На правах рукописи

Хорошилов Алексей Александрович

МЕТОДЫ, МОДЕЛИ, АЛГОРИТМЫ И ЭКСПЕРИМЕНТАЛЬНОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ АВТОМАТИЧЕСКОГО ВЫЯВЛЕНИЯ НЕЯВНО ВЫРАЖЕННЫХ ЗАИМСТВОВАНИЙ В НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТАХ

Специальность 05.13.17 – «Теоретические основы информатики»

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва - 2015

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН)

Научный руководитель	**Захаров Виктор Николаевич**, доктор технических наук, доцент
Официальные оппоненты:	**Максимов Николай Вениаминович**, доктор технических наук, профессор, Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Национальный исследовательский ядерный университет «МИФИ», профессор кафедры системного анализа
	Лукашевич Наталья Валентиновна, кандидат физико-математических наук, ведущий научный сотрудник Научно-исследовательского вычислительного центра Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова» (НИВЦ МГУ)
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Библиотека по естественным наукам Российской академии наук (БЕН РАН), г. Москва

Защита состоится «9» декабря 2015 г. в 15 часов 00 минут на заседании диссертационного совета Д 002.073.01 при Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д.44, корп.2 (конференц-зал).

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН и на официальном сайте ФИЦ ИУ РАН: http://www.ipiran.ru/announce/ .

Автореферат разослан « » 2015 г.

Ученый секретарь
диссертационного совета Д 002 073.01
доктор технических наук, профессор С.Н. Гринченко

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Стремительное развитие сети Интернет и информационных технологий многократно облегчили возможности доступа к разнообразным информационным ресурсам. Но, наряду с огромным позитивным влиянием этого явления на развитие современного общества, возникли серьезные проблемы, связанные с недобросовестным использованием информации. В частности, это незаконное присвоение авторства на чужое произведение или некорректное заимствование его части (так называемая проблема плагиата). Это явление особенно широко распространено среди студенчества. Не секрет, что среди студентов распространен метод написания работ, который получил название «copy paste» - копирование материалов из интернета с минимальным их редактированием. Такое использование информационных ресурсов можно расценить как «неприкрытое копирование» или плагиат[1]. Это же явление, судя по материалам СМИ, также серьезно затронуло и научно-педагогическую деятельность, связанную с подготовкой различного рода квалификационных работ, включая кандидатские и докторские диссертации.

В последние годы с этим негативным явлением ведется планомерная борьба. Так, например, для выявления заимствований в квалификационных работах используются различные средства автоматизации. Но имеющиеся на рынке IT-услуг системы поиска заимствований в документах способны достоверно выявлять, в основном, факты прямого заимствования. Это связано со сложностью анализа содержания текстов, обусловленной прежде всего тем, что в них одни и те же ситуации могут описываться в терминах различной степени общности и с помощью различных языковых средств.

Поэтому в настоящее время только человек-эксперт, анализирующий документы на предмет установления фактов плагиата, на основе результатов их автоматического анализа, способен, руководствуясь своими представлениями о содержании документов и средствах выражения этого содержания, а также опираясь на свои профессиональные знания и опыт, установить наличие или отсутствие такого факта. Но когда факты плагиата скрыты путем значительной переработки заимствованного текста, их невозможно выявить имеющимися в настоящее время средствами автоматизации.

Актуальность темы исследования определяется потребностью в получении информации обо всех возможных фактах незаконных заимствований в анализируемых документах, необходимой, в частности, для обеспечения более объективной оценки квалификационных работ различного уровня. А это, в свою очередь, поможет повысить уровень подготовки научных и

---

[1] Кичерова М.Н., Кыров Д.Н и др. Плагиат в студенческих работах: анализ сущности проблем //Интернет-журнал Науковедение, 2013, № 4.

профессиональных кадров и, в конечном итоге, улучшить качество высшего и среднего образования.

Используемые в этом исследовании подходы базируются на современных представлениях о смысловой структуре текстов и методах семантико-синтаксического и концептуального анализа смыслового содержания текстов.

Наибольший теоретический вклад в решение проблем семантического анализа текстов на естественном языке внесли такие ученые как Апресян Ю.Д., Белоногов Г.Г., Быстров И.И., Гиляревский Р.С., Добров Б.В., Звягинцев В.А., Лахути Д.Г., Лукашевич Н.В., Максимов Н.В., Мельчук И.А., Калинин Ю.П., Козеренко Е.Б., Кузнецов И.П., Осипов Г.С., Пиотровский Р.Г., Попов И.И., Поспелов Г.С., Рудаков К.В., Хорошевский В.Ф., Шемакин Ю.И., Шрайберг Я. Л., Broder A., Hartrumpf S., Salton G., Mooney R. J. и многие другие отечественные и зарубежные ученые.

Целью исследования является решение проблемы выявления неявно выраженных заимствований в текстах документов. В соответствии с указанной целью в работе поставлены следующие задачи:

1. Исследовать и разработать модели представления смыслового содержания текстов документов.
2. Исследовать и разработать методы и алгоритмы выявления наименований понятий в текстах документов и унификации их смыслового содержания.
3. Исследовать и разработать методы, модели и алгоритмы автоматического выявления заимствований в текстах документов, включая случаи неявно выраженных заимствований.
4. Разработать программное обеспечение для решения задачи автоматического выявления заимствований в текстах документов (включая неявно выраженные заимствования).
5. Провести экспериментальное исследование, устанавливающее достоверность теоретических концепций и эффективность разработанных методов выявления заимствований в текстах документов.

Объект исследования: понятийный состав и семантико-синтаксическая структура научно-технических текстов.

Предмет исследования: модели, методы и алгоритмы автоматической обработки, формализации и сопоставления смыслового представления содержания текстов.

Научная новизна. К основным результатам работы, отличающимся научной новизной относятся:

1. Методы, алгоритмы и экспериментальное программное обеспечение процесса формализации смыслового представления содержания документов.
2. Модели, методы, алгоритмы и экспериментальное программное обеспечение процесса автоматического выявления заимствований в текстах документов, включая случаи неявно выраженных заимствований.

3. Комплексное решение задачи автоматического выявления заимствований в текстах документов (включая неявно выраженные заимствования) на основе анализа их смысловой структуры.

Методы исследования базируются на использовании аппарата математической статистики, теории вероятностей, моделей представления знаний, моделей семантико-синтаксического и концептуального анализа текстов, методов формализации и кластеризации текстов.

Теоретическая ценность диссертации заключается в разработке решений, направленных на развитие моделей представления смыслового содержания текстов и построения на их основе моделей установления смысловой идентичности научно-технических текстов или их фрагментов.

Практическая ценность работы заключается в том, что научные и практические результаты диссертационных исследований были использованы в Федеральном государственном автономном научном учреждении «Центр информационных технологий и систем органов исполнительной власти» (ФГАНУ ЦИТиС) в рамках государственного задания на НИР в 2012-2014 гг. по теме «Исследование и разработка методов семантической экспертизы структуры и содержания научно-технических документов, а также наличия регламентированных для данного типа документов разделов и выявления несанкционированных заимствований (включая неявные заимствования)» при создании макета системы в подсистеме выявления заимствований в текстах.

Практические результаты также были использованы в рамках создания промышленной системы «Мониторинг СМИ» для Ситуационно-кризисного центра Госкорпорации Росатом (ФГУП «СКЦ Росатома»), реализующей функции сбора, консолидации, оперативной обработки поступающих документов для решения задачи формализации смыслового содержания и установления смысловой близости документов.

В настоящее время система «Мониторинг СМИ» функционирует в режиме промышленной эксплуатации. В ее базе данных накоплено более 37 млн. документов. Ежедневно в систему поступает и оперативно обрабатывается более 100 тыс. документов и новостных сообщений по различным тематикам.

На защиту выносятся следующие результаты:

1. Модель процесса выявления заимствований в документах (включая неявно выраженные) на основе анализа их смысловой структуры.
2. Метод установления смысловой близости и смысловой схожести фрагментов текста на основе анализа их смысловой структуры.
3. Алгоритм выявления наименований понятий в научно-технических текстах.
4. Алгоритм автоматического установления смысловых отношений между наименованиями понятий.
5. Алгоритм выявления заимствований в документах (включая неявно выраженные).

6. Экспериментальный программный комплекс выявления заимствований в научно-технических текстах (включая неявно выраженные).

7. Результаты исследования по автоматическому выявлению заимствований, подтверждающие достоверность и эффективность предложенных методов.

