ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА Д 002.073.01 НА БАЗЕ ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО УЧРЕЖДЕНИЯ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР «ИНФОРМАТИКА И УПРАВЛЕНИЕ» РОССИЙСКОЙ АКАДЕМИИ НАУК» ПО ДИССЕРТАЦИИ НА СОИСКАНИЕ УЧЕНОЙ СТЕПЕНИ КАНДИДАТА НАУК

аттестационное дело № _____

решение диссертационного совета от 9 декабря 2015 года № 12

О присуждении Хорошилову Алексею Александровичу, гражданину Российской Федерации, ученой степени кандидата технических наук.

Диссертация «Методы, модели, алгоритмы и экспериментальное программное обеспечение автоматического выявления неявно выраженных заимствований в научно-технических текстах» по специальности 05.13.17 – «Теоретические основы информатики» принята к защите 7 октября 2015 года, протокол № 8 диссертационным советом Д 002.073.01 на базе Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», 119333, Москва, Вавилова, д.44, кор.2, приказ № 714/нк от 02.11.2012.

Соискатель Хорошилов Алексей Александрович, 1988 года рождения. В 2011 году соискатель окончил Государственное образовательное учреждение высшего профессионального образования Московский авиационный институт (государственный технический университет) «МАИ», работает научным сотрудником в Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук».

Диссертация выполнена в Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук».

Научный руководитель – доктор технических наук, доцент, Захаров Виктор Николаевич, ученый секретарь Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук».

Официальные оппоненты:

Максимов Николай Вениаминович, доктор технических наук, профессор, профессор кафедры системного анализа Федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский ядерный университет «МИФИ»;

Лукашевич Наталья Валентиновна, кандидат физико-математических наук, ведущий научный сотрудник Научно-исследовательского вычислительного центра Московского государственного университета имени М.В. Ломоносова (НИВЦ МГУ), – дали положительные отзывы на диссертацию.

Ведущая организация - Федеральное государственное бюджетное учреждение науки Библиотека по естественным наукам Российской академии наук (БЕН РАН) в своем положительном заключении, подписанном Цветковой Валентиной Алексеевной, доктором технических наук, профессором, заместителем директора по научной работе БЕН РАН и Погорелко Константином Павловичем, кандидатом технических наук, ведущим научным сотрудником БЕН РАН, указала, что «диссертация А.А. Хорошилова является законченной научно-технической работой, выполнена на высоком научном уровне. Сформулированные положения подкрепляются экспериментальными исследованиями. Автореферат достаточно полно отображает содержание диссертации. Диссертационная работа отвечает критериям п. 9, 10 «Положения о присуждении ученых степеней», а ее автор, Хорошилов Алексей Александрович, заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики».

Соискатель имеет 14 опубликованных работ, все по теме диссертации, в том числе 5 работ (из них без соавторства – 1) в рецензируемых научных изданиях общим объемом 1.75 авт. листа, 9 работ общим объемом 1.7 авт. листа – в сборниках трудов (из них без соавторства – 4), по теме работы получено 6 свидетельств об официальной регистрации программ для ЭВМ (в соавторстве). Наиболее значимые работы соискателя:

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

3. Захаров В.Н., Хорошилов А.А. Автоматическое формирование визуального представления смыслового содержания документа // 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.

На диссертацию и автореферат поступили положительные отзывы от сотрудников Академии ФСО России к.т.н. Толкунова Александра Александровича и к.т.н. доцента Белякова Эдуарда Викторовича, профессора кафедры физики МАИ (Национальный технический университет) д.ф-м.н. Измайлова Георгия Николаевича, ведущего научного сотрудника Всероссийского института научной и технической информации РАН д.ф-м.н. Забежайло Михаила Ивановича, руководителя Центра по изучению проблем информатики Института научной информации по общественным наукам РАН к.философ.н. Чёрного Юрия Юрье-

вича. Авторы отзывов отмечают актуальность темы исследования, высоко оценивают практическую значимость полученных результатов и их новизну. В отзывах ведущей организации, официальных оппонентов и в отзывах на автореферат содержатся следующие критические замечания:

