

СРАВНЕНИЕ РУССКОЯЗЫЧНЫХ ТЕКСТОВ С ПОМОЩЬЮ ГРАФА СЛОВСОЧЕТАНИЙ ДЛЯ ВЫЯВЛЕНИЯ ОТЛИЧИТЕЛЬНЫХ СМЫСЛОВЫХ ФРАГМЕНТОВ¹

Н.В. Мелешенко (*meleshenko.nikolay@mail.ru*)

О.И. Федяев (*olegfedyayev@mail.ru*)

Донецкий национальный технический университет, Донецк

В работе рассматривается проблема обновления университетских учебных программ с учётом требований (рекомендаций) предприятий. Предложен подход, основанный на представлении текста в виде графа словосочетаний, который позволяет визуализировать связи между терминами и улучшить процесс анализа. Отличием от предыдущих работ является использование дерева составляющих вместо дерева зависимостей для формализации извлечения словосочетаний. Решены проблемы нормализации извлеченных терминов и обработки конъюнкций в тексте, что способствует более точному определению словосочетаний. Проведенные экспериментальные исследования подтверждают эффективность предложенного подхода.

Ключевые слова: естественный язык, сравнение текстов, требования предприятий, учебные программы дисциплин, граф словосочетаний, нормализация терминов, смысловые фрагменты.

Введение

Представленная работа затрагивает проблему инновации университетских учебных программ дисциплин путём учёта новых требований со стороны предприятий. Регулярное обновление учебных программ дисциплин обеспечивает высокий уровень профессиональной подготовки выпускников, востребованных на рынке труда. В связи с этим возникает необходимость в регулярной и профессионально-ориентированной кооперации кафедры и профильных предприятий для решения данной проблемы. Взаимодействие выпускающей кафедры с предприятиями осуществляется на

¹ Данная работа выполняется по плану Молодёжной научной лаборатории «Искусственный интеллект» ДонНТУ. Научная работа № FRRF-2024-0010.

уровне смыслового анализа текстовых документов (рекомендации предприятий, рабочие программы дисциплин) и извлечения новых компетенций (технологий, методов, инструментов и др.) для корректировки соответствующих рабочих программ дисциплин.

Появление новых методов обработки текстовой информации и соответствующих инструментов сделало возможным автоматизировать решение задачи по интеллектуальной поддержке формирования и обновления образовательных программ, в том числе и рабочих программ дисциплин (РПД). В одной из первых отечественных работ этого направления [Космачёва и др., 2016] была предложена интерактивная система формирования РПД, которая использовала простые трансформационные методы обработки информации и ограничивалась проверкой РПД на соответствие критериям качества ФГОС. Формальные методы описания текстового документа при помощи графа были рассмотрены в работе [Sheetal et. al., 2014]. В работе [Ботов, 2019] использовались современные нейросетевые модели языка word2vec, но они применялись только для оценки семантической близости анализируемых документов. Извлечению с помощью нейросетевой модели BERT коротких фрагментов знаний и навыков из текстов требований онлайн-вакансий посвящена интересная работа [Николаев, 2023]. Однако в ней не решён вопрос насколько эти знания будут новыми по отношению к РПД. В статье [Wu et. al., 2024] авторы предлагают семантику текста представлять с помощью графа знаний из фраз, полученных при помощи вероятностной контекстно-свободной грамматики и алгоритма СКУ. Однако данный подход был протестирован авторами только в задачах классификации и кластеризации текстов.

В предшествующей работе авторов [Федяев и др., 2025] рассмотрен один из подходов к решению задачи инновации РПД. Процесс решения заключался в сравнении семантик текстов требований предприятия и рабочих программ дисциплин с целью получения разницы в знаниях, представленных в этих документах. Идея решения основывается на представлении текста в виде графа – семантической сети, отражающей синтаксические связи между словами в тексте. Разница в знаниях рассматривается как разность графов двух документов, при этом учитываются не только отдельные слова, но и словосочетания. В используемом подходе было выявлено несколько не решённых проблемных вопросов:

- процесс извлечения новых терминов из текста не был формализован, а использование дерева зависимостей накладывает неопределённость из-за своей непостоянной структуры вследствие ошибок частеречной и морфологической разметок [Демидов, 2023];
- выделение и преобразование смысловых фрагментов осуществлялось только в тексте требований предприятий, но лучшим решением будет представление текста требований и РПД в одном формате для сравнения их при помощи вычисления разности графов;

- отсутствие нормализации извлеченных терминов приводит к ухудшению смысловой оценки текста.