Достоверность выводов и рекомендаций обусловлена корректностью применения методов математической статистики, методов обработки текстов, воспроизводимостью и проверяемостью теоретических и экспериментальных результатов, согласованностью с практикой, внутренней непротиворечивостью, практической реализацией полученных результатов.

Личный вклад соискателя. Все изложенные в диссертации результаты исследования получены соискателем лично с учетом замечаний и рекомендаций научного руководителя.

Апробация результатов диссертационного исследования. Материалы диссертации излагались и обсуждались на следующих научно-технических конференциях: “Инновации в авиации и космонавтике – 2011” (Москва, 2011 г.), "Современные технологии в задачах управления, автоматики и обработки информации" (Алушта, 2011 г.), НТТМ-2011 (Москва, 2011 г.), КИИ-2012 (г. Белгород, 2012), RCDL’2012 (Переславль-Залесский, 2012), RCDL’2013 (Ярославль, 2013), Proceedings of ICAI’14, WORLDCOMP’14 (Las Vegas, Nevada), RCDL’2014 (Дубна, 2014), DAMDID/RCDL’2015 (Обнинск, 2015).

Публикации. Материалы диссертации содержатся в отчетах ФГАНУ ЦИТиС по государственному заданию на 2012-2015 г, в тематических выпусках журнала «Системы и средства информатики» (Т. 23, № 1, 2013), «Информатизация и связь» (№8, 2012; №3, 2013), «Научно-техническая информация» (№7, 2011). В открытой печати по теме диссертации опубликовано 14 работ, из них 5 работ в изданиях, входящих в Перечень ВАК Минобрнауки РФ. Получено 6 свидетельств об официальной регистрации программ для ЭВМ в Роспатенте.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения и приложений.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертации, определены цель и предмет исследования, новизна, практическая значимость, обоснованность и достоверность научных результатов диссертационного исследования, выводов и рекомендаций.

Первая глава посвящена исследованию понятия плагиат и анализу существующих программных средств его выявления в текстах документов.

В первом разделе в результате исследования действующего законодательства было установлено, что к настоящему времени понятие плагиата в нашей стране не получило единого и однозначного юридического определения. С юридической точки зрения плагиат представляет собой одну из

форм противоправного поведения и, в зависимости от степени общественной опасности, влечет различные виды юридической ответственности. Тем не менее, научная работа, в которой заимствованные положения были перефразированы и не являлись прямым отражением оригинального текста, а авторство источников таких некорректных заимствований принадлежит не одному, а нескольким лицам, в соответствии с законодательством не может быть квалифицирована как нарушение авторского права, хотя и является, по сути, компиляцией и плагиатом. Между тем, в зарубежной академической практике западных университетов и научных журналов плагиатом, как правило, считается любое использование чужих идей и высказываний без должной отсылки к источнику. Поэтому сейчас решение вопроса о наличии плагиата в диссертациях отнесено к компетенции экспертных советов, которым необходимо выявить в оспариваемых научных работах некорректные заимствования, определить их источники и сопоставить эти тексты, сравнивая полноту и смысловое содержание спорных фрагментов. Такая работа, в отсутствии четких установленных нормативно-правовыми актами критериев, носит сугубо субъективный характер и, особенно в отсутствии развитых технических средств поиска и сопоставления информации, требует больших трудозатрат и привлечения экспертов наивысшей квалификации.

Во втором разделе на основе проведенного анализа существующих методов автоматического выявления заимствований в текстах документов было установлено, что на современном уровне развития программных средств выявления заимствований не существует надежных методов достоверного выявления всех случаев заимствований, включая сложные случаи, такие как выявление неявно выраженных заимствований. К перспективным методам можно отнести семантические методы анализа текстов, базирующиеся на анализе семантической структуры текстов с помощью инструментов установления смысловой связи слов и словосочетаний (онтологии, тезаурусы, семантические словари). Но, тем не менее, и эти методы сейчас несовершенны и их несовершенство обусловлено использованием недостаточно адекватного семантического инструментария.

В третьем разделе ставится задача создания средств выявления заимствований в текстах документов (включая неявно выраженные). Эту задачу, как показывает анализ существующих методов поиска заимствований, можно решить только путем сопоставления смысловой структуры текстов и разработки методов, моделей и алгоритмов, определяющих и детализирующих этот процесс.

В **главе 2** в первом разделе рассмотрены технологии и процедуры автоматической обработки текстовой информации, основным назначением которых является решение таких задач как структурирование и формализация смыслового содержания текстов, выявление их понятийного состава, установление парадигматических, синтагматических и ассоциативных связей между наименованиями понятий и приведение их к унифицированному

формализованному представлению. При решении задачи настоящего исследования возникала необходимость использовать эти средства для формализации смыслового содержания документов, в качестве которых автор настоящего исследования использовал процедурные и декларативные средства программно-лингвистической платформы МетаФраз, разработанные при его непосредственном участии в рамках научной школы проф. Г.Г. Белоногова. Это программное обеспечение создано на основе теоретической концепции фразеологического концептуального анализа текстов, предложенной профессором Г.Г. Белоноговым и базирующейся на широкомасштабных исследованиях и адекватных представлениях о смысловой структуре текстов. Основной идеей этой концепции является обоснование использования в качестве основных единиц смысла устойчивых фразеологических и терминологических словосочетаний, обозначающих понятия, отношения между понятиями и типовые ситуации, представленные в предметной области. Эта концепция определяет принципы и методы выявления статистически обоснованного понятийного состава предметной области.

Исследование и разработка средств автоматического выявления заимствований базируются на процедуре установления смысловых отношений между наименованиями понятий и процедуре концептуального анализа текстов, выделяющей эти понятия в текстах. Автор в процессе работы над диссертационным исследованием разработал алгоритмы таких процедур. Ниже приводится общая схема алгоритма установления смысловых отношений между наименованиями понятий по словарю унифицированных формализованных представлений наименований понятий (по словарю УФПНП).

Алгоритм 1 - установления смысловых отношений между наименованиями понятий по словарю УФПНП.

Шаг 1. Выполнить морфологический анализ первого и второго наименований понятий (НП).

Шаг 2. Выполнить поочередно автоматическую генерацию смысловых вариантов первого и второго наименований понятий путем последовательной замены каждого слова наименования понятия на его синонимы.

Шаг 3. Произвести поиск всех сгенерированных вариантов обоих наименований понятий в словаре УФПНП. В случае нахождения какого-либо словосочетания в этом словаре, присвоить ему номер элемента словаря УФПНП.

Шаг 4. Если один из номеров первого понятия совпадет с одним из номеров второго понятия, то считается, что смысловая связь между анализируемыми наименованиями понятий установлена. В случае отсутствия такого совпадения считается, что связь между наименованиями понятий не установлена.

На рисунке 1 приводится общая схема алгоритма выявления наименований понятий в научно-технических текстах. Этот алгоритм

разработан в двух вариантах: точный анализ на основе использования эталонного концептуального словаря (ЭКС) и приближенный - на основе использования словаря обобщенных синтагм (ОС) и словаря малоинформативных слов и словосочетаний (МСС).

