

Результаты экспериментов

В качестве входных данных были использованы три различных набора данных: научные статьи из eLibrary, статьи с сайта Lenta.ru и научные статьи из датасета Krapivin (табл. 1).

Набор данных с сайта Lenta.ru состоит из русскоязычных научно-популярных текстов и является размеченным набором, включающим пять категорий.

В набор данных Krapivin входят англоязычные статьи по тематике Computer science. Особенностью датасета является наличие полных текстов статей и эталонных ключевых слов для каждой статьи.

Целью первого эксперимента было выявление наилучшего способа предобработки текста. Для этого выбраны восемь алгоритмов предобработки текста, после чего произведена K-means кластеризация (метод векторизации TF-IDF) полученных текстов и оценена ее точность [2] (табл. 2).

По результатам первого эксперимента можно сделать вывод, что первый вариант предобработки текста оказался наилучшим. Повторное удаление стоп-слов дает небольшой при-

Таблица 1

Описание входных наборов данных

Table 1

Description of input datasets

Характеристика	Elibrary	Lenta.ru	Krapivin-A	Krapivin-T
Размер корпуса текстов	12 356	25 000	2 293	2 293
Тематика	Научные публикации	Различная	Computer science	
Состав документа	Аннотации	Полный текст	Аннотации	Полный текст
Среднее количество токенов в документе	149,4	260,31	169,1	8597,6
Среднее количество ключевых слов в тексте	n/a	n/a	5,34	
Среднее количество токенов в заголовке	8,1	5,9	8,78	

Таблица 2

Результаты первого эксперимента

Table 2

Results of the first experiment

Метод предобработки	Номер эксперимента							
	1	2	3	4	5	6	7	8
Приведение к нижнему регистру	+	+	+	+	+	+	+	+
Токенизация	+	+	+	+	+	+	+	+
Удаление стоп-слов и знаков пунктуации	+	+	+	+	-	-	+	-
Лемматизация	+	-	+	-	+	-	-	-
Стемминг	-	+	-	+	-	+	-	-
Повторное удаление стоп-слов и знаков пунктуации	+	+	-	-	-	-	-	-
Точность классификации K-means (ARI)	0,74	0,72	0,71	0,7	0,62	0,61	0,55	0,51

рост точности в пару процентов. Стемминг дает почти такую же точность, как и лемматизация, но лемматизация является более затратной операцией, из-за чего время обработки текстов увеличивается.

Для оценки эффективности извлечения ключевых слов было проведено сравнение его алгоритмов. Для оценки алгоритмов использовались метрика F-score и наборы данных Krapivin-A и Krapivin-T, включающие эталонные ключевые слова для каждой статьи. Итоги эксперимента приведены в таблице 3.

По результатам эксперимента можно сделать вывод, что статистические методы извлечения ключевых слов плохо подходят для решения задачи анализа состояния сложных технических систем. Ключевые слова, извлеченные с помощью таких алгоритмов, довольно сильно отличаются от эталонных ключевых слов и необъективно отражают содержание

текстов, поскольку извлеченные ключевые слова в большей степени являются общими для всех текстов и не дают описания для конкретных предметных областей.

Таблица 3

Результаты извлечения ключевых слов с помощью статистических методов

Table 3

Keyword extraction results using statistical methods

Алгоритм	F1-score	
	Krapivin-A	Krapivin-T
TF-IDF	0,35	0,28
Rake	0,29	0,31
Yake	0,27	0,28
TextRank	0,33	0,25
Биграммы PMI	0,39	0,34
Биграммы T-Score	0,32	0,27
Биграммы LLR	0,31	0,3

В эксперименте по кластеризации текстовых данных проводилось сравнение точности различных методов классификации с различными методами векторизации текста и с полученным в первом эксперименте методом предобработки текста. Для проведения экспериментов были выбраны пять тематик: экономика, мир, наука и техника, путешествия, ценности. По каждой тематике отобрано по 400 статей. Всего получено 2 000 статей.

Для эксперимента были использованы пять алгоритмов кластеризации: K-means, DBSCAN, Spectral, Agglomerative, MiniBatchK-means. Предобработка текста включала следующие методы: стемминг, лемматизация, стоп-листинг [3]. В качестве методов оценки точности кластеризации по размеченным данным использованы пять метрик: ARI, AMI, Homogeneity, Completeness, V-measure. Результаты данного эксперимента представлены в таблице 4.