Поэтому целью данной работы является представление текстов рабочих программ дисциплин и требований предприятий в новой форме – в виде графов словосочетаний, позволяющих повысить формализацию выявления и сравнения представленных в них знаний.

1. Особенности представления текста в виде дерева составляющих

Текст может быть представлен при помощи синтаксического анализа в двух видах: дерева зависимостей и дерева составляющих [Кравченко и др., 2024]. В работе [Федяев и др., 2025] текст изначально представлялся в виде дерева зависимостей, но в данной работе мы предлагаем использовать дерево составляющих, которое можно получить при помощи набора жадных (greedy) регулярных выражений (см. листинг 1), являющихся подобием контекстно-свободной грамматики.

Дерево составляющих основано на формализме контекстно-свободных грамматик и может быть построено автоматически, используя программные средства (язык программирования Python, библиотеки SpaCy и NLTK) синтаксического анализа на основе регулярных выражений. В таком дереве предложение делится на составляющие, т.е. фразы, которые относятся к определенной категории в грамматике.

Листинг 1

```
1: AP: {<A>+}
2: ACC: {<CC|COMMA><AP>}
3: APCC: {<AP><ACC>+}
4: NPG: {<NG><NG|FN>*}
5: NP: {<N><NPG|FN>*}
6: ANP: {<AP|APCC><NP>}
7: ANPG: {<AP|APCC><NPG>}
8: TERM: {<ANP|NP><ANPG>*}
9: TERMCC: {<CC|COMMA><TERM>}
10: TERMG: {<ANPG|NPG><ANPG>*}
11: TERMGCC: {<CC|COMMA><TERMG>}
12: COMTERM: {<TERM><TERMCC>+}
13: COMTERMEG: {<COMTERM><TERMGCC>+}
14: COMTERMG: {<TERM><TERMGCC>+}
15: FNP: {<FN>+}
16: AFNP: {<AP|APCC><FNP>}
17: FTERM: {<AFNP|FNP>}
18: FTERMCC: {<CC|COMMA><FTERM>}
```

По своей сути грамматика позволяет строить правильные предложения и извлекать их синтаксическую структуру [Кравченко и др., 2024], [Полетаев и др., 2023].

Подводя итог, можно отметить, что дерево составляющих, в отличие от дерева зависимостей, содержит синтаксическое представление предложения в соответствии с заданной контекстно-свободной грамматикой, что позволяет формализовать извлечение словосочетаний из текста. Такое представление имеет чёткую иерархию и делит предложения на отдельные фразовые составляющие [Wu et. al., 2024].

Приведенные выше регулярные выражения содержат правила, включающие разные синтаксические категории (табл. 1).

Таблица 1

Категория	Описание	Пример
CC	Союз	и, или
COMMA	Запятая	,
A	Прилагательное	синий
FN	Иностранное существительное	data, text
N	Существительное	метод
NG	Существительное в родительном падеже	метода
AP	Последовательность прилагательных	большой синий
ACC	Конъюнкция прилагательного	и большой синий
APCC	Конъюнкция набора прилагательных	черный средний и большой синий
NPG	Именная группа в родительном падеже	анализа данных
NP	Именная группа	метод анализа данных
ANP	Группа прилагательного	эффективный метод анализа данных
ANPG	Группа прилагательного в родительном падеже	больших данных
TERM	Термин	основные стратегии обучения нейронных сетей
TERMCC	Конъюнкция термина	и бинарное дерево
TERMG	Термин в родительном падеже	структур нейронных сетей
TERMGCC	Конъюнкция термина в родительном падеже	и структур данных
COMTERM	Конъюнкция терминов	обработка и анализ больших данных
COMTERMEG	Конъюнкция терминов и конъюнкция терминов в родительном падеже	разработка и тестирование прикладного программного обеспечения и системных решений
COMTERMG	Термин и конъюнкция терминов в родительном падеже	разработка моделей, алгоритмов и программ
FNPN	Иностранная именная группа	text mining

Дерево составляющих используется нами для выделения основных смысловых конструкций – терминов. Под словом «термин» мы условились понимать такое максимально длинное словосочетание, удовлетворяющее регулярному выражению, которое начинается с существительного в именительном падеже и заканчивается, по возможности, существительным в родительном падеже, включая прилагательные, описывающие каждое существительное в термине. Например, дерево составляющих для предложения «Основные стратегии обучения нейронных сетей» представлено на рис. 1.