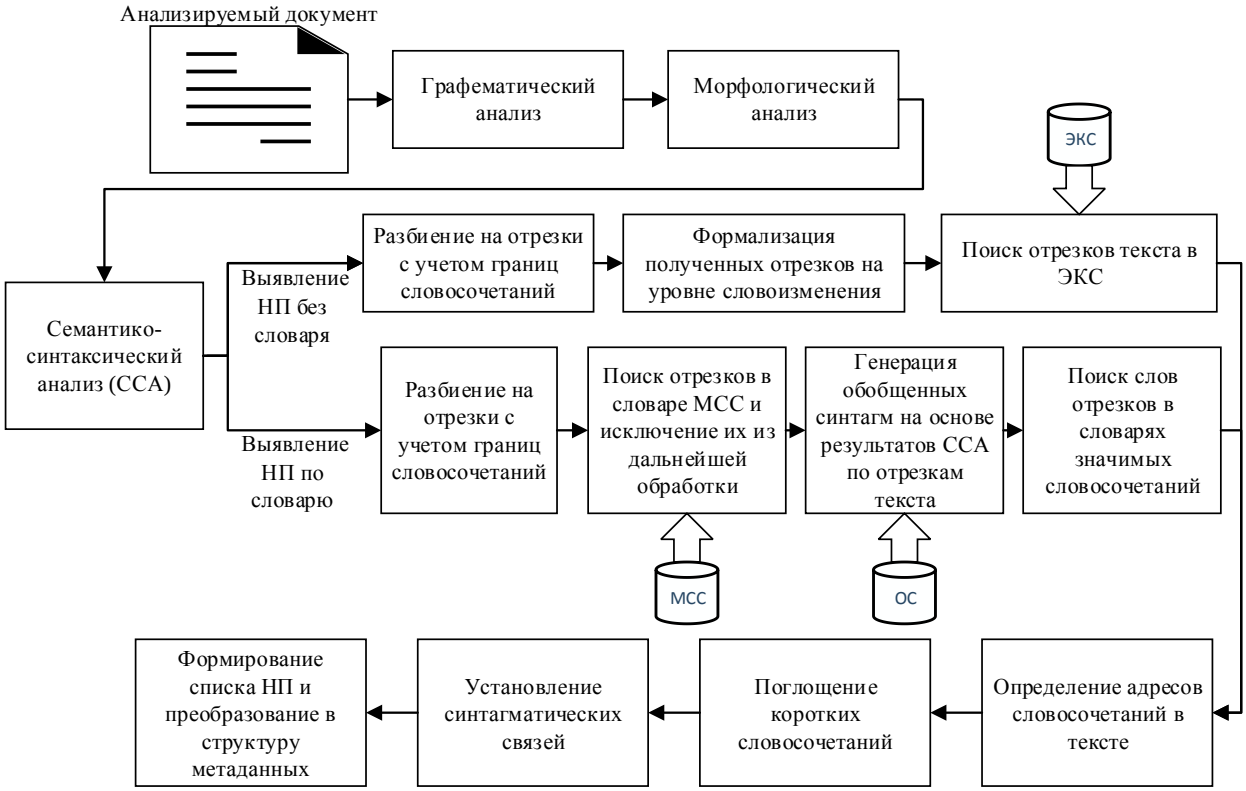

Рис. 1 - Общая схема алгоритма выявления наименований понятий

Во втором разделе рассматриваются методы и алгоритмы формализации наименований понятий. Формализация наименований понятий может выполняться на уровне словоизменения, на уровне словообразования и на уровне, учитывающем явления словоизменения, словообразования, а также явления синонимии и гипонимии. Этот вариант формализации может быть трансформирован в вариант формализации, обеспечивающий приведение наименований понятий к их унифицированным формализованным представлениям. Под унифицированным формализованным представлением наименования понятия (УФПНП) понимается одна из его форм, зафиксированная в словаре. Ниже представлен алгоритм приведения наименования понятия к его унифицированному формализованному представлению.

Алгоритм 2 - приведения наименования понятия к его унифицированному формализованному представлению.

Шаг 1. Произвести приведение буквенного состава слов наименования понятия к нижнему регистру. Выполнить морфологический анализ слов

наименования понятия и произвести пословную нормализацию на уровне словоизменения.

Шаг 2. Произвести поиск формализованного наименования понятия на уровне словоизменения в словаре УФПНП. В случае его совпадения с каким-либо элементом словаря соотнести его с номером элемента словаря УФПНП. Перейти к шагу 5. В случае отсутствия такого совпадения перейти к шагу 3.

Шаг 3. Произвести поиск формализованного наименования понятия на уровне словоизменения в словаре ЭКС. В случае его совпадения с каким-либо элементом словаря, соотнести анализируемое наименование понятия с этим элементом словаря и присвоить ему номер элемента словаря ЭКС. Перейти к шагу 5. В случае отсутствия такого совпадения перейти к шагу 4.

Шаг 4. Выполнить формализацию наименования понятия на уровне словообразования (с изменением порядка следования слов) и присвоить этой форме представления наименования понятия нулевой номер.

Шаг 5. Преобразовать полученное формализованное представление наименования понятия в структуру метаданных.

В третьем разделе описываются технологии автоматизированного формирования декларативных средств. Описан процесс автоматизированного составления словаря наименований понятий по репрезентативному корпусу текстов, представляющему конкретную предметную область.

В **главе 3** рассматриваются методы, модели и алгоритмы формализации смыслового представления текстов документов и процесса выявления всех случаев заимствования, включая случаи неявно выраженных заимствований.

В первом разделе приводится теоретическое обоснование методов обнаружения неявно выраженных заимствований в текстах документов. В качестве базовой теоретической концепции автор использовал концепцию проф. Г.Г. Белоногова и проф. Р.С. Гиляревского, констатирующую, что смысловое содержание текстов выражается с помощью единиц смысла, входящих в их состав[2]. По их мнению, наиболее устойчивыми единицами смысла являются понятия. Проф. Г.Г. Белоногов определяет термин *«понятие» как «социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания…»*[3].

Понятия занимают центральное место в языке и речи и являются теми базовыми строительными блоками, на основе которых формируются

---

[2] Белоногов Г.Г., Гиляревский Р.С. и др. Развитие систем автоматической обработки текстовой информации. - Нейрокомпьютеры: разработка, применение. - 2010, №8. - С. 4-13.

[3] Белоногов Г.Г. Теоретические проблемы информатики, Том 2. Семантические проблемы информатики. - М.: РЭА им. Г.В. Плеханова, 2008. - 342 с.

смысловые единицы более высоких уровней. Второй по значимости единицей смысла является предложение. Из предложений формируются различного рода сверхфразовые единства, которые представляются в виде последовательностей связного текста. В связном тексте предложения выступают не изолированно друг от друга, а в тесной смысловой связи. В основе этой связи лежат мыслительные образы тех конкретных или абстрактных объектов (ситуаций, явлений), которые человек имеет в виду, когда порождает текст. Образы этих объектов имеют определенную структуру. Кроме того, они дополнительно структурируются человеком при их описании на естественном языке. Соответственно этому структурируется и текст[4].

В работе Лукашевич Н.В.[5] вводятся понятия глобальной и локальной связности текстов. При этом констатируется, что глобальная связность обеспечивает раскрытие темы документа, а локальная связность проявляется во взаимосвязи между соседними единицами текста. В соответствии с нашей моделью под глобальной смысловой связностью текста или его фрагмента будем понимать смысловую связь совокупности наименований понятий текста или его фрагмента, расположенных в определённом порядке. Под локальной смысловой связностью текста или его фрагмента будем понимать смысловую связь конкретного наименования понятия и его контекстного окружения.

Преобразование текстового представления в его формализованное смысловое представление дает возможность сопоставления текстов по их смысловому содержанию. Такое сопоставление смыслового содержания текстов, обеспечивающее выявление идентичных по смыслу фрагментов текстов, на наш взгляд, должно удовлетворять следующим условиям:

1. В двух текстах должна быть пересекающаяся совокупность наименований понятий. Число понятий этой совокупности должно быть равно или превышать число наименований понятий, входящих в состав единичного высказывания.
2. В двух таких текстах должны быть фрагменты, в которых концентрация пересекающихся наименований понятий превышает пороговое значение. Эти фрагменты должны иметь соизмеримые размеры.
3. Эти фрагменты текстов должны быть сходными по составу наименований понятий и порядку их следования.

---

[4] Белоногов Г.Г., Калинин Ю.П. и др. Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации - М.: Русский мир, 2004. - 264 с.

[5] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М:, Изд. Моск. ун-та, 2011 г.- 508 с.

Определение схожего порядка следования наименований понятий в тексте или его фрагменте базируется на предположении, что смысл наименований понятий в значительной степени определяется их контекстным окружением. В нашей модели смысл текста определяется как смысловое содержание совокупности взаимосвязанных наименований понятий, расположенных в нем в определенном порядке. Идентичные по смыслу тексты или их фрагменты должны удовлетворять условиям локальной и глобальной смысловой схожести. Локальная смысловая схожесть (ЛСС) наименований понятий текста определяется как сходство контекстного окружения идентичных наименований понятий в двух текстах или их фрагментах. Глобальная смысловая схожесть (ГСС) текстов или их фрагментов определяется как сходство состава идентичных наименований понятий и порядка их следования в текстах или их фрагментах. Каждое понятие этого фрагмента также должно удовлетворять условию локальной смысловой схожести.

Во втором разделе описывается модель процесса выявления неявно выраженных заимствований в текстах документов. Заимствования будут выявляться в массиве МД, состоящем из $n_{МД}$ документов для документа $Д_р$, в отношении которого производится проверка на наличие неявно выраженных заимствований из одного или нескольких документов массива. Для моделирования процесса выявления возможных заимствований в анализируемом документе целесообразно разбить решаемую задачу на несколько этапов.