Наибольшую точность дает модель, включающая в себя векторизацию методом Word2Vec и агломеративную кластеризацию. Связка данных методов в среднем дает точность порядка 80 %.

В последнем эксперименте кластеризация научных статей проводилась с помощью методов тематического моделирования. В данной работе применялся метод LDA. Для эксперимента были выбраны статьи за 2017–2021 гг. и произведено тематическое моделирование на пять нечетких тематик. Результаты данного эксперимента представлены в таблице 5.

Из построенной диаграммы распределения статей по годам и тематикам (<http://www.swsys.ru/uploaded/image/2025-2/14.jpg>) можно сделать вывод о наиболее популярных в разные годы тематиках, а также о динамике каждой из них в разрезе нескольких лет.

сделать вывод о наиболее популярных в разные годы тематиках, а также о динамике каждой из них в разрезе нескольких лет.

Таблица 5

Тематика статей за 2017–2021 гг.

Table 5

Topics of articles for 2017–2021

Номер темы	Тематика
0	страниц, сайт, интернет, электрон, учет
1	мем, изображен, интернет, воздушн, текст
2	граф, вершин, оптимальн, матриц, уравнен
3	агент, модель, вычислен, возмущен, устойчив, интеллект
4	нейрон, сервис, преобразован, плоскост
5	запрос, кластер, классификац, обслуживан, центр

Проведенные эксперименты показали, что методы извлечения ключевых слов дают недостаточные точность и качество ключевых слов для дальнейшего построения OWL-онтологий. Классические методы кластеризации показывают довольно хорошую точность, но несколько сложны в интерпретации результата, поскольку после кластеризации требуется дополнительная обработка данных, чтобы получить осознанные тематики кластеров. По результатам эксперимента наиболее осознанные кла-

Таблица 4

Результаты эксперимента по кластеризации текстов

Table 4

Activity diagram of the software system for searching anomalies of technical time series

Алгоритм кластеризации	Метрика				
	ARI	AMI	Homogeneity	Completeness	V-measure
TF-IDF					
K-means	0,74	0,842	0,819	0,869	0,843
DBSCAN	0,237	0,238	0,238	0,241	0,239
Spectral	0,636	0,755	0,724	0,792	0,756
Agglomerative	0,578	0,745	0,695	0,811	0,747
MiniBatchK-means	0,506	0,685	0,644	0,738	0,686
Word2Vec					
K-means	0,722	0,782	0,747	0,825	0,783
DBSCAN	0,2366	0,26	0,25	0,513	0,262
Spectral	0,67	0,749	0,714	0,793	0,751
Agglomerative	0,808	0,849	0,813	0,893	0,85
MiniBatchK-means	0,808	0,821	0,819	0,825	0,822

стеры и тематики показал метод тематического моделирования.

Результаты позволяют сделать вывод, что методы тематического моделирования могут помочь в извлечении терминов предметных областей и онтологических структур из текстов с целью дальнейшего применения полученных онтологий при анализе состояния сложных технических систем.