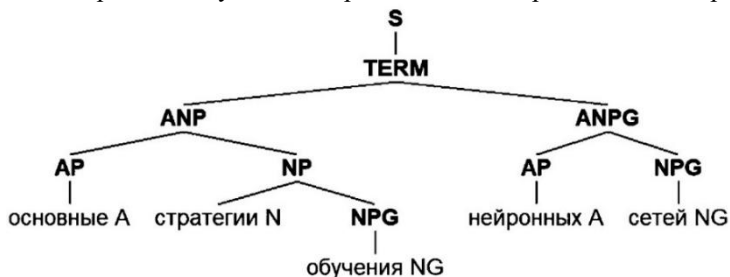


Рис. 1. Дерево составляющих предложения

Как видно из рисунка данное предложение содержит одну категорию TERM (термин). Термины рассматриваются нами как основные лексические единицы текста, поэтому текст можно представить как совокупность терминов [Wu et. al., 2024]. Эту совокупность можно в дальнейшем анализировать с целью выявления ключевых фраз и слов, а также осуществлять их поиск, если представить текст как граф терминов и их составляющих, что мы и используем в этой работе.

2. Преобразование дерева составляющих в граф словосочетаний

Главным недостатком, как дерева составляющих, так и дерева зависимостей является то, что они строятся для одного предложения, а не для всего текста [Wu et. al., 2024]. Поэтому необходимо определить структуру данных и алгоритм преобразования множества деревьев составляющих в общую структуру в качестве представления всего текста [Григорьева и др., 2023].

Представим текст в виде орграфа $G(V, E)$, где:

V – множество вершин, представляющих собой слова и образуемые из них словосочетания – термины, которые встречаются в исходном тексте;

E – множество ребер, которые указывают на формирование фразы $\langle \quad \rangle$, где pos – позиция части словосочетания a в словосочетании b . Графическое представление данного графа представлено на рис. 2.

Таким образом, данную структуру данных можно назвать графом словосочетаний, который представляет весь текст в виде набора терминов и словосочетаний, которые его формируют.

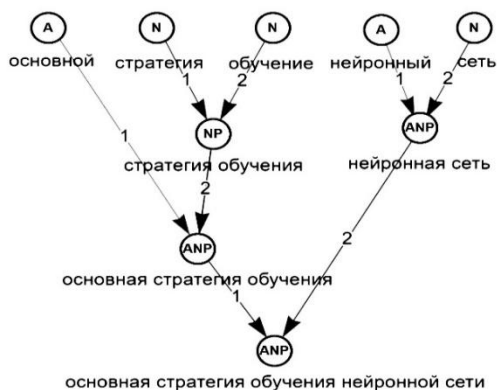


Рис. 2. Графическое представление графа словосочетаний

На рисунке видны существительные (стратегия, обучение и сеть), прилагательные (основной, нейронный), именная группа (стратегия обучения) и группы прилагательных (нейронная сеть, основная стратегия обучения, основная стратегия обучения нейронной сети). Данное представление текста имеет следующие преимущества:

- 1) возможность наглядной визуализации и простота интерпретации графа человеком;
- 2) данная структура данных позволяет применять различные алгоритмы на графах для анализа и обработки данных текста;
- 3) представление текста в виде графа даёт возможность при помощи операций над графами (разность, пересечение, объединение) определять различия в словосочетаниях, общие используемые термины и составить общий словарь словосочетаний для двух текстов [Sheetal et. al., 2014].

При построении графа для каждой вершины учитывается количество вclusions данного слова или словосочетания в другие словосочетания в тексте [Кравченко и др., 2024]. После данной операции мы можем для каждого ребра рассчитать частоту расширений словосочетаний, т.е., например, как часто слово «сеть» расширяется до «нейронная сеть» [Григорьева и др., 2023]. Таким образом, появляется возможность дальнейшего преобразования полученного графа путём выделения наиболее частых выражений в отдельные вершины. Программная модель графа реализована при помощи библиотеки NetworkX и визуализирована с помощью библиотеки PyVis.