На первом этапе необходимо определить подмножество документов массива, в которых возможны заимствования.

Для этого предварительно каждому документа массива, а также при обработке анализируемому документу ставятся в соответствие их формализованные смысловые описания. Эти описания представляют собой совокупность номеров наименований понятий, выявленных в тексте и идентифицированных по словарю УФПНП и сопровождаемых адресами их вхождения в текст, которые представлены идентификатором текста, номером предложения в этом тексте и позицией наименования понятия в предложении. Такое формализованное описание документа будем называть концептуальным образом документа (КОДом).

$$КОД = \{НП_i \mid i \in [1, n_{НП}]\} \quad (1)$$

$n_{НП}$ – количество элементов в концептуальном образе документа;

$$НП_i = (ННПС_i, Адр_i) \quad (2)$$

$НП_i$ – информация об i-ом наименовании понятия;

$ННПС_i$ – номер наименования понятия в словаре УФПНП;

$$Адр_i = \{ИТ_{ij}, ИП_{ij}, ПСП_{ij} \mid j \in [1, n_{Адр_i}]\} \quad (3)$$

$Адр_i$ – информация о местоположении наименования понятия в тексте;

$ИТ_{ij}$ – идентификатор текста, в котором находится наименование понятия;

$ИП_{ij}$ – идентификатор предложения, в котором находится наименование понятия;

$ПСП_{ij}$ – позиция наименования понятия в предложении.

Рассмотрим анализируемый документ $Д_p$ и массив имеющихся в хранилище документов МД.

$$МД = \{Д_j \mid j \in [1, n_{МД}]\} \quad (4)$$

$Д_j$ – j-ый документ массива;

$n_{МД}$ – количество документов в массиве.

Каждый документ $Д_j$ массива предварительно обработан при помощи функции получения КОДа: $КОД_{Д_j} = код(Д_j)$, в результате такому документу соответствует $КОД_j = \{НП_{ji} \mid i \in [1, n_{НП_j}]\}$.

$код(Д)$ – функция получения кода, аргументом которой является текст документа.

Аналогичным образом обрабатывается и анализируемый документ $Д_p$, которому соответствует $КОД_{Д_p} = код(Д_p)$.

Для ограничения числа рассматриваемых документов контрольного массива необходимо выявить только те документы, в которых возможны заимствования. Для этого нужно сопоставить полученный $КОД_{Д_p}$ анализируемого документа с КОДами всех контрольных документов массива. Процесс сопоставления выполняется путем попарного пересечения двух КОДов анализируемого документа и документа массива. Найденные элементы КОДа ($КОД_{Д_pД_j}$), должны включать наименования понятий, входящие как в документ $Д_p$, так и в документ $Д_j$, а адреса наименований понятий записываются из обоих текстов.

На рисунке 2 проиллюстрирован вышеописанный процесс сопоставления КОДов - анализируемого документа и документов контрольного массива.

После этого формируется массив документов $МД' \in МД$, состоящий только из тех документов $Д_j$, для которых выполняется следующее условие:

$$n_{КОД_{Д_pД_j}} > k_{мспп} \quad (5)$$

$n_{КОД_{Д_pД_j}}$ – размерность кода $КОД_{Д_pД_j}$, полученного при пересечении $КОД_{Д_p}$ и $КОД_{Д_j}$.

$k_{мспп}$ – коэффициент, соответствующий необходимому минимальному числу пересекающихся понятий в текстах документов.

На втором этапе процесса необходимо в каждом документе $Д_j \in МД'$ выявленного подмножества документов определить фрагменты текстов, в которых наиболее вероятны заимствования. Предварительно фрагменты текста автоматически выделяются в документе путем его разбиения на всевозможные отрезки текста различной длины, состоящие из контактно расположенных предложений, минимальный размер, которых может быть равен одному предложению, а максимальный - не превышает пороговую величину $k_{\max\phi}$.

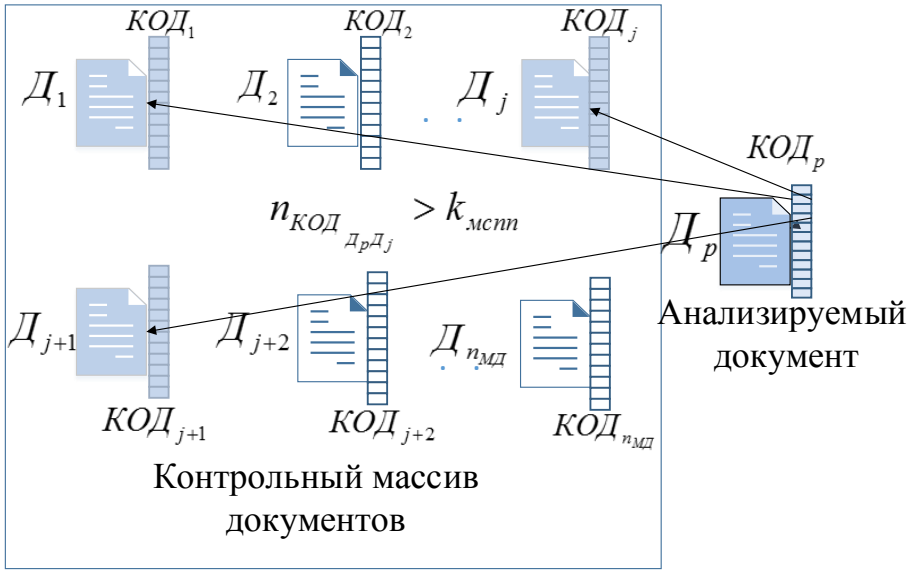

Рис. 2 – Иллюстрация процесса сопоставления КОДов анализируемого документа и документов контрольного массива

Возьмем один из фрагментов текста $Д_p$ длиной в t предложений и начинающийся с l-ого предложения и один из фрагментов текста $Д_j$ длиной в y предложений и начинающийся с z-ого предложения. Этим фрагментам будут соответствовать списки информации (2) о наименованиях понятия фрагментов (СИНПФ) $СИНПФ_{П_{p\,l} П_{p\,l+t}}$ и $СИНПФ_{П_{j\,z} П_{j\,z+y}}$ соответственно.

$$СИНПФ_{П_{p\,l} П_{p\,l+t}} = \{НП_{p\,l\,1}, .., НП_{p\,l\,n_l}, .., НП_{p\,l+t\,n_{l+t}}\} \quad (6)$$

$$СИНПФ_{П_{j\,z} П_{j\,z+y}} = \{НП_{j\,z\,1}, .., НП_{j\,z\,n_z}, .., НП_{j\,z+y\,n_{z+y}}\} \quad (7)$$

Затем производится проверка условия сопоставимости фрагментов текстов:

$$\max(\text{len}(ФТ_{p\,l\,t}), \text{len}(ФТ_{j\,z\,y})) < k_{мдо} \min(\text{len}(ФТ_{p\,l\,t}), \text{len}(ФТ_{j\,z\,y})) \quad (8)$$

$ФТ_{p\,l\,t}$ – фрагмент текста $Д_p$;

$ФТ_{j\,z\,y}$ – фрагмент текста $Д_j$;

$k_{мдо}$ – коэффициент, определяющий во сколько раз могут отличаться размеры сравниваемых фрагментов текста;

$\text{len}()$ – функция определения длины фрагмента текста (количество понятий);

$\min()$, $\max()$ – функции выбора максимального и минимального значений соответственно.

Далее осуществляется сопоставление совокупностей наименований понятий, соответствующих этим фрагментам текста. Для этого воспользуемся мерой Сёренсона. Мера смысловой близости этих двух фрагментов будет определяться по следующей формуле:

$$k_{сб} = \frac{2N(\text{СИНПФ}_{П_{pl}П_{pl+t}} \cap \text{СИНПФ}_{П_{jz}П_{jz+y}})}{N(\text{СИНПФ}_{П_{pl}П_{pl+t}}) + N(\text{СИНПФ}_{П_{jz}П_{jz+y}})} \quad (9)$$

$N(\text{СИНПФ}_{П_{pl}П_{pl+t}})$ – количество понятий во фрагменте $П_{pl} \cap П_{pl+1} \cap .. \cap П_{pl+t}$;

$N(\text{СИНПФ}_{П_{jz}П_{jz+y}})$ – количество понятий во фрагменте $П_{jz} \cap П_{jz+1} \cap .. \cap П_{jz+y}$;

$N(\text{СИНПФ}_{П_{pl}П_{pl+t}} \cap \text{СИНПФ}_{П_{jz}П_{jz+y}})$ - количество общих понятий для обоих фрагментов.

