**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

На правах рукописи

Андриевская Наталия Климовна

СОВЕРШЕНСТВОВАНИЕ МОДЕЛЕЙ И АЛГОРИТМОВ ОБРАБОТКИ ИНФОРМАЦИИ В СИСТЕМАХ ОРГАНИЗАЦИОННОГО СОПРОВОЖДЕНИЯ ДЕЯТЕЛЬНОСТИ НАУЧНО-ОБРАЗОВАТЕЛЬНЫХ УЧРЕЖДЕНИЙ

Специальность 05.13.01 – Системный анализ, управление и обработка информации (по отраслям) (технические науки)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Донецк – 2021

Работа выполнена в ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ», Министерства образования и науки Донецкой Народной Республики, г. Донецк.

Научный руководитель: кандидат технических наук, доцент
Секирин Александр Иванович,
ГОУ ВПО «ДОННТУ» (г. Донецк),
заведующий кафедрой «Автоматизированные системы управления»

Официальные оппоненты: **Моисеев Дмитрий Владимирович**
доктор технических наук, профессор,
ФГАОУ ВО «Севастопольский государственный университет» (г. Севастополь),
профессор кафедры «Информационные технологии и компьютерные системы»

Ермоленко Татьяна Владимировна
кандидат технических наук, доцент,
ГОУ ВПО «ДОННУ» (г. Донецк),
доцент кафедры компьютерных технологий

Ведущая организация: **ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «ИНСТИТУТ ПРОБЛЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА» (ГУ ИПИИ)** (г. Донецк)

Защита состоится «29» марта 2022 г. в 14 час. 00 мин. на заседании диссертационного совета Д 01.024.04 при ГОУВПО «ДОННТУ» и ГОУВПО «ДОННУ» по адресу: 283001, г. Донецк, ул. Артема, 58, корп.1, ауд. 203. Тел./факс: 380(62) 304-30-55, e-mail: uchensovet@donntu.org.

С диссертацией можно ознакомиться в библиотеке ГОУВПО «ДОННТУ» по адресу: 283001, г. Донецк, ул. Артема, 58, корп. 2. Адрес сайта университета: http://donntu.org

Автореферат разослан «___» _____2022 г.

Ученый секретарь
диссертационного совета Д 01.024.04
кандидат технических наук, доцент Т.В. Завадская

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Организационное сопровождение деятельности научно-образовательных учреждений характеризуется непрерывным ростом количества электронных документов и их общедоступности, в том числе и в среде Интернет. Слабая структурированность информационных фондов осложняет управление информацией и работу пользователей с ней, что напрямую относится к потокам информации, с которыми встречаются сотрудники вузов. В настоящее время кафедрами вузов накоплен большой объем знаний и информационных ресурсов (ИР) по различным курсам и результатам научно-методической работы. Однако отсутствие связанности ИР и унифицированного доступа к ним приводят к возникновению проблем поиска, учета и систематизации как существующих знаний, так и новых.

В связи с этим возникает необходимость создания современного интеллектуального инструмента, поддерживающего повседневную профессиональную деятельность преподавателя. Для решения этой задачи необходим переход на качественно новый уровень представления и обработки информации – семантический, что позволит учитывать смысл документов, извлекая из них важные для пользователя знания.

Таким образом, совершенствование моделей и алгоритмов обработки информации для реализации в системах организационного сопровождения деятельности научно-образовательных учреждений является актуальной научно-технической задачей, имеющей отраслевое значение.

Связь работы с научными программами, планами, темами.

Работа выполнена в соответствии с тематическим планом ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»: Н-17-12 «Разработка основ, методов и средств проектирования информационных управляющих систем»; Н-8-18 «Развитие научных основ, методов и средств проектирования информационных систем и технологий»; Н-2020-16 «Методы и средства построения информационных систем с использованием технологий интеллектуального анализа данных», в которых соискатель являлся исполнителем (справка №06/4-328 от 28.09.2021 г.).

Степень разработанности темы исследования. Среди систем организационного сопровождения деятельности научно-образовательных учреждений в последние годы все чаще встречаются системы управления знаниями (СУЗ), которые явно ориентированы на эффективную работу с ИР и знаниями.

Эта тенденция обусловлена главным образом тем, что знания все больше становятся организационным активом, существование которого позволяет повторно использовать знания, избежать «испарения» знаний и поддерживать принятие решений в учреждении.

Перспективность данного направления подтверждается результатами исследований таких зарубежных учёных, как D. Ameller and X. Franch, M. Bhat,

P. Kruchten, R. Capilla, I. Lytra, H. Tran, U. Zdun и др., а также российских ученых: А.Ф. Тузовского, В.З. Ямпольского, В.А. Лапшина, А.Г. Олейника, Т.А. Гавриловой и др.

Таким образом, можно констатировать, что важность проблематики разработки и совершенствования моделей и алгоритмов обработки информационных ресурсов научно-образовательных учреждений, построенных на основных принципах систем управления знаниями, осознается специалистами, занимающимися информационными технологиями и корпоративным управлением. Тем не менее, до сих пор не существует полного набора моделей и алгоритмов, позволяющих поддерживать работу с явным описанием семантики ИР вузов и других научно-образовательных учреждений, что определяет необходимость дальнейших исследований.

Цель и задачи исследования. Целью диссертационной работы является повышение эффективности системы управления ИР научно-образовательных учреждений за счет применения онтологического подхода, разработки новых и усовершенствования существующих моделей и алгоритмов поиска, хранения и классификации данных.

Для достижения цели поставлены и решены следующие задачи.

1. Провести системный анализ процессов с целью формализации исследуемого объекта и обосновать возможность использования онтологического подхода к построению системы управления ИР учреждения.
2. Разработать онтологическую модель объектов знаний – аппарат для описания семантики области профессиональной деятельности сотрудников научно-образовательных учреждений.
3. Усовершенствовать гибридную меру оценки семантической близости (СБ).
4. Модифицировать векторную модель представления текстов на базе известных подходов bag-of-words и bag-of-concepts, улучшив ее за счет использования онтологической модели предметной области.
5. Усовершенствовать модели и алгоритмы поиска, хранения и классификации данных на основе разработанной онтологической модели.
6. Разработать прототипы программных модулей, реализующие предложенные модели и алгоритмы в виде фреймворка, а также выполнить на базе фреймворка реализацию и тестирование программного модуля учета научной деятельности сотрудников учреждения.

Объект исследования – информационные процессы поиска и обработки ИР научно-образовательного учреждения.

Предмет исследования – модели и алгоритмы реализации информационных процессов и концепции информационного поиска на семантическом уровне с использованием онтологий в системах организационного сопровождения деятельности научно-образовательных учреждений.

Научная новизна полученных результатов заключается в следующем:

1. Впервые разработана онтологическая модель научно-образовательной деятельности сотрудников вуза.

2. Усовершенствована гибридная мера определения семантической близости на базе модифицированной N-мерной модели представления знаний RDF-графа, использование которой повысило качество поиска, выраженное F-мерой, на 10.7% по сравнению с мерой «косинусного сходства».

