ОТЗЫВ

официального оппонента на диссертационную работу

Хорошилова Алексея Александровича

"Методы, модели, алгоритмы и экспериментальное программное обеспечение автоматического выявления неявно выраженных заимствований в научно-технических текстах"

по специальности 05.13.17 - "Теоретические основы информатики"

Развитие сети Интернет в значительной мере способствовало возможности несанкционированного использования чужих авторских произведений при создании своих учебных и научных работ. Поскольку с такого рода плагиатом ведется борьба, то для сокрытия заимствования скопированные фрагменты могут перефразироваться, производятся синонимические замены, и, таким образом, создается неявный плагиат, который значительно сложнее выявить.

В диссертационной работе Хорошилова Алексея Александровича исследуется подход к выявлению неявного плагиата, на основе формирования списков понятий, обсуждаемых в тексте.

Диссертационная работа А.А. Хорошилова состоит из введения, четырех глав, заключения, списка использованной литературы, включающего 117 названий, и пяти приложений.

Во **введении** обоснована актуальность темы, определен предмет исследования, сформулированы цель и задачи исследования, приведены методы исследования, изложены основные результаты и их научная новизна, обоснована теоретическая и практическая значимость полученных результатов, а также приведены данные о структуре и объеме диссертации

В **первой главе** рассматривается понятие плагиата, а также подходы, которые используются для обнаружения плагиата, включая подходы на основе векторов, сигнатур и шинглов. Делается вывод о важности использования не пословных, а смысловых подходов к обнаружению плагиата, для чего необходима разработка методов смыслового представления текстов.

Во **второй главе** рассматриваются процедуры автоматической обработки текстов, направленных на структурирование содержания текстов, выявление их

понятийного состава, отношений между понятиями. В частности, описывается процедура установление связей синонимии и гипонимии между словосочетаниями. Важной частью работы является унификация выделенных понятий, которая заключается в том, что главное слова словосочетания записывается первым, все зависимые слова нормализованы на уровне словообразования, и записаны в алфавитном порядке. Таким образом, предложено интересное решение к унификации вариантов написания понятий в текстах.

В **третьей главе** описываются разработанные методы и алгоритмы для выявления всех случаев заимствований, включая неявно выраженные заимствования. На первом этапе определяется подмножество документов массива, в которых возможны заимствования. Это производится на основе построенных понятийных индексов документов при достижении некоторого порога по пересечению множеств понятий документов. Далее сравниваемые тексты разбиваются на фрагменты разной длины, и между фрагментами производится поиск пересечений по упоминаемым понятиям. На третьем этапе процесса выявления заимствований устанавливается смысловая схожесть фрагментов двух текстов, мера смысловой близости которых превышает пороговое значение. Для этого производится определение локальной смысловой схожести наименований понятий, входящих в состав этих фрагментов, путем сопоставления окружающих их справа и слева понятий.

В **четвертой главе** рассматривается разработанный программный комплекс автоматического выявления заимствований. Описывается эксперимент по оценке качества предложенного подхода на основе массива текстов статей Научной электронной библиотеки eLIBRARY.RU. При сравнении предложенного метода с методом шинглов было показано, что точность метода немного ниже, зато полнота значительно более высокая, что приводит к более высокой комбинированной мере качества (F-мера).

В **заключении** описаны основные результаты диссертационной работы и направления дальнейших исследований. Основными новыми научными результатами, полученными в диссертации являются следующие:

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).

В качестве замечаний к тексту диссертации следует отметить следующие:

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 процедуры 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.

Указанные недостатки не являются принципиальными и не умаляют достоинств диссертации.

Всего по теме диссертации опубликовано 14 работ: 5 из них в рецензируемых изданиях из списка ВАК РФ, получено 6 свидетельств об официальной регистрации программ для ЭВМ в Роспатенте. Опубликованные работы полностью отражают результаты диссертации.

Автореферат соответствует диссертации, отражает её содержание и дает представление об актуальности темы, целях, задачах, объекте и методах исследования, научной новизне, практической ценности, реализации, апробации, объеме, кратком содержании и результатах работы.

Практическая ценность работы заключается в том, что научные и практические результаты диссертационного исследования 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 2012-2014гг. 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.

Исходя из вышеизложенного, можно утверждать, что диссертация Хорошилова А.А. на соискание ученой степени кандидата технических наук является законченной научно-квалификационной работой, в которой содержится описание новой модели интегрированного семантико-синтаксического анализа текстов на русском языке.

Считаю, что диссертационная работа Хорошилова Алексея Александровича отвечает всем требованиям Положения ВАК о порядке присуждения ученых степеней, а её автор заслуживает присуждения ему учёной степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики».

Кандидат физико-математических наук,
ведущий научный сотрудник
Научно-исследовательского
вычислительного центра
Московского государственного университета
им. М.В. Ломоносова

20.11.2015

Лукашевич Наталья Валентиновна

Контактные данные:
E-mail: louk_nat@mail.ru
Тел. +7(926)1446163
Адрес: Москва, Ленинские горы, 1, стр. 4

Подпись Лукашевич Н.В. заверяю

Директор НИВЦ МГУ,
проф. Тихонравов А.В.