3. Обработка конъюнкций

Конъюнкция – это операция образования сложных высказываний из более простых и по смыслу эквивалентная соединительному союзу «и» в естественном языке [Студеникина, 2018]. Конъюнкты нами считаются альтернативными составляющими для исходных словосочетаний, поэтому создаются различные вершины для одного и того же исходного выражения, но с различными составляющими так, как если бы это были отдельные выражения в тексте. На рис. 3 продемонстрировано дерево составляющих для предложения «Методы качественной оценки и способы обеспечения безопасности программ и данных».

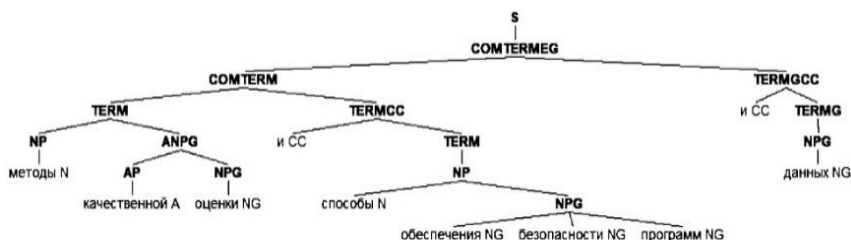


Рис. 3. Дерево составляющих для предложения с конъюнкциями

На рисунке видно, что в предложении обнаружен один комбинированный термин, состоящий из набора терминов в именительном падеже («методы качественной оценки», «способы обеспечения безопасности программ») и термина в родительном падеже («данных»), между терминами стоят союзы «и». Используя наш подход, из данного предложения можно выделить 4 термина (см. рис. 4): «метод качественной оценки безопасности программы», «способ обеспечения безопасности программы», «метод качественной оценки безопасности данных» и «способ обеспечения безопасности данных».

Для обработки конъюнкций использовались такие операции сложения словосочетаний [Студеникина, 2018]:

Во всех иных случаях словосочетания считаются не принадлежащими конъюнкции и данные операции для формирования отдельных словосочетаний не применяются.

Данный подход позволяет выделять обособленные, несвязанные между собой термины, которые в тексте могли находиться в конъюнкции и, вследствие этого, могли быть захвачены как один связанный термин, что затрудняет смысловую оценку текста из-за наличия сочинительных союзов или запятых в термине.

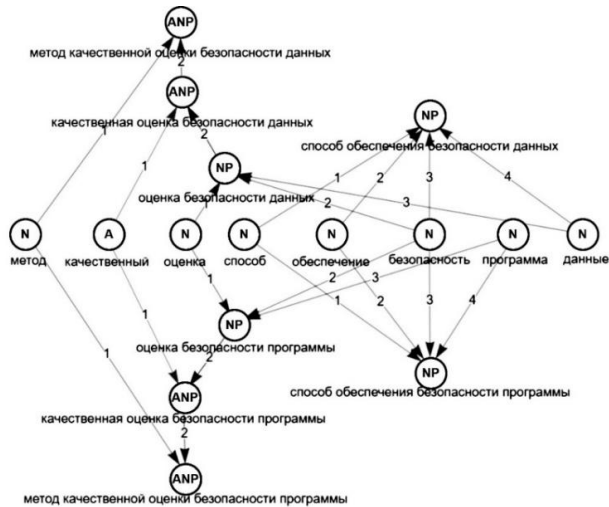


Рис. 4. Граф словосочетаний для предложения с конъюнкциями

4. Нормализация выделенных словосочетаний

Важным аспектом для понимания человеком результатов выделения словосочетаний является их нормализация. В работе [Федяев и др., 2025] при формировании графа смысловых фрагментов используются только исходные выражения из текста требований, что может привести к ошибочной интерпретации терминов из-за омонимии [Демидов, 2023], поэтому необходимо нормализовать данные выражения. Для этого в работе использовался морфологический анализатор Руморфhy2, написанный на языке Python. Схема нормализации представлена на рис. 5.

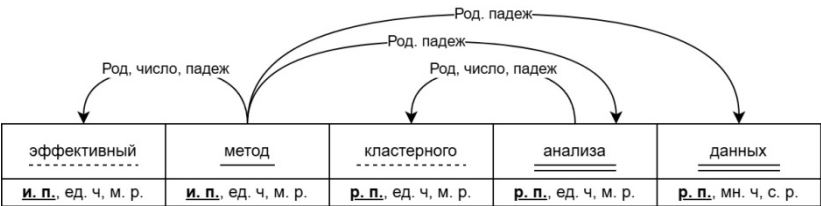


Рис. 5. Схема нормализации именной группы

На рисунке одной чертой выделено главное слово именной группы – существительное, оно всегда в именительном падеже, т.к. принимает форму леммы (нормальной формы слова), пунктиром выделены прилагательные, а двойной чертой – остальные существительные.