1. Н.В. Максимов отмечает: 1) В работе приводится сравнение эффективности разработанного метода и распространённого метода выявления заимствований – метода «шинглов» с точки зрения полноты и точности выявления заимствования. Но при этом не приводятся сравнительные характеристики по их быстродействию и затратам вычислительных ресурсов. 2) В работе приводится три метода формализации наименований понятий, но из текста работы не совсем понятно для каких задач и в каких случаях необходимо использовать каждый из предлагаемых методов. 3) Пороговое значение смысловой близости автором выбрано экспериментально, однако неясно надо ли и как будет он определяться для других предметных областей или массивов документов другого вида. 4) В работе имеются некоторые, на мой взгляд, неудачные формулировки. Некорректно представлено выражение для меры смысловой близости. Также представляется, что предложенные модели и алгоритмы инвариантны по отношению к предметной области (а не специфичны для научно-технических текстов): по крайней мере, из изложенного материала диссертации не следует явных ограничений на область применения.
2. Н.В. Лукашевич отмечает: 1) Во фразе ".. семантические профили слов выражаются в терминах явных (LSA), неявных (ESA) и характерных (SSA) понятий" (стр. 28) допущена неточность. Должно быть " в терминах неявных (LSA), явных (ESA) и характерных (SSA) понятий. 2) В тексте работы (глава 2) единым образом перечисляются различные процедуры для построения концептуального содержания текста, начиная с графематического и морфологического анализа, и при изложении не выделено, с какого момента начинаются этапы, относящиеся к защищаемым результатам работы, выполненным лично автором. Только в заключении главы сказано, какие именно про-

цедуры относятся к авторству диссертанта. Данная особенность затрудняет восприятие предложенных лично автором методов. 3) Стр. 54. Непонятна фраза: "Далее на основе результатов семантико-синтаксического анализа текстов устанавливаются синтагматические связи между наименованиями понятий, и с помощью процедуры установления смысловых связей между понятиями производится замена родовых понятий на их видовые понятия". Казалось бы более естественно выполнять процедуру обобщения, но написано "родовых понятий на видовые". А если видовых понятий - несколько? 4) Стр. 61-62. Указана неправильная ссылка на таблицу. Написано таблица 2.10, на самом деле 2.11. Это затрудняет чтение текста.

3. Ведущая организация отмечает: 1) Автор в работе подробно описывает программно-лингвистический инструментарий, используемый им в дальнейшем для построения модели процесса выявления заимствований. При этом не совсем понятно, возможно ли использование другого инструментария для реализации поставленных автором задач исследования. Например, инструментария, базирующегося на ситуационно-реляционном подходе или расширенных сетях переходов. 2) Общеизвестно, что в зарубежной литературе часто излагаются идеи, методы, требующие решения в отечественной практике. Некоторые недобросовестные авторы иногда выдают такие методы и идеи за свои собственные. Современные системы выявления заимствований не способны выявлять такие заимствования. Автор не указал, возможно ли использование предложенного метода для выявления заимствований из зарубежных источников. 3) Из описания алгоритма выявления заимствований видно, что процесс, выполняемый этим алгоритмом достаточно сложный и требует значительных вычислительных ресурсов. Следовало бы для обеспечения приемлемой скорости работы описанных алгоритмов предусмотреть распараллеливание обработки при работе с большим числом документов. 4) В тексте диссертации на рисунке 2.1 (стр. 57) допущена опечатка, перепутаны местами надписи «Выявление НП без словаря» и «Выявление НП по слова-

рю». 5) Не описана методика работы экспертов, анализирующих результаты работы программного комплекса по выявлению фактов заимствований в документах, в которых возможны неявно выраженные заимствования. 6) Следует добавить процедуру проверки орфографии анализируемых документов.

4. Толкунов А.А. и Беляков Э.В. отмечают: 1) Из описания разработанной модели в тексте автореферата не ясно, как предлагается определять значения некоторых коэффициентов модели: $k_{мспп}$ (стр. 13), $k_{мдо}$ и $k_{\max ф}$ (стр. 14), $k_{псб}$ (стр. 15) , $k_{псх}$ (стр. 17), - которые в совокупности могут оказывать существенное влияние на решение о наличие в тексте неявно выраженных заимствований. 2) Из описания экспериментального исследования в тексте автореферата не ясно, обладают ли предложенные решения устойчивостью по выявлению заимствований к n-кратному последовательному изменению и предъявлению системе проверяемого на заимствования текста, т.е. способностью к адаптации по отношению к наиболее распространенной модели поведения «злоумышленника», осуществляющего заимствование научно-технического текста.
5. Г.Н. Измайлов отмечает: 1) Не совсем понятно, какие результаты покажет разработанный метод, если в тексте будет содержаться большое количество формул. 2) Не показано, как изменятся результаты работы метода выявления заимствований при использовании в формуле (9) вместо меры Сёренсона другой меры близости.
6. М.И. Забежайло отмечает: 1) При проведении экспериментов сравнение предлагаемого соискателем метода проводилось лишь с методом “шинглов”. Полагаю, было бы полезно сопоставить разработанный метод и с другими подходами, на практике использующимися при компьютерном сравнении текстов. 2) Методику выбора порогового значения смысловой схожести $k_{псх}$ следовало бы описать более детально. Нет достаточно полной аргументации, на основании чего в диссертационном исследовании в качестве единой мет-

рики, объединяющей полноту и точность, используется F1– мера. Возможно, стоило придать большую значимость значению параметра полноты. 3) По-видимому, диссертационная работа А.А. Хорошилова и автореферат только выиграли бы, если бы соискатель нашел возможным представить также и оценки вычислительной сложности предложенных им алгоритмов смыслового анализа текстов. (Возможно, это станет одним из направлений дальнейших исследований соискателя в изучаемой им предметной области).