3. Получила дальнейшее развитие векторная модель представления текстов на базе известных подходов bag-of-words и bag-of-concepts, улучшенная за счет применения онтологии и тематической редукции векторного пространства, что позволило при уменьшении размерности пространства с 2250 терминов до 30 терминов повысить скорость выполнения тестируемых алгоритмов более чем на порядок при незначительном снижении меры семантической близости на 6.2%.

4. Усовершенствована модель классификации данных, основанная на применении гибридной меры определения СБ, что привело к повышению качества классификации, выраженного F-мерой, по сравнению с алгоритмами, использующими меру, вычисленную только по онтологии на 45.4%, «косинусную» меру – на 5.3% и «мягкую косинусную» меру – на 9.5%.

Теоретическая значимость работы.

Теоретическая значимость результатов исследований заключается в развитии моделей и алгоритмов обработки ИР научно-образовательных учреждений и переходу к онтологическому и семантическому моделированию.

Практическая значимость работы.

1. На основе проведенных ранее исследований, разработанных моделей и алгоритмов выполнена программная реализация системы управления ИР в рамках кафедры вуза.

2. Использование документированной прикладной онтологии дает возможность разработчикам систем повторно использовать и развивать данную онтологию, а различным ИС – интегрировать данные и обеспечивать обмен данными на основе онтологии.

3. Предложенные подходы и математические модели, разработанные программные модули фреймворка могут быть применены при создании различных СУЗ, а также систем обработки ИР любых учреждений.

4. Разработанные в ходе выполнения диссертационной работы модели и методы использованы в учебном процессе кафедры автоматизированных систем управления ГОУВПО «ДОННТУ» при выполнении курсовых работ и выпускных квалификационных работ студентов.

5. Разработанный программный модуль «Наука» успешно прошел тестирование в ГУ «Автоматгормаш им. В.А. Антипова» (г. Донецк) в условиях отдела систем управления.

Практическая реализация результатов работы подтверждается справкой №30-12/214 от 10.12.2021 г. о внедрении в учебный процесс ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»,

справкой № 12-319 от 9.06.2021 г. о внедрении в ГУ «Автоматгормаш им. В.А. Антипова» (г. Донецк).

Методология и методы исследования. Для решения поставленных задач использованы методы исследования, основанные на методах системного анализа, онтологического инжиниринга, семантического моделирования, теории графов и множеств, теории экспертных оценок, а также теории нечеткой логики и генетических алгоритмов.

Научные положения, выносимые на защиту.

1. Доказано, что применение усовершенствованной гибридной меры определения семантической близости, использующей модифицированную N-мерную модель представления знаний RDF-графа и генетический алгоритм определения весовых коэффициентов базовых мер, позволило повысить точность определения сходства концептов и улучшить качество поиска, выраженного F-мерой, на 10.7% по сравнению с мерой «косинусного сходства».

2. Определено, что использование техники снижения размерности векторного пространства по тематическим векторам предметной онтологии для векторной модели представления текста при размере онтологии, равном 2250 терминов, и длине контекстного вектора в 30 элементов, приводит к снижению вычислительной сложности тестируемых алгоритмов в десятки раз при незначительном снижении коэффициента СБ с 0.83 до 0.778.

3. Установлено, что при классификации текстовых ИР результаты, полученные с помощью усовершенствованной модели классификации, алгоритм которой использует гибридную меру определения СБ, более точны, чем вычисленные по онтологии, по «косинусной» и «мягкой косинусной» мерам на 45.4%, 5.3% и 9.5% соответственно.

Степень достоверности результатов. Обоснованность и достоверность научных положений, выводов и практических результатов подтверждается полнотой анализа теоретических и практических исследований, разработкой и тестированием программного модуля системы, о чем свидетельствуют справки о внедрении, выполненными публикациями и положительной оценкой на научно-технических конференциях.

По направлению исследований, содержанию научных положений и выводов, существу полученных результатов диссертационная работа соответствует паспорту специальности 05.13.01 – Системный анализ, управление и обработка информации (по отраслям) (технические науки) в частности: п.4 «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации»; п.5 «Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации»; п.8 – «Теоретико-множественный и теоретико-информационный анализ сложных систем».

Апробация. Основные положения диссертационной работы апробированы на научно-технических конференциях: VII Международной

научно-технической конференции студентов, аспирантов и молодых ученых «Информатика и компьютерные технологии» г. Донецк, 22-23 ноября 2011 г.; XI Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование» (ИУСМКМ-2020), г. Донецк, 27-28 мая 2020 г.; 16-й Юбилейной Международной молодежной научно-технической конференции «Современные проблемы радиоэлектроники и телекоммуникаций» (РТ-2020), г. Севастополь, с 12 по 16 октября 2020 г.; III Международной научно-практической конференции «Программная инженерия: методы и технологии разработки информационно-вычислительных систем» (ПИИВС-2020), г. Донецк, 25–26 ноября 2020 г.

Личный вклад соискателя. Все результаты и положения, составляющие основное содержание диссертации, вынесенные на защиту, получены автором самостоятельно. Личный вклад соискателя заключается в обосновании идеи работы и ее реализации, цели и задач работы, в выборе методов и направлений исследований, выполнении теоретических, аналитических и экспериментальных исследований, разработке положений и рекомендаций по использованию результатов работы.

Публикации. По теме диссертационной работы всего было опубликовано 10 научных работ. Из них 6 работ в изданиях, рекомендованных ВАК ДНР, 4 по материалам научно-технических конференций.

Структура и объем диссертации. Диссертационная работа содержит 199 страниц машинописного текста и состоит из введения, четырех разделов, заключения, списка литературы из 141 источника и 4 приложений. Основной текст диссертации иллюстрируется 82 рисунками и содержит 47 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение содержит общую характеристику работы. Обоснована актуальность темы диссертации, сформулированы цели и задачи исследования. Показана научная новизна и практическая значимость полученных результатов.

В первом разделе работы «Анализ состояния вопроса, цель и задачи исследования» описывается необходимость совершенствования существующих и разработка новых подходов к сбору, хранению и обработке ИР.

В процессе исследования были изучены различные аспекты применения онтологий при проектировании различных систем управления знаниями, в том числе и научно-образовательных учреждений, обоснована целесообразность онтологического подхода. Выполнен анализ существующих моделей, методов и алгоритмов решения задач обработки информации: моделей представления знаний; моделей представления текста; видов информационного поиска и метрик оценки его качества; мер оценки семантической близости.

На основе проведенного анализа поставлена цель диссертационной работы, сформулированы основные задачи и выбраны основные направления для достижения поставленной цели.

Во втором разделе работы «Онтологический подход в системах организационного сопровождения деятельности научно-образовательных учреждений» разработан гибридный подход к формированию онтологии, заключающаяся в том, что на различных этапах создания онтологии были использованы различные способы ее создания. На начальном этапе создания были изучены и адаптированы следующие онтологии верхнего уровня: Dublin Core, FOAF, VIVO, BIBO, DBpedia ontology, TEACH, VCard.