При нормализации в нашем случае используется две связи слов в словосочетаниях:

- 1) согласование, при котором зависимое слово согласуется в роде, числе и падеже с главным (например, «эффективный метод»);
- 2) управление, при котором зависимое слово ставится в том падеже, которого требует главное (например, «метод анализа»).

Главное слово управляет остальными существительными группы, поэтому они употребляются в родительном падеже. Все прилагательные согласуются со следующим существительным по тексту в роде, числе и падеже.

5. Определение новых словосочетаний

Для апробации рассмотренной идеи проведём эксперимент с текстовыми данными, рассмотренными в предыдущей статье авторов [Федяев и др., 2025]. Определим новые словосочетания для одного предложения из текста требований к специалисту по интеллектуальному анализу данных: «Знать методы дискриминантного и кластерного анализа данных». Результаты извлечения новых словосочетаний представлены на рис. 6. Серым выделены известные словосочетания по тексту рабочей программы, а черным – новые словосочетания. Можно сделать два вывода, во-первых, теперь для выявления новых словосочетаний достаточно вычислить разницу графов требований и рабочей программы дисциплины без промежуточных преобразований графа требований, т.к. оба графа имеют одинаковую структуру. Во-вторых, структура графа даёт чёткое понимание, что в тексте рабочей программы нет точной формулировки «методы дискриминантного и кластерного анализа данных», однако есть «кластерный анализ данных».

Это позволяет сказать, что несмотря на то, что все слова, из которых состоит данный термин, присутствуют в рабочей программе, как таковые понятия «метод кластерного анализа данных» и «метод дискриминантного анализа данных» не фигурируют в тексте рабочей программы. В отличие от предыдущей работы, где весь этот термин считается известным из текста рабочей программы (табл. 2).

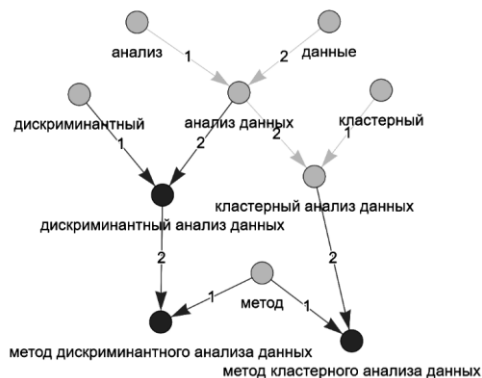


Рис. 6. Определение новых словосочетаний

Таблица 2

Пример анализируемого предложения в тексте требований: «Знать методы дискриминантного и кластерного анализа данных»	Выделенные словосочетания в предложении	
	Подход на основе дерева составляющих	Подход на основе дерева зависимостей
Новые словосочетания для РПД	Метод кластерного анализа данных, метод дискриминантного анализа данных, дискриминантный анализ данных	Не обнаружены
Известные словосочетания для РПД	Кластерный анализ данных, анализ данных	Методы дискриминантного и кластерного анализа данных

Таким образом, данный подход позволяет точнее идентифицировать части словосочетаний, которые присутствуют в обоих документах, и выделять как общие термины, так и те, которыми документы отличаются по смысловому содержанию. Кроме того, он выявляет в 2–3 раза больше словосочетаний, что подтверждает его более высокую детализацию разбиения на смысловые фрагменты.

Заключение

В работе предложен улучшенный способ сравнения текстовых документов, проиллюстрированный на примере извлечения новых смысловых фрагментов из текста требований предприятий по отношению к знаниям и навыкам, представленным в текстах рабочих программах дисциплин выпускающей кафедры. В качестве решения предлагается однородное представление сопоставляемых текстов в виде графов словосочетаний, которые позволяют при помощи разности графов получить отличительные смысловые фрагменты.

Извлечение словосочетаний формализовано при помощи набора правил с использованием регулярных выражений. Для представления всего текста в виде графа словосочетаний разработана структура самого графа и алгоритм преобразования набора деревьев составляющих в единый граф, представляющий весь текст.