Аналогичным образом производится попарное сравнение всех фрагментов текста. После этого производится выбор текстовых фрагментов максимальной длины с коэффициентом смысловой близости больше порогового значения $k_{псб}$.

На рисунке 3 проиллюстрирован вышеописанный процесс сопоставления фрагментов текстов анализируемого документа и документа контрольного массива.

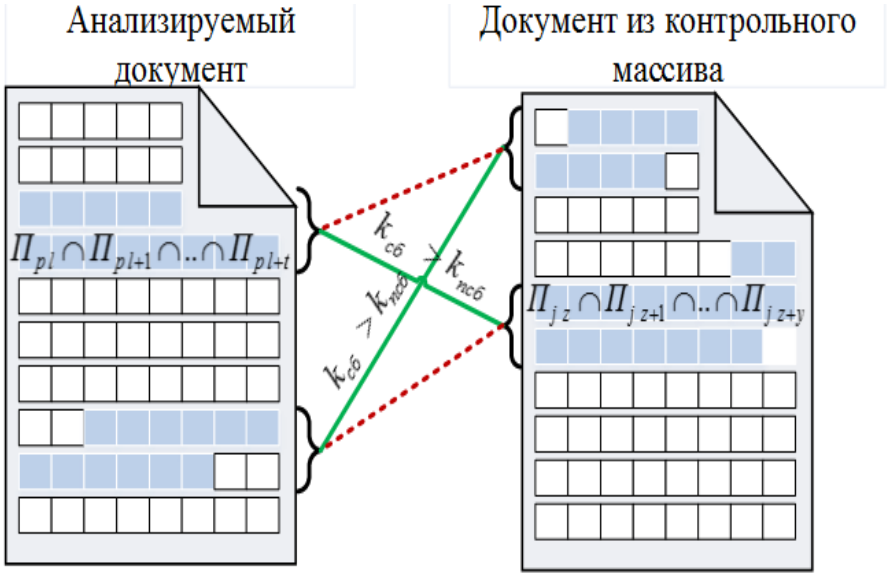

Рис. 3 – Иллюстрация процесса сопоставления фрагментов текстов анализируемого документа и документа контрольного массива

На третьем этапе процесса выявления заимствований необходимо установить смысловую схожесть фрагментов двух текстов, мера смысловой близости которых превышает пороговое значение. Для этого необходимо определить локальную смысловую схожесть наименований понятий, входящих в состав этих фрагментов, путем сопоставления окружающих его справа и слева понятий. В используемой модели для этого необходимо вычислить меру сходства элементов формализованного описания текстовых фрагментов, которые включают в себя контекстное окружение наименования понятия. После этого вычисляется мера глобального смыслового сходства, которая зависит от мер локального смыслового сходства.

На данном этапе, используется усовершенствованная модель представления смыслового содержания текста, где, в отличие от обычного

КОДа, информация о наименованиях понятий дополняется контекстом. Такую модель будем называть концептуальным образом документа, дополненным контекстным окружением (КОДКО).

$$КОДКО = \{НП_i, К_i \mid i \in [1, n_{НП}]\} \quad (10)$$

$$НП_i = (ННПС_i, Адр_i, \text{ОСРНП}_i) \quad (11)$$

$НП_i$ – информация об i-ом наименовании понятия;

$ННПС_i$ – номер наименования понятия в словаре УФПНП;

$Адр_i$ – адреса вхождений наименования понятия в тексте;

$n_{НП}$ – количество наименований понятий;

ОСРНП_i – символ обобщенной синтаксической роли i-ого наименования понятия;

$К_i$ – множество контекстов i-ого наименования понятия, где эти контексты описываются аналогичным образом:

$$К_i = \{НПК_{ik} \mid k \in [1, n_{НПК_i}]\} \quad (12)$$

$$НПК_{ik} = (ННПС_{ik}, Адр_{ik}, \text{ОСРНП}_{ik}, \text{КЗК}_{ik}) \quad (13)$$

$$\text{КЗК}_{ik} = \begin{cases} 1 & \text{, если в разных предложениях с наименованием понятия } \text{НП}_i \\ 2 & \text{, если в одном предложении с наименованием понятия } \text{НП}_i \end{cases} \quad (14)$$

КЗК_{ik} – коэффициент значимости контекста.

Затем вычислим значение меры $м_{ik}$ выполнения условия локального смыслового сходства для каждого наименования понятия из КОДКО сравниваемых документов. В случае $м_{ik} = 0$ данное условие – не выполнено, при $м_{ik} > 0$ – выполнено частично, а при $м_{ik} = 1$ – выполнено полностью.

Если $снп(\text{НП}_{pi}, \text{НП}_{jk}) = 0$, то $м_{ik} = 0$, иначе

$$м_{ik} = \frac{\text{снп}(\text{НП}_{pi}, \text{НП}_{jk})}{3} + \frac{2\text{ско}(\text{К}_{pil}, \text{К}_{jk\text{m}})}{3} \quad (15)$$

$\text{ско}()$ – функция сравнения контекстного окружения наименований понятий;

$$\text{ско}(\text{К}_\text{a}, \text{К}_\text{b}) = \begin{cases} 1 & \text{, фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) > 1 \\ \text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) & \text{, фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) < 1 \end{cases} \quad (16)$$

$\text{фвзбк}()$ – функция вычисления значения близости контекстов;

$$\text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) = \frac{\sum_{c=0}^{n_{НПК_a}} \sum_{d=0}^{n_{НПК_b}} \text{фвппэ}(НПК_{ac}, НПК_{bd})}{4k_{мспп}} \quad (17)$$

$$\text{фвппэ}(НПК_{ac}, НПК_{bd}) = \begin{cases} 0 & , ННПС_{ac} \neq ННПС_{bd} \\ \frac{\text{КЗК}_{ac} + \text{КЗК}_{bd}}{2} & , (ННПС_{ac} = ННПС_{bd}) \wedge (\text{ОСРНП}_{ac} \neq \text{ОСРНП}_{bd}) \\ \text{КЗК}_{ac} + \text{КЗК}_{bd} & , (ННПС_{ac} = ННПС_{bd}) \wedge (\text{ОСРНП}_{ac} = \text{ОСРНП}_{bd}) \end{cases} \quad (18)$$

$\text{фвппэ}()$ – функция вычисления параметра похожести элементов контекстного окружения;

$\text{снп}(\text{НП}_{\text{pi}}, \text{НП}_{jk})$ – функция определения эквивалентности наименований понятий, причем $\text{снп}(\text{НП}_{\text{pi}}, \text{НП}_{jk}) \in \{0,1\}$, НП_{pi} – i-ый элемент формализованного смыслового описания рассматриваемого документа, НП_{jk} – k-ый элемент формализованного смыслового описания j-ого документа контрольного массива.

Условием глобального смыслового сходства является сходство порядка следования наименований понятий, но, поскольку порядок следования наименований понятий учтен при подсчете коэффициентов $м_{ik}$, с точностью до перестановок слов и словосочетаний, которые возможны в идентичных по смыслу текстах, после этого производится поиск последовательностей наименований понятий, у которых значения локальной смысловой схожести $м_{ik}$ выше некого заданного порога $k_{псх}$. Эта последовательность и будет отрезком текста, для которого выполняется условие глобального смыслового сходства, а мера его выполнения вычисляется как среднее значение характеристик выполнения условия локального смыслового сходства, содержащихся в этих последовательностях наименований понятий. Эта величина и будет являться коэффициентом смыслового сходства фрагментов текстов:

$$k_{сх} = \frac{\sum_{i=0}^{n_{\text{НП}_\text{р}}} \max_k(м_{ik})}{n_{\text{НП}_\text{р}}} \qquad (19)$$

$\max_k(м_{ik})$ – максимальное значение $м_{ik}$, при $k \in [1, n_{\text{НП}_j}]$; $n_{\text{НП}_\text{р}}$ – число элементов в КОДКО рассматриваемого документа; $n_{\text{НП}_j}$ – число элементов в КОДКО j-ого документа контрольного массива.

На рисунке 4 проиллюстрирован вышеописанный процесс установления смысловой схожести фрагментов анализируемого документа и документа контрольного массива.