При формировании «базовой» онтологии использовался экспертный способ. Для более наглядного представления онтологии применялся модульный принцип построения. «Базовая» онтологическая модель системы включает следующие модели: обобщенная «Онтомодель», «Научные мероприятия», «Академические звания и должности», «Персоналии», «Научно-образовательное учреждение», «Информационный ресурс», «Документ», «Образовательный ресурс».

На рисунке 1 приведена обобщенная «Онтомодель», разработанная в Protege:

$$Ontomodel = <Addr, Cond, Org, Eve, Profile, Themes> \quad (1)$$

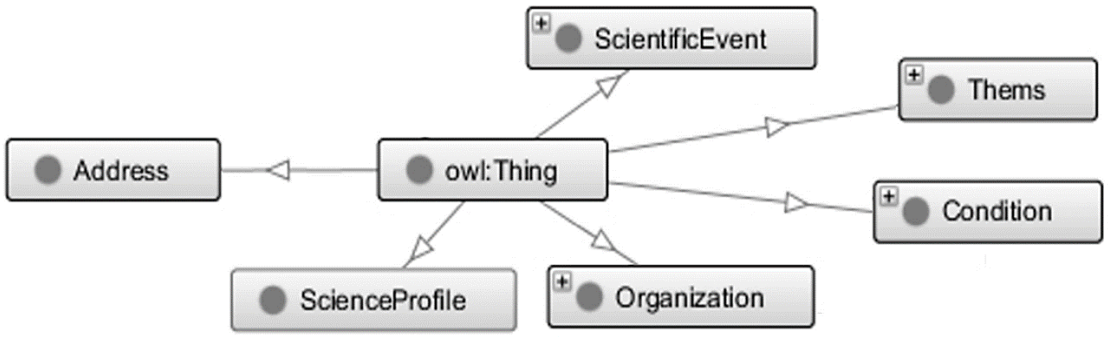

Рисунок 1 – Обобщенная «Онтомодель»

В таблице 1 описаны основные классы обобщенной модели «Онтомодель».

Таблица 1 – Классы модели «Онтомодель»

Параметр	Имя класса	Класс в онтологии	Описание класса
Addr	Address	vCard:Address	Адрес
Cond	Condition		Состояние/Статус
Org	Organization	foaf:Organization	Учреждение
Eve	ScientificEvent		Событие
Profile	ScienceProfile		Наукометрический профиль
Thems	Thems	Vivo:ReseachArea	Тематика ИР

При формировании онтологий предметных областей реализованы: экспертный режим; способ полуавтоматической обработки документов; словарные способы пополнения; импорт из универсальной онтологии DBpedia.

При извлечении из документов знаний модулем полуавтоматической обработки текста использовались регулярные выражения и библиотека

морфологического анализа PHPMorphy. В результате обработки документа был сформирован предварительный список ранжированных по частоте слов-кандидатов в термины, который может быть изменен экспертом.

Реализован способ пополнения онтологии, базирующийся на использовании словарных концептов из электронных словарей, тезаурусов (WordNet, RussNet, Википедиа, Wiktionary). Как показали эксперименты (Таблица 2), словарь Wiktionary, для которого показатель качества поиска, выраженного F-мерой, равной 0.87, обладает наибольшей терминологической полнотой для тестируемой предметной области и в среднем возвращает большее количество релевантных результатов. (Рисунок 2).

Кривая «полнота-точность» для словаря Wiktionary с уточнением предметной области поиска приведена на рисунке 2.

Таблица 2 – Расчеты для словаря Wiktionary с уточнением предметной области поиска

№	TP	TN	FP	FN	Recall (полнота)	Precision (точность)	F-мера
1	9	1	5	0	1	0,64	0,78
2	7	2	1	0	1	0,87	0,93
3	8	3	0	2	0,8	1	0,89
4	10	1	0	0	1	1	1,00
5	7	1	0	3	0,96	1	0,98
-	-	-	-	-	-	-	-
20	6	4	0	2	0,75	1,00	0,86
В среднем по коллекции					0,82	0,92	0,87

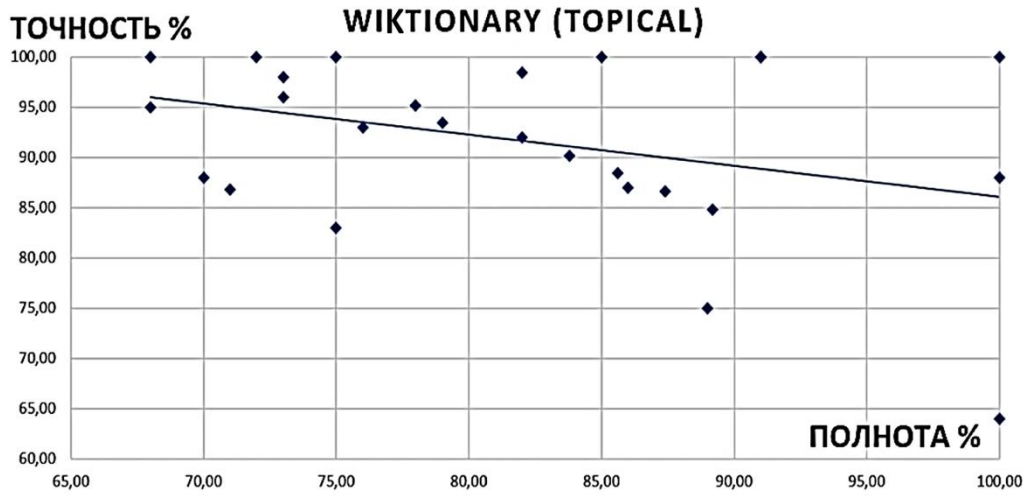

Рисунок 2 – Кривая «полнота-точность» для словаря Wiktionary с уточнением предметной области поиска

Разработан способ пополнения онтологии с использованием междоменной онтологии DBpedia. Автоматическая обработка текста проводилась двумя модулями: первый выполнял автоматический разбор текста и его аннотирование (Рисунок 3), а второй извлекал информацию о подобных концептах по выбранному пользователем термину (Рисунок 4).