Также было проведено экспериментальное сравнение предложенного подхода с ранее опубликованными результатами авторов данной работы. Эксперимент показал, что предложенный подход позволяет более правильно оценить смысловую новизну термина при сравнении документов требований предприятий и рабочей программы дисциплины. В отличие от подхода, используемого авторами в предыдущей работе [Федяев и др., 2025], данный подход благодаря разделению термина на словосочетания в четкой синтаксической иерархии позволяет выделить конкретный фрагмент термина, который является новым по смыслу словосочетанием для текста рабочей программы дисциплины (см. пункт 5). К тому же, обработка конъюнкций позволила разделить термин с включением союза «и» на два отдельных обособленных термина, что было невозможно в прежнем подходе.

Таким образом, новизна и практическая значимость предложенного подхода заключается в применении дерева составляющих вместо дерева зависимостей для извлечения словосочетаний, обработке конъюнкций в тексте и нормализации извлечённых словосочетаний, что в целом позволяет более правильно определять новые смысловые фрагменты для текста рабочей программы дисциплины и повышает их интерпретируемость человеком (лектором).

Список литературы

[Ботов, 2019] Ботов Д.С. Интеллектуальная поддержка формирования образовательных программ на основе нейросетевых моделей языка с учетом требований рынка труда // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2019. – Т. 19, № 1. – С. 5-19. – doi: 10.14529/ctcr190101.

- [Григорьева и др., 2023] Григорьева Е.Г., Клячин В.А., Помельников Ю.В., Попов В.В. Алгоритм выделения ключевых слов на основе графовой модели лингвистического корпуса // Вестник Волгоградского государственного университета. Серия 2: Языкознание. – 2017. – Т. 16, № 2. – С. 58-67. – doi: 10.15688/jvolsu2.2017.2.6.
- [Демидов, 2023] Д.В. Демидов. Представление синтаксических структур с сочинительными конструкциями и омонимией // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2023. – Т. 21, № 4. – С. 17-45. – doi: 10.25205/1818-7900-2023-21-4-17-45.
- [Космачёва и др., 2016] Космачёва И.М., Квятковская И.Ю., Сибикина И.В. Автоматизированная система формирования рабочих программ учебных дисциплин // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2016. – № 1. – С. 90-97.
- [Кравченко и др., 2024] Кравченко Д.Ю. [и др.]. Алгоритм оптимизации извлечения ключевых слов на основе применения лингвистического парсера // Информатика и автоматизация. – 2024. – Т. 23, № 2. – С. 467-494. – doi: 10.15622/ia.23.2.6.
- [Николаев, 2023] Метод извлечения знаний и навыков/компетенций из текстов требований вакансий // Онтология проектирования. – 2023. – Т. 13, № 2(48). – С. 282-293. – doi: 10.18287/2223-9537-2023-13-2-282-293.
- [Полетаев и др., 2023] Полетаев А.Ю., Парамонов И.В., Бойчук Е.И. Алгоритм построения дерева синтаксических единиц русскоязычного предложения по дереву синтаксических связей // Информатика и автоматизация. – 2023. – Т. 22, № 6. – С. 1323-1353. – doi: 10.15622/ia.22.6.3.
- [Студеникина, 2018] Студеникина К.А. Синтаксис сочинения русских именных групп: эллипсис или малые конъюнкты? // Типология морфосинтаксических параметров. – 2018. – Т. 1, № 2. – С. 115-133.
- [Федяев и др., 2025] Федяев О.И., Мелешенко Н.В. Ролевые модели агентов системы моделирования процесса обновления учебных дисциплин с учётом требований предприятий // Проблемы искусственного интеллекта. – 2025. – Т. 36, № 1. – С. 12-25. – doi: 10.24412/2413-7383-12-25.
- [Sheetal et. al., 2014] Sheetal S., Kulkarni P. Graph based Representation and Analysis of Text Document: A Survey of Techniques // International Journal of Computer Applications. – 2024. – Vol. 96. – P. 1-8. – doi: 10.5120/16899-6972.
- [Wu et. al., 2024] Wu Y., Pan X., Li J., Dou S., Dong J., Wei D. Knowledge Graph-Based Hierarchical Text Semantic Representation // International Journal of Intelligent Systems. – 2024. – P. 1-14. – doi: 10.1155/2024/5583270.