В третьем разделе приводится алгоритм процесса выявления неявно выраженных заимствований в текстах. Этот алгоритм был разработан на основе вышеописанной модели процесса выявления неявно выраженных заимствований в текстах документов. Необходимым условием для реализации этого алгоритма является наличие процедур семантико-синтаксического и концептуального анализа текстов и предварительно адаптированного к предметной области ЭКС. Исходными данными для работы алгоритма являются: массив документов МД и документ $Д_р$, в отношении которого необходимо установить, существуют ли в нем фрагменты, идентичные по своему смысловому содержанию каким-либо фрагментам документов массива МД. Порядок работы алгоритма следующий.

Алгоритм 3 - выявления неявно выраженных заимствований (ВНВЗ) в текстах документов.

Шаг 1. Все документы массива МД и анализируемый документ $Д_p$ обработать процедурами семантико-синтаксического и концептуального анализа текстов и получить их КОДы.

Шаг 2. Попарно пересечь элементы КОДов анализируемого документа с КОДами документов массива. Выявить подмножество документов $МД' \in МД$, в которых возможны заимствования.

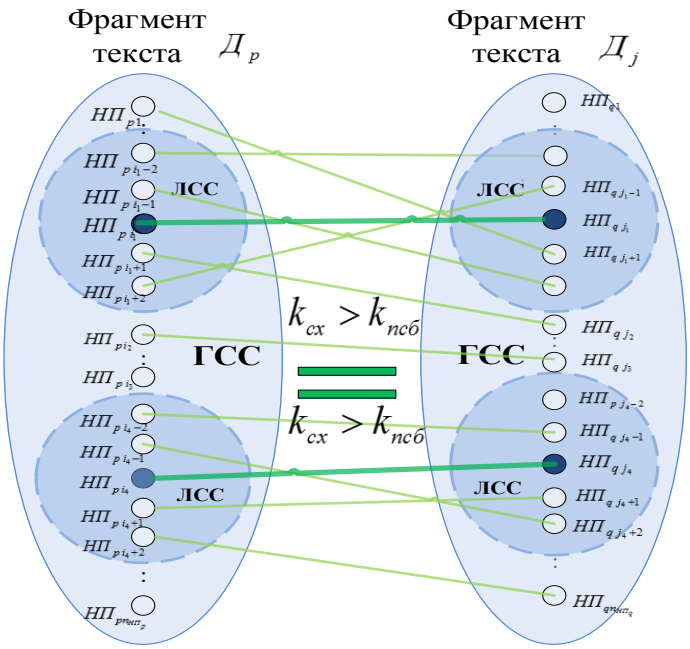

Рис. 4 – Иллюстрация процесса установления смысловой схожести фрагментов анализируемого документа и документа контрольного массива

Шаг 3. Разделить тексты подмножества документов и текст анализируемого документа на фрагменты, состоящие из контактно расположенных предложений, минимальный размер которых может быть равен одному предложению, а максимальный - не превышает пороговую величину $k_{\max\phi}$.

Шаг 4. Попарно сравнить понятийный состав фрагментов, полученных для рассматриваемого документа $Д_p$ и $Д_j \in МД'$, после чего вычислить для них коэффициент смысловой близости и установить пары наиболее близких по смысловому содержанию фрагментов анализируемых текстов.

Шаг 5. Для каждой установленной на шаге 4 пары близких по смыслу фрагментов текстов определить меру локальной смысловой схожести всех наименований понятий этих фрагментов по формуле (15).

Шаг 6. Вычислить степень глобальной смысловой схожести фрагментов по формуле (19).

Шаг 7. Для всех фрагментов анализируемого текста и текстов массива, удовлетворяющих условиям смысловой близости и глобальной схожести, определить адреса их вхождений в эти тексты.

Шаг 8. Полученные данные о наличии возможных заимствований в анализируемом тексте преобразовать в общую структуру метаданных системы.

На основе этого алгоритма было разработано экспериментальное программное обеспечение системы выявления неявно выраженных заимствований в научно-технических документах.

В **главе 4** приводятся результаты исследований эффективности описанных методов, моделей и алгоритмов. Автор исследования разработал

экспериментальное программное обеспечение, реализующее комплекс процедур автоматического выявления неявно выраженных заимствований в текстах научно-технических документов**.**

В первом разделе описана программно-лингвистическая платформа МетаФраз, построенная на базе теоретической концепции фразеологического концептуального анализа текстов. Платформа МетаФраз обеспечивает реализацию всего технологического цикла преобразования текстового представления документа в его формализованное смысловое представление. Она разработана в виде единого интегрированного многофункционального программного комплекса, состоящего из нескольких подсистем, предназначенных для решения отдельных функциональных задач по обработке, формализации и анализу смыслового содержания разноязычных документов.

Во втором разделе описан разработанный автором экспериментальный программный комплекс выявления неявно выраженных заимствований в научно-технических текстах. Архитектура этого комплекса приведена на рисунке 5.

В соответствии с этой архитектурой программный комплекс выявления неявно выраженных заимствований в текстах базируется на программно-лингвистической платформе МетаФраз и включает три подсистемы:

Подсистема выявления неявно выраженных заимствований в текстах предназначена для реализации всего цикла автоматической обработки текста, формализации его смыслового представления и обеспечения возможности сопоставления этого представления с аналогичными представлениями других текстов.

Подсистема управления и визуализации процесса выявления заимствований в текстах предназначена для обеспечения управления процессами обработки текстов, настройки параметров обработки и визуализации основных этапов и результатов этой обработки. Подсистема включает комплекс графических интерфейсов процессов обработки текстов.

Подсистема хранения текстов документов и их формализованных смысловых представлений предназначена для обеспечения процессов загрузки, обработки, формализации и хранения текстов и их формализованных смысловых представлений, а также поиска по этим представлениям.

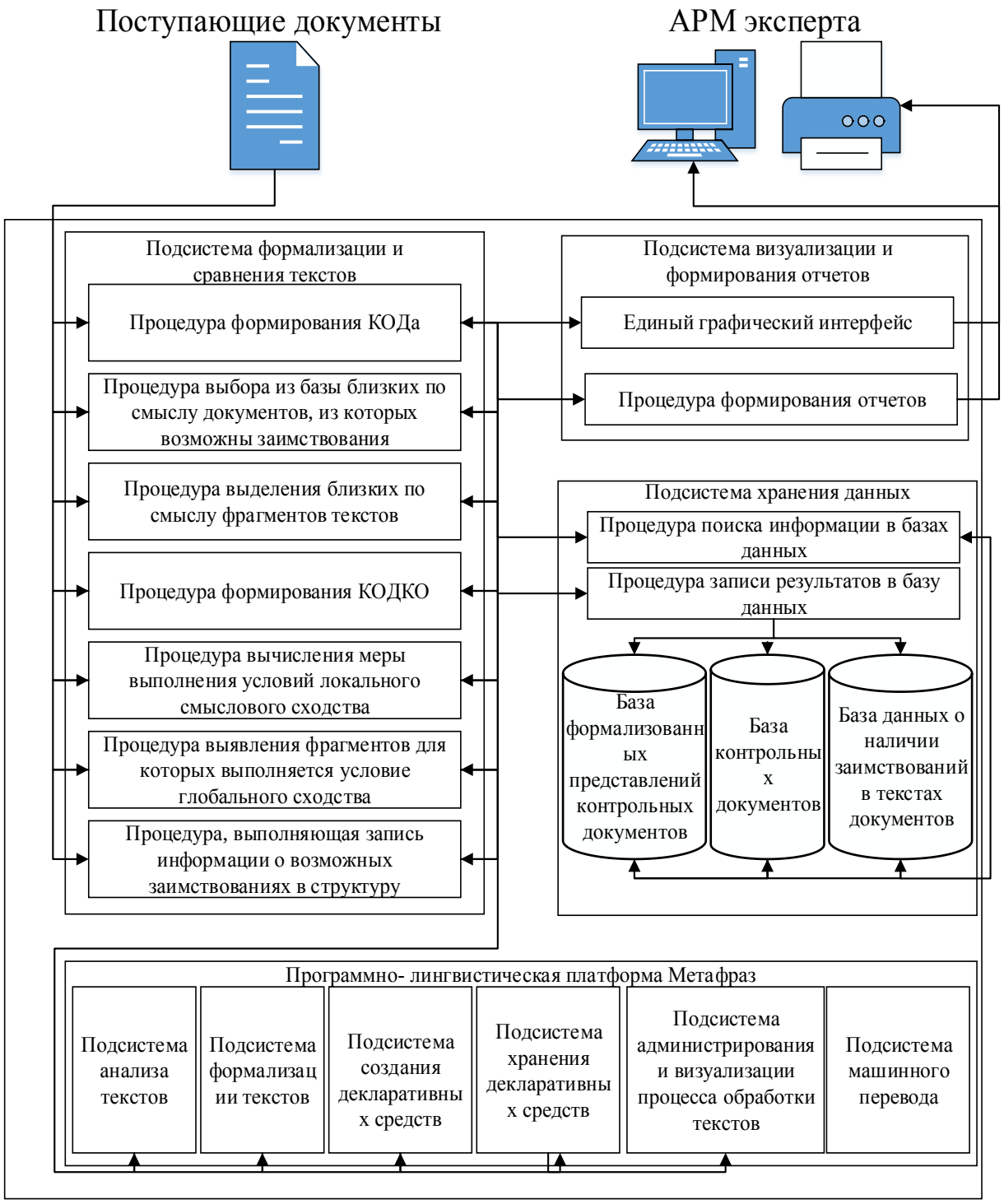

Рис. 5 – Архитектура экспериментального программного комплекса выявления неявно выраженных заимствований