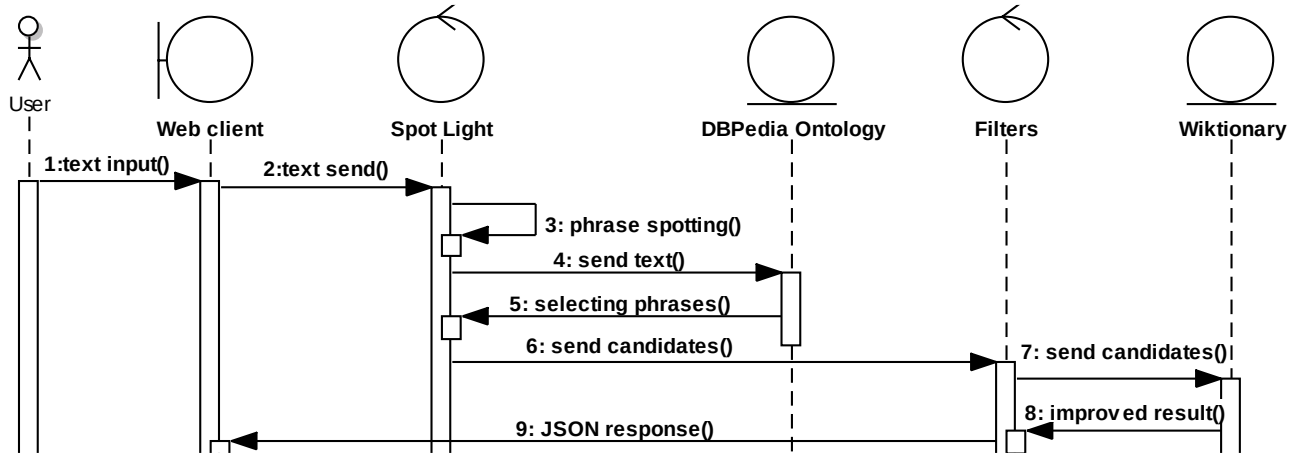

Рисунок 3 – Диаграмма последовательности процесса аннотирования документа

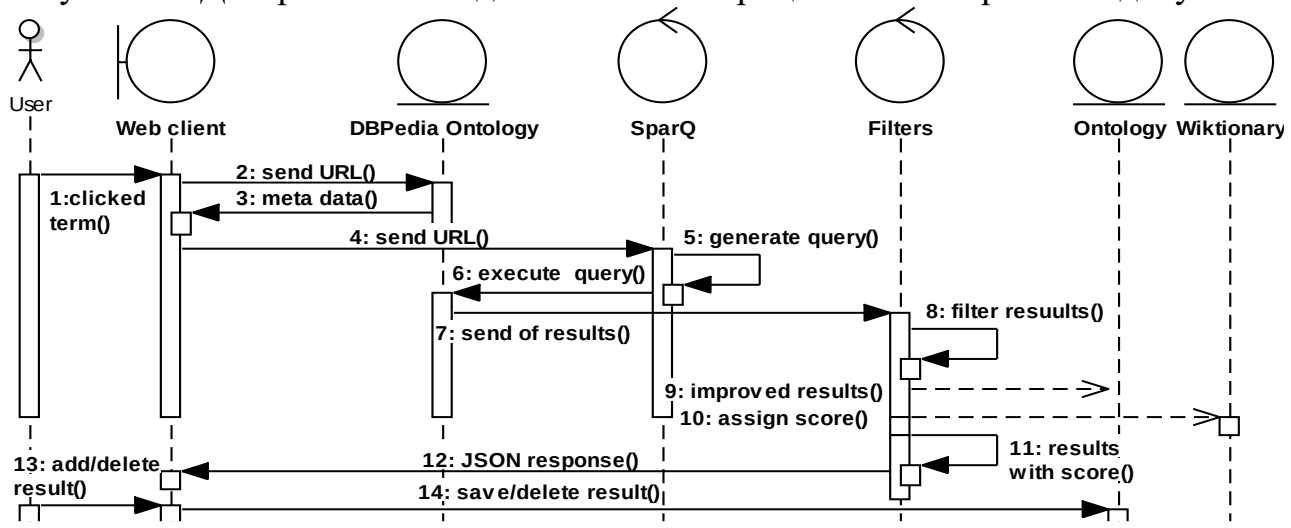

Рисунок 4 – Диаграмма последовательности процесса извлечения данных

В рамках тестового примера 266 понятий были вручную размечены экспертом. Программный модуль с точностью 0.87, полнотой 0.78 и F-мерой 0.82 автоматически аннотировал 241 элемент. Результаты тестирования показывают, что концепты могут быть извлечены из документов с использованием общедоступной онтологии DBpedia.

В третьем разделе работы «Разработка моделей, методов и алгоритмов выявления, приобретения и классификации знаний» приведены основные модели и алгоритмы для работы с ИР.

Для улучшения семантики предложено использовать графовую модель представления знаний (RDF) в виде трехмерного тензора семантических связей, значения которого определять в диапазоне [0; 1], а не только значениями «0» или «1» (Рисунок 5).

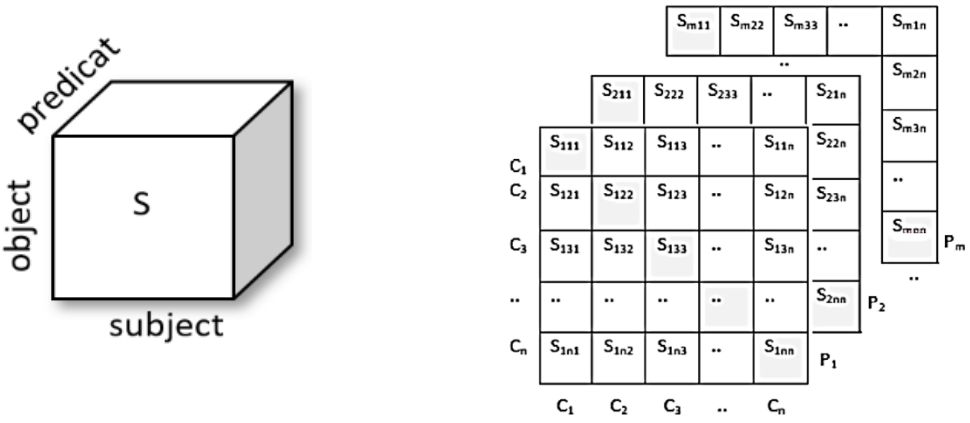

Рисунок 5 – Трехмерный тензор семантических связей RDF-графа знаний

Представление RDF-графа знаний в виде тензора семантических связей позволяет эффективным образом вычислить оценки семантической близости между двумя концептами i и j через факторизацию срезов тензора. Выполнив свертку тензора $S^{ij}{}_k$ с векторами коэффициентов значимости для каждого типа отношения W^k получаем:

$$R^{ij} = \sum_{k=1}^{p} W^k S_k^{ij} \qquad (2)$$

где W^k – вес, который определяет относительную важность каждого типа отношения; p – число отношений; R – матрица семантических связей.

Используемые при расчете гибридной меры отношения, виды оценок и формулы расчета СБ приведены в таблице 3.

Для нахождения весовых коэффициентов W^k разработан генетический алгоритм. Получено некоторое субоптимальное решение, для которого значение ЦФ≈0.27 и точность вычисления целевой функции равна 0.00028. Весовые коэффициенты модели W^1=0.13; W^2=0.18; W^3=0.18; W^4=0.04; W^5=0.1; W^6=0.12; W^7=0.05; W^8=0.2 установлены в качестве параметров модели в настройках системы.