7. Ю.Ю. Черный отмечает: 1) Не описан технологический процесс использования программного комплекса в различных организациях, являющихся прямыми потребителями результатов исследования (ВУЗы, библиотеки, экспертные советы). 2) В автореферате следовало бы представить более полное описание архитектуры, программных модулей и входных и выходных данных программного комплекса, а также системные требования к нему.

Авторы отзывов отмечают, что замечания являются рекомендательными и не снижают высокой оценки проделанной соискателем работы. Авторы отзывов указывают, что диссертация А.А. Хорошилова является законченной научно-квалификационной работой, выполнена на высоком научном уровне и удовлетворяет требованиям ВАК РФ, а ее автор, Хорошилов Алексей Александрович, заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики».

Выбор официальных оппонентов и **ведущей организации** обосновывается их компетенцией и достижениями в соответствующей отрасли науки, наличием публикаций в сфере исследований.

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

- Разработаны методы, алгоритмы и экспериментальное программное обеспечение процесса формализации смыслового представления содержания документов.

– Разработаны модели, методы, алгоритмы и экспериментальное программное обеспечение процесса автоматического выявления заимствований в текстах документов, включая случаи неявно выраженных заимствований.
– Разработана совокупность технических решений, имеющих важное значение для области автоматической обработки неструктурированной текстовой информации.
– Реализовано комплексное решение задачи автоматического выявления заимствований в текстах документов (включая неявно выраженные заимствования) на основе анализа их смысловой структуры.

Теоретическая значимость исследования обоснована тем, что были созданы новые методы автоматического выявления всех видов заимствований (включая неявные), которые базируются на трехэтапной модели процесса обработки текстов, причем на этих этапах используются модели представления смыслового содержания текстов различной степени сложности. При этом был использован комплекс современных методов, среди которых можно выделить методы семантической обработки текстов, статистические методы, методы работы с декларативными средствами.

Значение полученных соискателем результатов исследования для практики подтверждается тем, что:

– результаты исследований по теме диссертационной работы 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 г.г. 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;

- разработанные методы и алгоритмы 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;
- определена возможность использования разработанных методов в автоматизированных информационных системах, предназначенных для хранения и обработки текстовой информации.

Оценка достоверности результатов исследования выявила:

- идея исследования базируется на анализе практики применения известных методов сравнения текстов, созданных ведущими отечественными и зарубежными учеными;
- экспериментальное исследование, проведенное для оценки эффективности метода, было максимально приближено к условиям его использования в реальных условиях. При этом метод сравнивался с методом «шинглов», который, в настоящее время, наиболее часто используется в системах, позволяющих выявлять заимствования в текстах документов. Для такого сравнения использовались стандартные показатели эффективности – полнота, точность и F1-мера, были получены высокие значения этих параметров;
- теоретические положения опираются на известные, проверяемые данные и факты и согласуются с опубликованными экспериментальными данными по теме диссертации;
- использованы современные методики сбора, подготовки и анализа исходных данных.

Личный вклад соискателя состоит в следующем. Соискателем лично были исследованы использующиеся в настоящее время методы формализации и сравнения текстов, разработаны методы унификации наименований понятий и установления смысловых связей между ними, также соискателем был разработан метод выявления значимых наименований понятий в тексте. Были разработаны модели представления смыслового содержания текста, использующиеся на различных этапах процесса выявления заимствований. Была разработана модель процесса выявления всех видов заимствований (в том числе неявных), состоящая из трех этапов. Выполнена программная реализация разработанных методов, спланирован и проведен эксперимент, подтверждающий их эффективность.

Результаты, относящиеся к теме диссертации, получены автором лично и подготовлены им к публикации самостоятельно.

На заседании «9» декабря 2015 года диссертационный совет принял решение присудить А.А. Хорошилову ученую степень кандидата технических наук.

При проведении тайного голосования диссертационный совет в количестве 21 человек, из них 7 докторов наук по специальности 05.13.17 – «Теоретические основы информатики», участвовавших в заседании, из 24 человек, входящих в состав совета, дополнительно введены на разовую защиту 0 человек, проголосовали: за присуждение ученой степени 21, против присуждения ученой степени 0, недействительных бюллетеней 0.

Председатель диссертационного совета И.А. Соколов

Ученый секретарь совета С.Н. Гринченко

« 11 » декабря 2015 года

М.П.