Функционирование системы выявления неявно выраженных заимствований в текстах документов начинается с загрузки контрольных текстов в базу данных системы. В процессе загрузки эти тексты автоматически обрабатываются, их смысловое содержание выявляется, формализуется и заносится в базу данных. Анализируемый документ подвергается аналогичной

обработке и формализации его смыслового представления. Процесс выявления заимствований включает несколько этапов: выявление документов-кандидатов, в которых возможны заимствования, определение в них фрагментов текстов, в которых наиболее вероятны заимствования и установление смысловых отношений и смыслового сходства между этими фрагментами текстов. В случае если эти фрагменты текстов удовлетворяют условиям смысловой близости и сходства, то такие документы с обозначенными в них фрагментами подаются эксперту для принятия решения.

На рисунках 6, 7, 8 проиллюстрирована работа метода ВНВЗ в случае явного заимствования (Рис. 6), в случае неявного заимствования (Рис. 7) и в случае, когда заимствование отсутствует, но совпадает лексический состав фрагментов текстов (Рис. 8).

Для оценки эффективности метода ВНВЗ были посчитаны параметры полноты, точности и F_1 – меры процесса выявления заимствований в текстах документов и эти результаты были сопоставлены с аналогичными параметрами, посчитанными для выявления плагиата методом «шинглов». В общем случае параметр точности определяется как отношение числа релевантных документов к общему числу найденных документов. Для нашего случая точность будет определяться как отношение суммы размеров найденных заимствованных фрагментов текстов к сумме размеров найденных фрагментов.

$$\text{Precision} = \frac{L_{rel} \cap L_{retr}}{L_{retr}} \tag{20}$$

L_{rel} - сумма размеров заимствованных из контрольного массива фрагментов;

L_{retr} - сумма размеров найденных заимствованных фрагментов текстов;

В общем случае параметр полноты определяется как отношение числа найденных документов к общему числу релевантных документов в базе. Для нашего случая полнота будет определяться как отношение суммы размеров найденных заимствованных фрагментов к сумме размеров заимствованных из контрольного массива фрагментов.

$$\text{Recall} = \frac{L_{rel} \cap L_{retr}}{L_{rel}} \tag{21}$$

L_{rel} - сумма размеров заимствованных из контрольного массива фрагментов;

L_{retr} - сумма размеров найденных заимствованных фрагментов текстов;

Значение F_1 – меры будет определяться по следующей формуле:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

На рисунке 9 показано изменения значения F_1 - меры для различных коллекций документов при разных пороговых значениях смысловой схожести (k_{ncx}=0.55, k_{ncx}=0.65 , k_{ncx} = 0.75).

Обобщенные результаты оценки эффективности разработанных методов представлены в таблице 1.

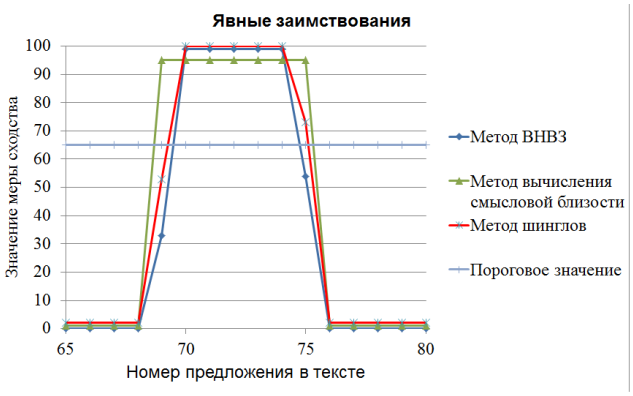

Рис. 6 - Иллюстрация работы метода ВНВЗ в случае явного заимствования

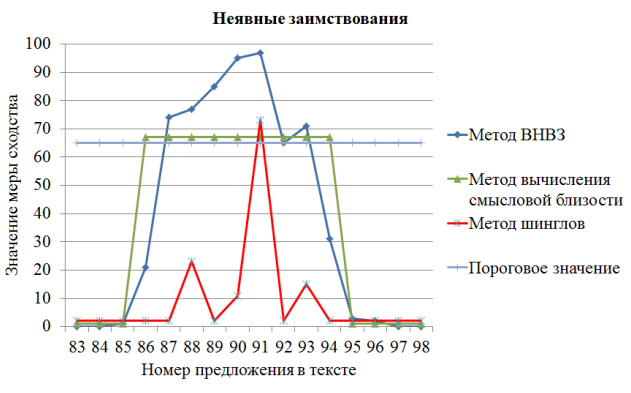

Рис. 7 - Иллюстрация работы метода ВНВЗ в случае неявного заимствования

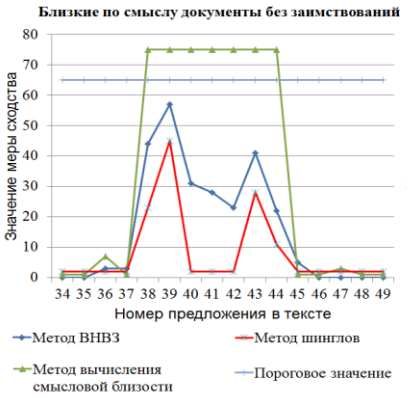

Рис. 8 - Иллюстрация работы метода ВНВЗ в случае, когда заимствование отсутствует, но совпадает лексический состав фрагментов текстов

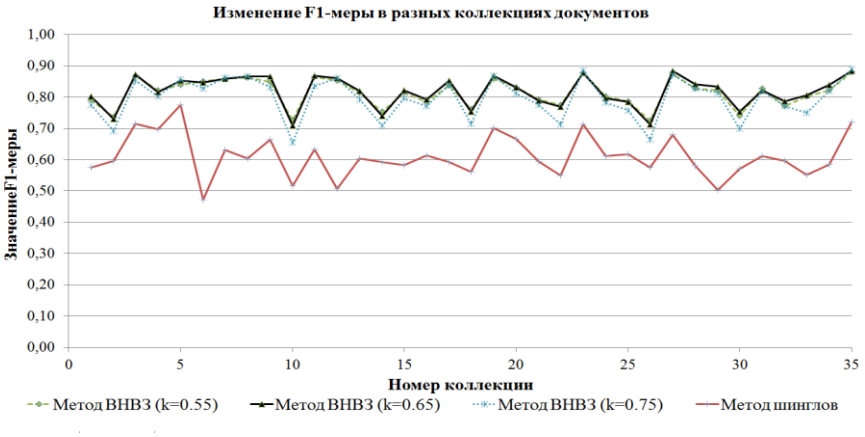

Рис. 9 - Изменение значения F_1 - меры для различных коллекций документов при разных пороговых значениях смысловой схожести (k_{ncx} =0.55, k_{ncx} =0.65 , k_{ncx} = 0.75)

Таблица 1.

Среднее значение полноты и точности выявления заимствований методом ВНВЗ и методом «шинглов»

Полнота		***Точность***		***F1 – мера***	
метод ВНВЗ	метод «шинглов»	метод ВНВЗ	метод «шинглов»	метод ВНВЗ	метод «шинглов»
0.73	**0.45**	**0.94**	**0.96**	**0.82**	**0.61**

Анализ результатов эксперимента (размер коллекции - 5398 документов) показал, что параметры полноты и F_1 – меры выявления заимствований в текстах массива научно-технических документов, полученные с помощью разработанного метода, основанного на концептуальном анализе текстов, в среднем выше на 28 и 21%, чем такие параметры, полученные с помощью метода «шинглов» (при длине шингла равной 3), но при этом наблюдается незначительное снижение (в среднем на 2%) параметра точности выявления заимствований. Наилучшие результаты выявления заимствований были получены при значении коэффициентов смысловой близости и смысловой схожести равном 0.65.