Таблица 3 – Отношения, виды оценок и формулы расчета СБ

Отношения	Виды оценок	Формулы
P1 – таксономические P2 – партономические P3 – родовидовые	по иерархии онтологии	Длина кратчайшего пути
P4 – синонимы, P5 – экземпляры	по свойствам концептов	Длина кратчайшего пути
P6 – ассоциации	по горизонтальным отношениям	Заполняется экспертом [0;1]
P7 – семантические	по общим атрибутам	Атрибутивная мера
P8 – семантические	на векторном представлении	Косинусная мера Мягкая косинусная мера

Процесс информационного поиска включает ряд операций, направленных на сбор, обработку и предоставление пользователю знаний (Рисунок 6).

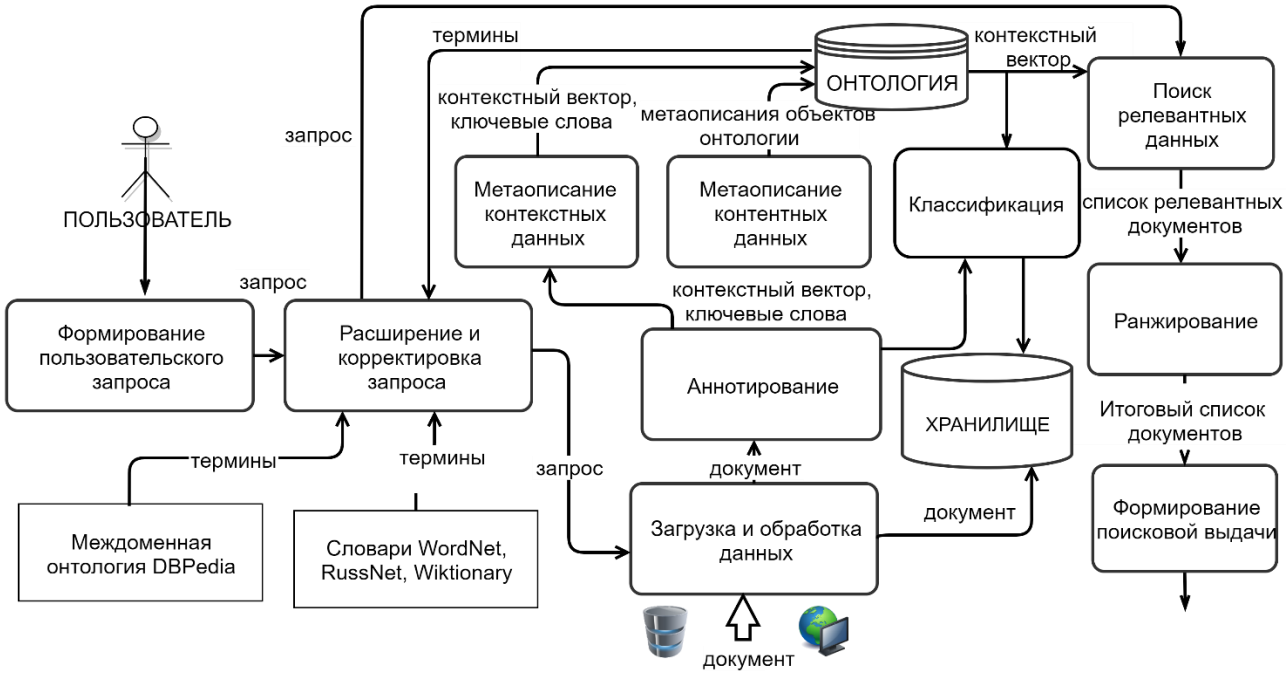

Рисунок 6 – Процесс информационного поиска

Разработка математической модели информационного поиска состоит из разработки нескольких моделей: представления текстов; формирования поисковых запросов; извлечения релевантных данных по запросу.

Представлением документа и поисковым образом ИР является контекстный вектор, элементы которого формируются на этапе автоматической обработки текста и хранятся в онтологии одновременно со списком ключевых слов.

Укрупненный алгоритм формирования векторного представления текста в виде контекстного вектора V(k) следующий:

1. Из модуля автоматического разбора текста передаются все необходимые данные.
2. Формируется вектор из ключевых слов, полученных с помощью частотного анализа.
3. Вектор дополняется термами из онтологии.
4. Поисковый запрос расширяется терминами со сторонних ресурсов.
5. Полученный вектор в случае необходимости корректируется экспертом или автором ресурса.
6. Для каждого термина вычисляется TF.
7. Элементы вектора сортируются в зависимости от значения TF.
8. Вектор ограничивается по количеству элементов k.
9. Полученный контекстный вектор сохраняется в онтологии.
10. Онтология наполняется новыми терминами из массива ключевых слов.

Каждый документ соотносится при помещении в онтологию с определенной темой, что позволяет сформировать из контекстных векторов

документов контекстные векторы тематических пространств и использовать их при поиске и классификации документов.

Модель Bag-of-concepts была модифицирована и обобщена для использования в пространствах третьего порядка. В качестве оператора перехода из одного пространства дескрипторов в другое использовался 3-х мерный тензор семантических связей S_p^{nn}. Выполнив аддитивную свертку тензора T_p^{nn} по формуле (2), получили значение обобщенной матрицы семантических связей R^{nn}.

Пусть текст представлен контекстным множеством в виде тензора первого ранга V_n. Полученная матрица семантических связей R^{nn} позволяет отобразить V_n – исходное представление текста в новое представление тензора первого ранга U^n, уже отражающее семантические связи между словами:

$$U^n = V_n \cdot R^{nn} \quad (3)$$

Чтобы справиться с большим количеством RDF-троек, с разреженностью данных, а также для устранения проблемы большой вычислительной сложности, была использована техника редукции признакового пространства, т.е. переход к пространствам более низких порядков.

Стандартные подходы снижения размерности исходного признакового пространства могут быть разделены на два класса: с трансформацией признакового пространства (PCA, SVD, NMF); без трансформации исходного пространства с исключением неинформативных признаков.

Были разработаны две редуцированные модели представления текстовых информационных ресурсов. Первая модель без трансформации признакового пространства получила название ONTO, вторая с трансформацией признакового пространства получила название TEN.

Для модели ONTO семантический образ каждого тематического раздела формируется на базе образов отдельных концептов, принадлежащих данному разделу.

Пусть пользователь сформулировал некоторый запрос Q_k. Тензор 2-ранга Z_t^k задает соответствие между концептами, являющимися ключевыми словами и темами. Для определения подходящей тематики сначала выполнили аддитивную свертку тензора Z_t^k с тензором первого ранга поискового запроса Q_k:

$$Y_t = Z_t^k \cdot Q_k \quad (4)$$

Редукция была выполнена по тематике с максимальным значением Y_t. Сформированный путем отбора на базе исходного тензора тематически редуцированный тензор 3-го ранга G_p^{kk} содержал только семантически близкие термины.

Для модели TEN редукцию выполнили с помощью подхода, основанного на тензорной факторизации. Графическая интерпретация одного из тензорных разложений Tucker2 приведена на рисунке 7.