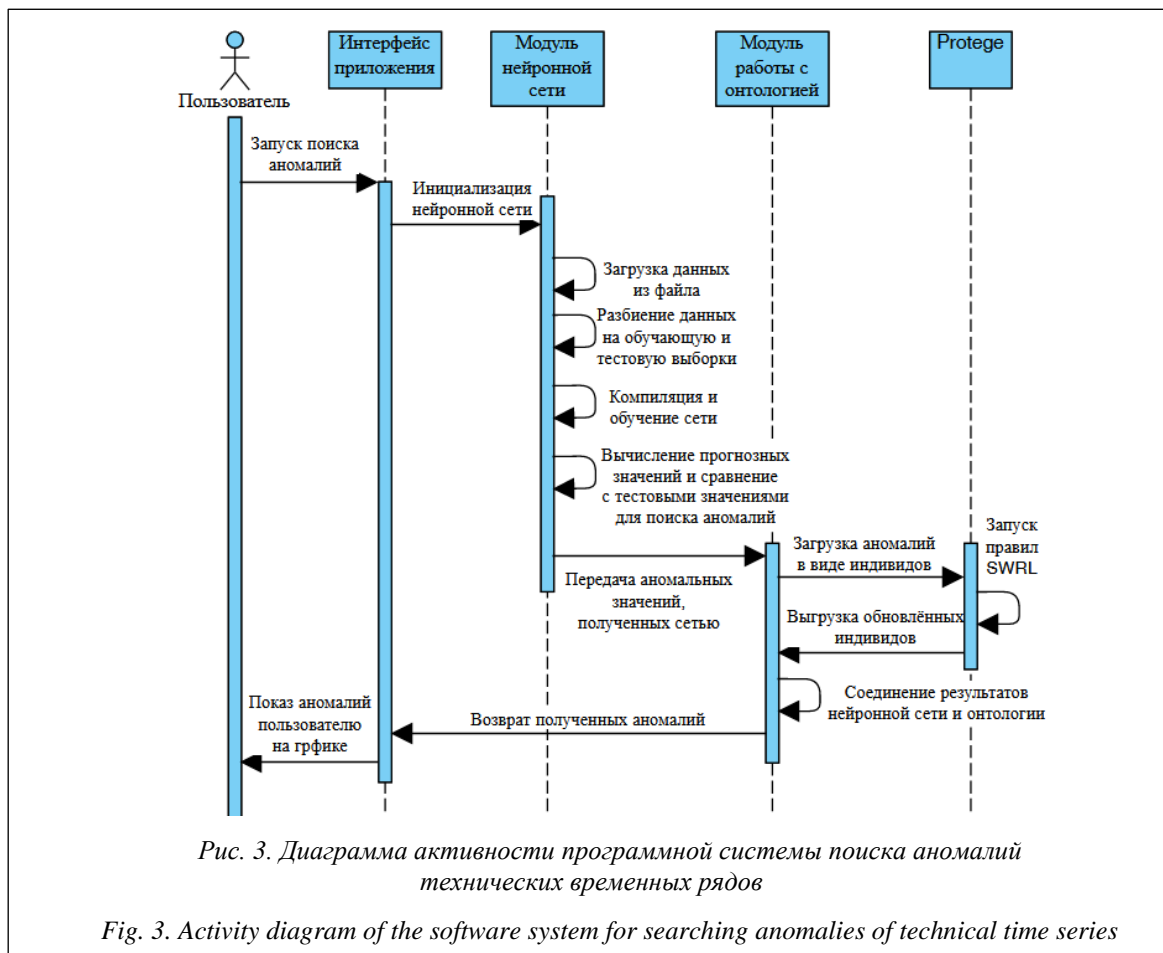
В настоящее время OWL-онтологии часто используются для решения прикладных задач по интерпретации результатов анализа числовых данных, полученных с помощью классических математических или нейросетевых методов, посредством учета особенностей предметной области, описываемой в онтологии. Например, в [11] рассматривается подход к решению задачи обнаружения аномалий временных рядов с учетом специфики предметной области. В работе предлагается метод, основанный на интеграции нейронной сети с долговременной краткосрочной памятью (LSTM) и онтологии Fuzzy OWL. На рисунке 3 представлена диаграмма активности программной системы, реализующей предложенный подход.

На первом этапе для математического поиска аномалий используется сеть LSTM. На втором нечеткая онтология фильтрует результаты обнаружения и формирует вывод для принятия решения. Онтология содержит формализованное представление объектов предметной области и правила вывода, которые отбирают только те значения аномалий, которые соответствуют данной предметной области.

Вычислительные эксперименты проводились на свободных данных технических характеристик буровых установок и показали достаточно высокую эффективность, что говорит о применимости предметных OWL-онтологий совместно с иными.

Заключение

В работе рассмотрена проблема реализации различных форм автоматического анализа текстовых данных для решения задачи построения OWL-онтологий предметной области. Дальнейшими целями проекта являются проверка возможности применения дообученных больших языковых моделей для извлечения терми-



нов предметных областей, а также извлечение семантических отношений между полученными терминами и определение OWL-аксиом.

Ожидается, что результаты данного исследования позволят создать эффективную систему автоматического анализа текстовых дан-

ных и построения OWL-онтологий в различных предметных областях. Это может способствовать автоматизации процесса создания онтологий, сократить затраты времени и ресурсов на этапе их разработки, а также улучшить качество полученных моделей.