Заключение

В процессе выполненных исследований были получены следующие результаты:

1) Исследованы и модернизированы методы унификации смыслового представления наименований понятий (с учетом явлений словоизменения и словообразования, а также синонимии и гипонимии).

2) Разработана модель процесса выявления заимствований в документах (включая неявно выраженные) на основе анализа их смысловой структуры.

3) Разработан метод установления смысловой близости и смысловой схожести фрагментов текста на основе анализа их смысловой структуры.

4) Разработан алгоритм выявления наименований понятий в научно-технических текстах.

5) Разработан алгоритм автоматического установления смысловых отношений между наименованиями понятий.

6) Разработан алгоритм процесса выявления неявно выраженных заимствований в текстах документов.

7) Разработан экспериментальный программный комплекс выявления заимствований (включая неявно выраженные заимствования) в научно-технических документах, базирующийся на использовании ПО платформы МетаФраз.

8) Проведено экспериментальное исследование, устанавливающее достоверность теоретических концепций и эффективность разработанных методов выявления заимствований в текстах документов.

Основные публикации по теме диссертации

[Личный вклад соискателя в опубликованные работы]

Статьи в журналах из Перечня ВАК

1. Хорошилов А. А. Автоматическое распознавание смысловой близости документов/ Белоногов Г. Г., Гиляревский Р. С., Хорошилов А. А.// Научно-техническая информация, сер. 2. Информационные процессы и системы/ Всероссийский институт научной и технической информации РАН.–

2011 № 7.– С. 15-22. [Соискателем предложен метод автоматического построения концептуального образа документа].

2. Хорошилов А.А. Методы автоматической кластеризации документов в хранилищах научно-технической информации для решения задачи поиска плагиата в текстах документов/ Борзых А.И., Брагина Г.А.// Информатизация и связь, вып. 8, 2012 г., С. 33-37. [Соискателем предложен метод автоматического установления смысловой близости документов кластера].

3. Хорошилов А.А. Автоматическое формирование визуального представления смыслового содержания документа/ Захаров В. Н. // Системы и средства информатики. 2013. Т. 23. No 1. С. 143-158. [Соискателем написан раздел статьи, посвященный автоматической формализации смыслового представления документа].

4. Хорошилов А.А. Системы обнаружения плагиата нового поколения, базирующиеся на методах концептуального анализа текстов и использовании предметно ориентированных концептуальных словарей // Информатизация и связь, вып. 3, 2013 г., С. 112-118. [Соискателем предложена концепция систем обнаружения плагиата нового поколения].

5. Khoroshilov Alexey A. On The Method for Automatic Determination of Semantic Similarity if the Document Text/ Zakharov Victor N., Khoroshilov Alexander A. // Proceedings of ICAI'14, WORLDCOMP'14, July 21-24, 2014, Las Vegas, Nevada, USA-CRSEA Press, USA, 2014, Vol.II. P. 68-73. [Соискателем предложен метод автоматической формализации смыслового представления документа].

Статьи в научных журналах и сборниках трудов

6. Хорошилов А.А. Автоматическое сопоставление смыслового содержания документов авиационно-космической тематики на основе методов концептуального анализа/ Головачев А.Г. // Научно-практическая конференция студентов и молодых учёных МАИ «Инновации в авиации и космонавтике – 2011». 26 – 30 апреля 2011 года. Москва. Сборник тезисов докладов. – М.: МЭЙЛЕР, 2011 – с. 75-76. [Соискателем предложена концепция автоматического сопоставления смыслового содержания документов].

7. Хорошилов А.А. Автоматическое установление смысловой близости документов в технологиях обработки информации / Головачев А.Г. // Труды X международного научно-технического семинара "Современные технологии в задачах управления, автоматики и обработки информации" - М.: МАИ, Алушта, 2011 г. [Соискателем разработана концепция установления смысловой близости документов].

8. Хорошилов А.А. Использование методов концептуального анализа для определения семантической структуры текстов/ Головачев А.Г. // Научно-техническое творчество молодежи – путь к обществу, основанному на знаниях: Сб. докладов III Межд. науч.-практ. конф. в рамках XI Всеросс.

выставки научно-технического творчества молодежи НТТМ-2011, Москва, ВВЦ, 28 июня-1 июля 2011г. – М.: МГСУ, 2011. [Соискателем предложена концепция определения семантической структуры текстов].

9. Хорошилов А.А. Автоматическое сопоставление смыслового содержания документов на основе методов концептуального анализа // Научная сессия НИЯУ МИФИ-2012. Аннотации докладов. В 3 т. Т.2: Проблемы фундаментальной науки. Стратегические информационные технологии.—М.: НИЯУ МИФИ, 2012. – с 235.

10. Хорошилов А.А. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний/ Захаров В. Н. // Труды XIV-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2012, г. Переславль-Залесский, Россия, 15-18 октября 2012 г. [Соискателем предложен метод автоматической оценки подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний].

11. Хорошилов А.А. Решение проблемы многоязычного поиска текстовой информации на основе использования систем фразеологического машинного перевода // Труды XIII-ой нац. конференции по искусственному интеллекту КИИ-2012, г. Белгород, Россия, 16-20 октября 2012 г.

12. Хорошилов А.А. Методы решения задачи автоматического выявления заимствований в структурированных научно-технических документах на основе их семантического анализа/ Захаров В. Н. // Труды XV-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14 – 17 октября 2013 года. [Соискателем предложены методы решения задачи автоматического выявления заимствований в структурированных научно-технических документах на основе их семантического анализа].

13. Хорошилов А.А. Методы автоматического установления смысловой близости документов на основе их концептуального анализа // Труды XV-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14 – 17 октября 2013 года.

14. Хорошилов А.А. Методы выявления имплицитно выраженных заимствований в научно-технических текстах на основе их концептуального анализа // Труды XV-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – DAMDID/RCDL'2015, Обнинск, 14 – 17 октября 2015 года.

Свидетельства об официальной регистрации программ для ЭВМ

15. Хорошилов А.А. Система перевода МетаФраз R10 (MF Translation System R10) / Никитин Ю.В., Смирнов М.В., Садовников Д.А. и др. // Свидетельство о гос. регистрации программы для ЭВМ № 2014663082 от

15.12.2014. – 1 с.[Соискатель разработал семантико-синтаксический анализ для русских и английских текстов].

16. Хорошилов А.А. Надстройки системы перевода МетаФраз R10 (MF Lingware Add-in R10) / Никитин Ю.В., Смирнов М.В., Садовников Д.А. и др. // Свидетельство о гос. регистрации программы для ЭВМ № 2014662528 от 02.12.2014. – 1 с. [Соискатель разработал надстройку для семантико-синтаксического анализа для русских и английских текстов].

17. Хорошилов А.А. Лингвистический комплекс МетаФраз R10 (MF Lingware Complex R10) / Никитин Ю.В., Смирнов М.В., Садовников Д.А. и др. // Свидетельство о гос. регистрации программы для ЭВМ № 2014663079 от 15.12.2014.– 1 с.[Соискатель разработал семантико-синтаксический анализ для русских и английских текстов].

18. Хорошилов А.А. Система семантической обработки текстов МетаФраз R10 (MF Text Analyst R10) / Никитин Ю.В., Смирнов М.В., Садовников Д.А. и др. // Свидетельство о гос. регистрации программы для ЭВМ № 2014663081 от 15.12.2014. – 1 с.[Соискатель разработал семантико-синтаксический анализ для русских и английских текстов].

19. Хорошилов А.А. Сервер лингвистического ПО МетаФраз R10 (MF Lingware Server R10) / Никитин Ю.В., Смирнов М.В., Садовников Д.А. и др. // Свидетельство о гос. регистрации программы для ЭВМ № 2014662743 от 08.12.2014.– 1 с.[Соискатель разработал серверное приложение семантико-синтаксического анализа для русских и английских текстов].

20. Хорошилов А.А. Лингвистический интеграционный комплект МетаФраз R10 (MF Lingware Integration Kit R10 – MF LIK R10) / Никитин Ю.В., Смирнов М.В., Садовников Д.А. и др. // Свидетельство о гос. регистрации программы для ЭВМ № 2014662529 от 02.12.2014. – 1 с. [Соискатель разработал средства интеграции модулей семантико-синтаксического анализа для русских и английских текстов].