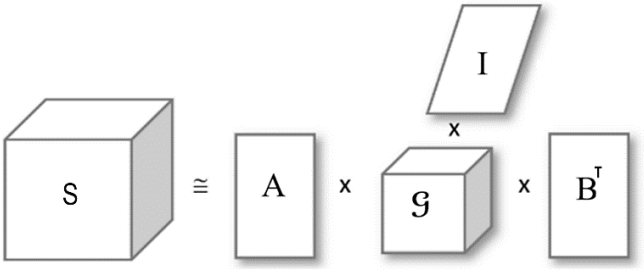

Рисунок 7 – Тензорное разложение Tucker2

Факторизация данных в более низкое размерное пространство вводит компактный базис, который может описать исходные данные в сжатой форме.

Воспользовавшись вариацией разложения Tucker2 RESCAL, для которого одна из матриц разложения единичная (I), а две другие матрицы (A, B) равны между собой, получаем:

$$S \cong A \times_1 G \times_2 A^T = [G; A, A^T, I] \quad (5)$$

В индексном виде, исходя из параметров тензоров S и G:

$$S_{oop} \cong \sum_{k=1}^{K} \sum_{k=1}^{K} \sum_{p=1}^{P} A_{ok} G_{kkp} A_{ok}^{T}, \quad (6)$$

где: S(OxOxP) – исходный тензор 3-го ранга; O – количество терминов; P – количество отношений; G(KxKxP) – ядро, тензор 3-го ранга, A – тензор 2-го ранга; K – количество ключевых терминов, K <<O.

Поскольку K намного меньше по значению, чем O, то тензор G можно рассматривать как сжатую версию S, что существенно уменьшает размерность пространства.

Перед реализацией программных модулей был выполнен ряд экспериментов на прототипах для тестирования разработанных алгоритмов и моделей. В процессе экспериментов изменялись такие параметры, как количество концептов модели, количество семантических мер модели, размер контекстного вектора с целью оценить, насколько эти параметры влияют на скорость выполнения базовых операций: ввода и обработки данных.

Экспериментальные исследования показали целесообразность использования модели без трансформации признакового пространства ONTO, которая использует редукцию по тематическим разделам онтологии и снижает размерность задачи до размера контекстного вектора тематического раздела, что позволило для количества концептов онтологии 2250 объектов и размера контекстного вектора 30 элементов уменьшить вычислительную сложность более чем на порядок практически без ослабления семантики (на 6.2%).

В ходе изучения проблем построения поисковых систем было обнаружено, что поисковые запросы возвращают недостаточно релевантные результаты и поисковый запрос требует расширения и семантического обогащения. Для устранения избыточности поисковых запросов необходимо ранжирование результатов поиска.

Модель формирования семантически обогащенного поискового запроса:

1. Поисковый запрос расширяется терминами сторонних ресурсов (WordNet, Wiktionary), параллельно наполняя онтологию новыми концептами.
2. Поисковый запрос наполняется множеством понятий из онтологии.
3. Выполняется расчет весов терминов по мере TF и их ранжирование.
4. Формируется вектор запроса на базе контекстных векторов терминов.

Для создания эффективного алгоритма извлечения релевантных данных по запросу необходимо искать только те документы, которые лежат в той же области пространства понятий, что и запрос. Тематические разделы описаны в онтологии и имеют свои контекстные множества и вектора длиной k. Сначала определяется близость поискового запроса к тематическим разделам, затем сравниваются контекстные вектора запроса и документов внутри раздела.

Разработан алгоритм извлечения релевантных данных по запросу.

1. Определяем тематический раздел по формуле (4), получаем тематически редуцированный тензор.
2. Получаем значение обобщенной семантической матрицы R^{kk} по формуле (2).
3. Получаем семантически обогащенный поисковый образ документа:

$$C^k = Q_k R^{kk}, \quad (7)$$

где: R(KxK) – матрица семантических связей терминов; k – число ключевых терминов; Q – поисковый запрос.

4. Получаем СБ запроса к другим документам тематического раздела:

$$X^l = \mathrm{C}_k B^{lk}, \quad (8)$$

где: B(LxK) – тензор 2-го ранга, хранящий контекстные вектора терминов тематического раздела; K – число ключевых терминов; L – общее число терминов тематического раздела, при чем L <=O; если O – общее количество терминов онтологии; C – поисковый образ документа.

5. Определяем релевантные запросу документы. Чем больше значение СБ между поисковым образом документа и представлением документа, тем выше позиция документа в финальной поисковой выдаче.

В результате исследований получили, что качество извлечения релевантных данных по запросу для разработанной модели, выраженное F-мерой, в среднем составляет 0.83, что на 10.7% больше, чем для модели с использованием меры «косинусного сходства».

Для решения задачи классификации и соотнесения ИР к тематическому разделу разработан алгоритм классификации:

1. Определение тематического раздела для классификации производим по формуле (4) и в результате получаем тему, максимально подходящую документу.
2. Рассчитываем обобщенную матрицу R^{kk} семантической близости через факторизацию срезов тензора по формуле (2).

3. Получаем семантически обогащенное представление документа C^k по формуле (7).

4. Получаем СБ документа к другим документам X^l по формуле (8).

5. Находим документ с максимальным весом X^l, и соответственно, определяем место для нового документа.

6. Пополняем онтологию метаданными и помещаем новый документ в хранилище.

Тестирование алгоритма показало повышение качества модели классификации, выраженного F-мерой, на 45.4% (F=0.8) по сравнению с алгоритмом, использующем онтологические меры СБ (F=0.55), на 5.3% по сравнению с алгоритмом, использующим меру «косинусного сходства» (F=0.76) и на 9.5% по сравнению с алгоритмом по «мягкой косинусной» мере (F=0.73).

Четвёртый раздел работы «Разработка системы учета информационных ресурсов» описывает архитектурные модели фреймворка системы управления ИР, например модели компонентов (Рисунок 8) и размещения (Рисунок 9). На базе фреймворка реализован модуль «Наука», функциональная модель которого приведена на рисунке 10.