Список литературы

1. Николаев А.В., Жуков В.В. Тематическая кластеризация научной литературы // ИТТММ: матер. конф. 2021. С. 268–273.
2. Юферев В.И., Разин Н.А. Векторизация текстов на основе word-embedding моделей с использованием кластеризации // Моделирование и анализ информационных систем. 2021. Т. 28. № 3. С. 292–311. doi: 10.18255/1818-1015-2021-3-292-311.
3. Кравченко Ю.А., Мансур А.М., Хуссейн М.Ж. Векторизация текста с использованием методов интеллектуального анализа данных // Изв. ЮФУ. Технич. науки. 2021. № 2. С. 154–167.
4. Мусаев А.А., Григорьев Д.А. Обзор современных технологий извлечения знаний из текстовых сообщений // Компьютерные исследования и моделирование. 2021. Т. 13. № 6. С. 1291–1315. doi: 10.20537/2076-7633-2021-13-6-1291-1315.
5. Леонов Н.А., Захарова И.Г. Методы и алгоритмы для построения онтологии для QA-системы // Математическое и информационное моделирование: матер. конф. 2023. № 21. С. 86–93.
6. Астанин П.А., Раузина С.Е., Зарубина Т.В. Автоматизированная система извлечения клинически релевантных терминов UMLS из текстов англоязычных статей на примере аксиального спондилоартрита // Социальные аспекты здоровья населения. 2023. Т. 69. № 3. Ст. 14. doi: 10.21045/2071-5021-2023-69-3-14.
7. Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon. Proc. LREC'16, 2016, pp. 1171–1176.
8. Moshkin V.S., Zarubin A.A., Koval A.R., Filippov A.A. Construction of the problem area ontology based on the syntagmatic analysis of external wiki-resources. Proc. Int. Conf. DS-ITNT, 2017, pp. 128–134. doi: 10.18287/1613-0073-2017-1903-128-134.
9. Solovyev V., Loukachevitch N. Comparing similarity of words based on psychosemantic experiment and RuWordNet. Proc. 11th Global Wordnet Conf., 2021, pp. 199–206.
10. Гусев П.Ю. Разработка системы классификации текстов по научным специальностям с применением методов машинного обучения // Вестн. НГУ. Сер.: Информационные технологии. 2021. Т. 19. № 1. С. 39–47.
11. Moshkin V., Kurilo D., Yarushkina N. Integration of fuzzy ontologies and neural networks in the detection of time series anomalies. Math., 2023, vol. 11, no. 5, art. 1204. doi: 10.3390/math11051204.

Constructing a subject OWL ontology: Comparing the effectiveness of different text data mining algorithms

Aleksandr A. Dyrnochkin ¹, Vadim S. Moshkin [✉], Nadezhda G. Yarushkina ¹

¹ Ulyanovsk State Technical University, Ulyanovsk, 432027, Russian Federation

For citation

Dyrnochkin, A.A., Moshkin, V.S., Yarushkina, N.G. (2025) 'Constructing a subject OWL ontology: Comparing the effectiveness of different text data mining algorithms', *Software & Systems*, 38(2), pp. 234–242 (in Russ.). doi: 10.15827/0236-235X.150.234-242

Article info

Received: 14.02.2024

After revision: 23.09.2024

Accepted: 03.10.2024

Abstract. The paper describes a software service for analyzing textual information in order to form OWL-ontology when analyzing the state of complex technical systems, for example, the subject domain of oil production systems that output the real-time data. The main direction of the service is processing textual information to form OWL-ontology. OWL-ontology is a formal descriptive model that allows semantical structuring and formalizing information. The paper presents the basic principles of this service functioning. The software service performs preprocessing of textual data, including tokenization, stop word removal, lemmatization and key term extraction. These operations are necessary to improve the quality and integrity of data before further analysis. The main stage of the service is to extract keywords and terms from texts using machine learning algorithms and statistical methods. The service groups texts by similarity and forms clusters. This stage allows revealing a hidden structure in textual data and identifying common themes or concepts. The described experimental

results confirm that the software service is effective for forming ontological structures. They demonstrate the ability of the service to classify and group text data, which is an important step in the process of creating semantic models of complex technical systems. The paper also provides a comparative study of three different concepts of keyword and term extraction: statistical keyword extraction methods, clustering and topic modeling. The authors evaluate the performance of each method based on keyword extraction accuracy, relatedness and understandability of key terms. The study results conclude the advantages and disadvantages of each approach. In addition, they allow determining the most productive method for specific tasks of ontology building in information systems.