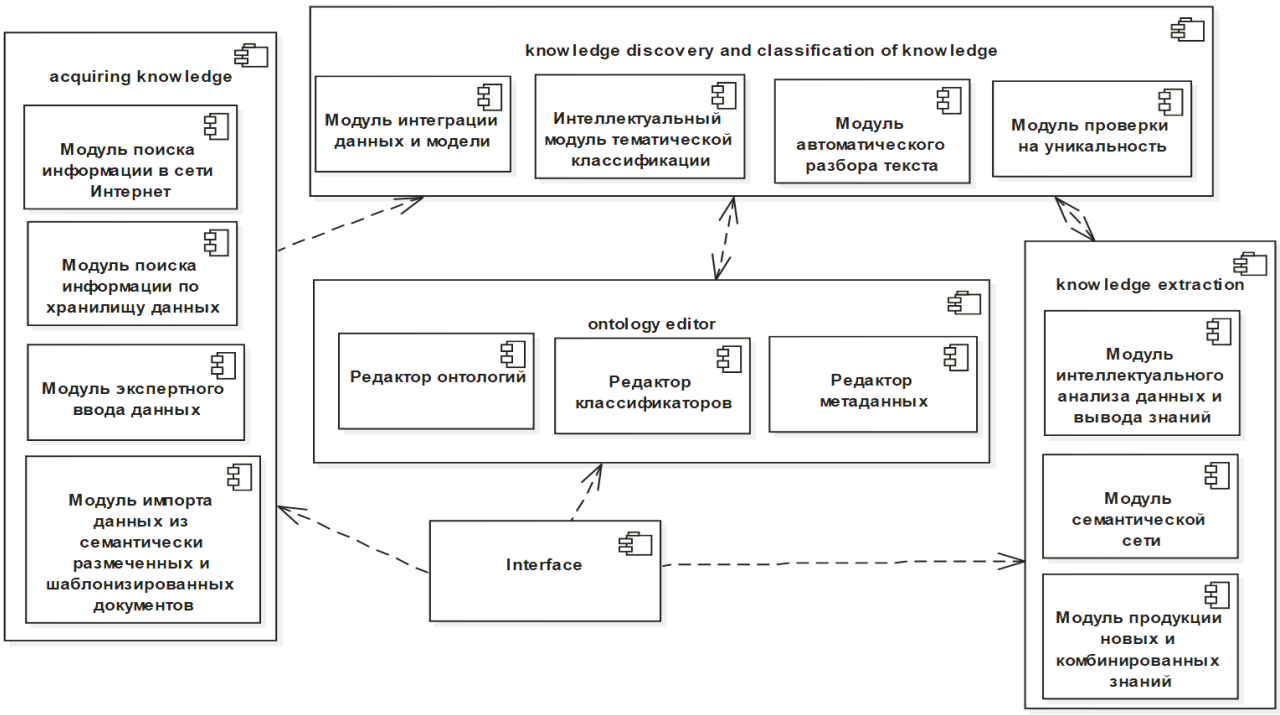

Рисунок 8 –Модель компонентов

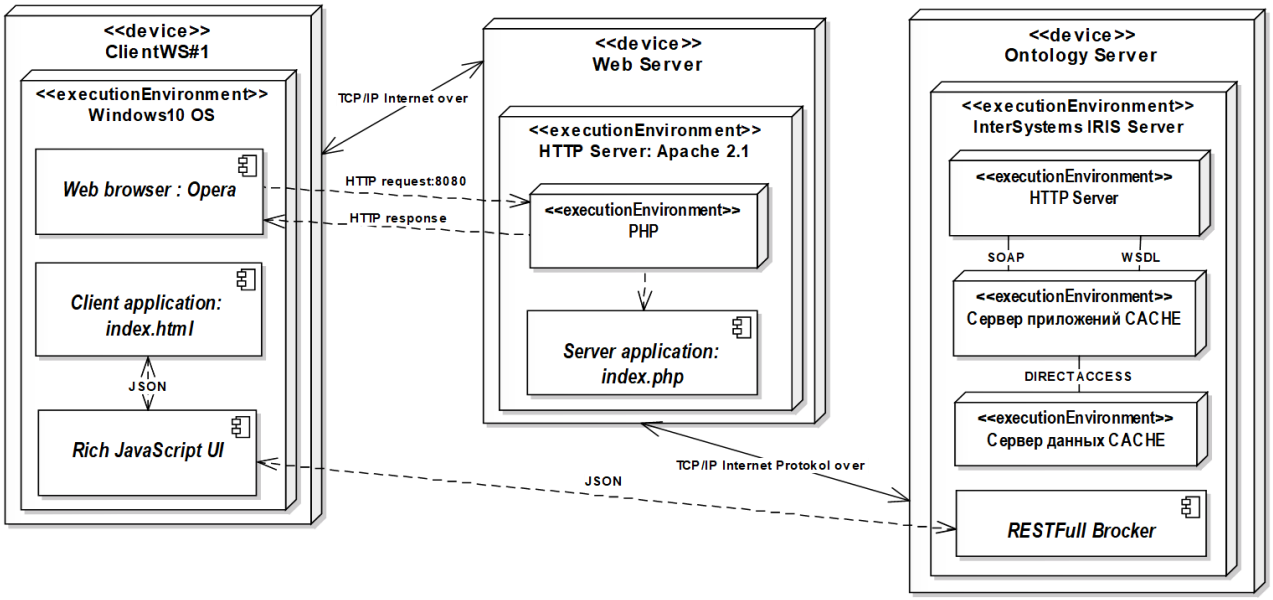

Рисунок 9 – Модель размещения

Рисунок 10 – Функциональная модель модуля «Наука»

Тестирование разработанных моделей и алгоритмов показало корректность полученных результатов и приемлемое время решения задач обработки информации.

ЗАКЛЮЧЕНИЕ

Диссертационная работа является законченной научно-исследовательской работой, в которой дано новое решение актуальной научно-технической задачи

повышения эффективности системы управления информационными ресурсами научно-образовательных учреждений за счет применения онтологического подхода, разработки новых и усовершенствования существующих моделей и алгоритмов поиска, хранения и классификации данных.

Основные результаты и выводы, полученные при выполнении работы состоят в следующем.

1. На основе анализа существующих разработок и исследований сделан вывод об актуальности выбранной тематики, о перспективности использования онтологий в качестве модели представления знаний и о целесообразности онтологического подхода к разработке системы управления ИР.
2. Разработана гибридная процедура формирования онтологии, когда на начальном этапе создания «базовой» онтологии использованы различные онтологии «верхнего уровня» и метод «экспертного создания». Для построения и пополнения нижних уровней онтологии предметных областей обоснован выбор словаря Wiktionary, для которого показатель качества поиска, выраженного F-мерой, достиг наибольшего среди тестируемых значения 0.87. Разработан способ пополнения онтологии с использованием DBpedia, при этом концепты могут быть извлечены из документов с точностью 87% и полнотой 78%.
3. Разработана концептуальная метамодель для учета связанности знаний и обеспечения однородности представления данных в рамках единой тематики проектируемой системы, ядром которой является онтология.
4. Модифицирована модель N-мерного представления знаний на базе RDF-графа в виде трехмерного тензора семантических связей, значения которого определяются различным образом для различных отношений RDF-графа знаний и содержат значения в диапазоне [0; 1], что позволило для отдельных видов отношений учитывать не только наличие связей, но и их силу.
5. Усовершенствована гибридная мера оценки семантической близости на базе модели N-мерного представления RDF-графа знаний, что дало возможность определять сходство с учетом семантики, частотных характеристик текста, контекста и структуры онтологии и улучшить качество поиска. Для определения весовых коэффициентов интеллектуальной гибридной меры оценки СБ использован генетический алгоритм.
6. Разработаны две редуцированные модели представления текстовых информационных ресурсов: без трансформации признакового пространства и с трансформацией признакового пространства. Экспериментальные исследования показали целесообразность использования модели без трансформации признакового пространства, которая использует редукцию по тематическим разделам онтологии и снижает размерность задачи до размера контекстного вектора тематического раздела, что позволило для количества концептов онтологии, равном 2250 объектов, и размеру контекстного вектора, равному 30 элементов повысить скорость выполнения базовых алгоритмов более чем в 40 раз практически без потерь качества поиска (на 6.2%).