Keywords: text mining, ontology, preprocessing, clustering, vectorization, topic modeling, keyword extraction, thematic clusters, semantic model, information structuring

Acknowledgements. The work was supported by RSF, project no. 23-71-01101, <https://rscf.ru/project/23-71-01101/>

References

1. Nikolaev, A.V., Zhukov, V.V. (2021) 'Thematic clustering of scientific literature', *Proc. Conf. ITTM*, pp. 268–273 (in Russ.).
2. Yuferev, V.I., Razin, N.A. (2021) 'Word-embedding based text vectorization using clustering', *Modeling and Analysis of Information Systems*, 28(3), pp. 292–311 (in Russ.). doi: 10.18255/1818-1015-2021-3-292-311.
3. Kravchenko, Yu.A., Mansur, A.M., Khussain, M.Z. (2021) 'Text vectorization using methods of intelligent data analysis', *Izv. SFeDU. Eng. Sci.*, (2), pp. 154–167 (in Russ.).
4. Musaev, A.A., Grigoriev, D.A. (2021) 'Extracting knowledge from text messages: Overview and state-of-the-art', *Computer Research and Modeling*, 13(6), pp. 1291–1315 (in Russ.).
5. Leonov, N.A., Zakharova, I.G. (2023) 'Methods and algorithms for constructing a QA system ontology', *Proc. Conf. Mathematical and Information Modeling*, (21), pp. 86–93 (in Russ.).
6. Astanin, P.A., Rauzina, S.E., Zarubina, T.V. (2023) 'Automated system for recognizing clinically relevant UMLS terms in texts of the english-language articles exemplified by axial spondyloarthritis', *Social Aspects of Population Health*, 69(3), art. 14 (in Russ.). doi: 10.21045/2071-5021-2023-69-3-14.
7. Loukachevitch, N., Levchik, A. (2016) 'Creating a general Russian sentiment lexicon', *Proc. LREC'16*, pp. 1171–1176.
8. Moshkin, V.S., Zarubin, A.A., Koval, A.R., Filippov, A.A. (2017) 'Construction of the problem area ontology based on the syntagmatic analysis of external wiki-resources', *Proc. Int. Conf. DS-ITNT*, pp. 128–134. doi: 10.18287/1613-0073-2017-1903-128-134.
9. Solovyev, V., Loukachevitch, N. (2021) 'Comparing similarity of words based on psychosemantic experiment and RuWordNet', *Proc. 11th Global Wordnet Conf.*, pp. 199–206.
10. Gusev, P.Yu. (2021) 'Development of a text classification system by scientific specialties using machine learning methods', *Vestn. NSU. Ser.: Information Technologies*, 19(1), pp. 39–47 (in Russ.).
11. Moshkin, V., Kurilo, D., Yarushkina, N. (2023) 'Integration of fuzzy ontologies and neural networks in the detection of time series anomalies', *Math.*, 11(5), art. 1204. doi: 10.3390/math11051204.

Авторы

Дырnochкин Александр Александрович¹,
аспирант, ассистент, dyrno4kin@gmail.com
Мошкин Вадим Сергеевич¹, к.т.н.,
доцент, postforvadim@ya.ru
Ярушкина Надежда Глебовна¹,
д.т.н., профессор,
профессор кафедры, jng@ulstu.ru

Authors

Aleksandr A. Dyrnochkin¹, Postgraduate Student,
Assistant, dyrno4kin@gmail.com
Vadim S. Moshkin¹, Cand. of Sci. (Engineering),
Associate Professor, postforvadim@ya.ru
Nadezhda G. Yarushkina¹,
Dr.Sci. (Engineering), Professor,
Professor of Department, jng@ulstu.ru

¹ Ульяновский государственный технический университет, г. Ульяновск, 432027, Россия

¹ Ulyanovsk State Technical University, Ulyanovsk, 432027, Russian Federation