7. Усовершенствован алгоритм поиска данных с использованием онтологии, гибридной меры оценки СБ и векторной модифицированной модели представления текстов, что привело к повышению качества модели информационного поиска, выраженного F-мерой, на 10.7% по сравнению с алгоритмом, использующим меру «косинусного сходства».

8. Усовершенствован алгоритм классификации данных с использованием гибридной меры оценки СБ, что привело к повышению качества модели классификации, выраженного F-мерой, на 45.4% по сравнению с алгоритмом, для которого СБ вычислялась только по онтологии, на 5.3% по сравнению с алгоритмом, использующим меру «косинусного сходства» и на 9.5% по сравнению с алгоритмом на базе «мягкой косинусной меры».

9. Разработана структурная архитектурная модель фреймворка системы управления информационными ресурсами научно-образовательных учреждений, реализующая предложенные модели и алгоритмы. Тестирование созданного на базе фреймворка программного приложения для учета научной деятельности сотрудников «Наука» показало корректность полученных результатов и приемлемое для пользователя время решения поисковых классификационных задач.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

- в рецензируемых научных изданиях, рекомендованных ВАК ДНР:

1. **Андриевская, Н.К.** Основные принципы и подходы при разработке системы управления профессиональными знаниями вуза / **Н.К. Андриевская** // Научный журнал «Информатика и кибернетика». – 2019. – №4(18). – С.49–56.

2. **Андриевская, Н.К.** Онтологический подход в системах обработки данных научных и научно-образовательных организаций / **Н.К. Андриевская** // Международный научно-теоретический журнал «Проблемы искусственного интеллекта». – 2020. – №1(16). – С.23–36.

3. **Андриевская, Н.К.** Разработка прикладной онтологии в системах обработки данных научных и научно-образовательных организаций / **Н.К. Андриевская** // Вестник ДонНУ. Сер.Г:Технические науки. – 2020. – №3. – С.43-51.

4. **Андриевская, Н.К.** Анализ возможностей использования существующих словарей для пополнения онтологии / **Н.К. Андриевская**, А.И. Секирин, С.В. Канатуш // Научный журнал «Информатика и кибернетика». – 2020. – №2(20). – С.13–20.

5. **Андриевская, Н.К.** Обобщенная модифицированная модель представления текстовых информационных ресурсов. / **Н.К. Андриевская** // Научный журнал «Информатика и кибернетика». – 2020. – №4(22). – С.21–30.

6. **Андриевская, Н.К.** Гибридный интеллектуальный способ оценки семантической близости / **Н.К. Андриевская** // Международный научно-

теоретический журнал «Проблемы искусственного интеллекта». – 2021. – №1(20). – С.4–17.

- в других изданиях:

7. Бажанова, А.И. Разработка морфологического анализатора для построения понятийного аппарата электронной библиотеки кафедры АСУ / А.И. Бажанова, Т.В. Мартыненко, **Н.К. Андриевская** // Информатика и компьютерные технологии / Сборник трудов VII международной научно-технической конференции студентов, аспирантов и молодых ученых – 22–23 ноября 2011 г., Донецк, ДонНТУ. – 2011. В 2-х томах. Т.1 – 417 с. – С.326–330.

8. Чайка, В.А. Обзор средств разработки онтологических моделей / В.А. Чайка, С.Ю. Землянская, **Н.К. Андриевская** // Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ–2020): XI Международная научно-техническая конференция, 27-28 мая 2020, г. Донецк: / Донец. национал. техн. ун-т; редкол. Ю.К. Орлов и др. – Донецк: ДОННТУ, 2020. – С.233–237.

9. Канатуш, С.В. Онтологический подход к веб-поиску / С.В. Канатуш, **Н.К. Андриевская** // Современные проблемы радиоэлектроники и телекоммуникаций: сб. науч. тр. / под ред. Ю.Б. Гимпилевича. – Москва-Севастополь: Изд-во: РНТОРЭС им. А.С. Попова, СевГУ, 2020. – №3. – 247с.– С.205.

10. **Андриевская, Н.К.** Разработка архитектурной модели системы управления информационными ресурсами организаций / **Н.К. Андриевская**, А.И. Секирин, О.В. Ченгарь // В сборнике: Программная инженерия: методы и технологии разработки информационно-вычислительных систем (ПИИВС-2020). Сборник научных трудов III Международной научно-практической конференции. – Донецк, 2020. – С.46–54.

Личный вклад соискателя в публикациях: [4] – предложен подход для формирования нижних уровней онтологии; [7] – разработана общая постановка проблемы и подход к построению онтологической модели; [8] – поставлена проблема извлечения и использования полезной информации из документов; [9] – предложено использование онтологического подхода к формированию поисковых запросов; [10] – разработаны основные архитектурные модели.

АННОТАЦИЯ

Андриевская Н.К. Совершенствование моделей и алгоритмов обработки информации в системах организационного сопровождения деятельности научно-образовательных учреждений. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.01 – Системный анализ, управление и обработка информации (по отраслям) (технические науки). – ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ», Донецк, 2021.

В диссертационной работе дано теоретическое обоснование и приведено решение актуальной научно-практической задачи повышения эффективности системы управления информационными ресурсами научно-образовательных учреждений за счет применения онтологического подхода, разработки новых и усовершенствования существующих моделей и алгоритмов поиска, хранения и классификации данных.

Для решения поставленной задачи предложен базирующихся на онтологии подход к построению системы управления информационными ресурсами учреждений, обеспечивающий представление и интерпретацию информации в виде знаний. Создан ряд моделей и алгоритмов на базе онтологии для решения задач поиска, классификации и обработки информационных ресурсов.

Разработан программный модуль для учета научной-исследовательской деятельности преподавателей. Тестирование показало корректность полученных результатов.

Ключевые слова: онтология, поиск, классификация, семантическая близость, векторная модель представления текста, словарь, RDF, OWL.

ABSTRACT

Andrievskaya N. Improvement of information processing models and algorithms for in the systems of institutions organizational support of scientific and educational activities. As the manuscript.

Ph.D. (Candidate's) Thesis in Engineering Science by specialty 05.13.01 - System analysis, management and information processing (by industry) (technical sciences) – STATE HIGHER EDUCATIONAL ESTABLISHMENT "DONETSK NATIONAL TECHNICAL UNIVERSITY", Donetsk, 2021.

The thesis provides a solution to the actual scientific and practical problem of improving the efficiency of the information resources management system of scientific and educational institutions through the use of an ontological approach, the development of new and improvement of existing models and algorithms for data search, storage and classification.

To solve this problem, an ontology-based approach to building an information resource management system of institutions is proposed.

A number of models and algorithms based on ontology have been created.

A software module was developed to take into account the scientific and research activities of teachers. The testing showed the correctness of the results obtained of the developed algorithms.

Keywords: ontology, retrieval, classification, semantic proximity, vector model of text representation, dictionary, RDF, OWL.