Структуры заимствований в диссертациях по историческим наукам

© П.В. Ботов © Д.В. Вьючнов © Н.С. Суровенко
© А.С. Хританков © С.В. Царьков © Ю.В. Чехович
ЗАО «Анти-Плагиат»
Москва
khritankov@antiplagiat.ru

Аннотация

В работе описано исследование структуры взаимных заимствований текстовых фрагментов в диссертациях кандидатов и докторов наук по историческим специальностям рубрикатора ВАК (07.xx.xx). С помощью алгоритмических, статистических методов и методов анализа графов и сетей были обнаружены группы сильно связанных по заимствованиям между собой диссертаций, обнаружены «скомпилированные» работы и указаны предполагаемые источники таких компиляций.

1 Введение

В данной статье представлены результаты исследования диссертаций на соискание степеней кандидатов и докторов наук по историческим наукам (коды специальностей ВАК: 07.xx.xx), проведенного по заказу Российской Государственной Библиотеки с использованием Электронной библиотеки диссертаций РГБ (ЭБД РГБ), системы «Антиплагиат» и специального программного обеспечения обработки данных и машинного обучения.

ЭБД РГБ [7] содержит библиографические описания и полные тексты авторефератов и диссертаций по различным специальностям ВАК, полученные путем сканирования текстовых документов.

Система «Антиплагиат» [1, 4, 6, 20] позволяет проводить для текста проверяемого документа и произвольной коллекции источников сравнительный анализ. Результатом такого анализа является список всех значимых фрагментов проверяемого документа, совпадающих полностью или частично с фрагментами в коллекции источников. Совпадения фрагментов текстов документа и источников обозначаются как «заимствования». При этом практически совпадения могут иметь различную интерпретацию: цитирование источника, цитирование третьего неизвестного текста в обеих работах, академический плагиат, использование общеупотребимых словосочетаний, случайное совпадение и т.д. Результат работы системы обычно анализируется экспертом, который и принимает решение о том, как квалифицировать обнаруженные системой заимствования и об академической ценности работы в целом [21]. Работа эксперта требует значительных затрат времени для квалифицированного анализа объемной диссертации – от нескольких часов до нескольких дней на одну работу. С учетом того, что в России ежегодно защищается около 25 тысяч диссертаций, проверка всего потока работ оказывается практически неподъемной задачей.

Основной целью проведенного исследования, таким образом, стала проверка технической возможности глубокого автоматического анализа заимствований в больших текстовых коллекциях для формирования «грубого фильтра» работ для последующего экспертного анализа. Такой фильтр позволил бы выделять часть работ, проведение экспертного анализа которых необходимо. В настоящем исследовании авторы главным образом сосредоточились на выборе процедур предобработки исходных данных, постобработки результатов и настройках параметров системы, с целью автоматизации и уточнения результатов последующей экспертной обработки.

Инициатором и заказчиком исследования выступила РГБ. Основные направления исследования были сформулированы в виде нескольких гипотез. В данной статье представлены результаты по гипотезам и исследовательским вопросам, приведенным в разделе 2.

Для корректного учета заимствований необходимо было исключить из состава обнаруженных совпадений корректно оформленные цитаты (см. раздел 3) и технические заимствования – общие фрагменты диссертаций вследствие использования общего формата, шаблона и правил

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

оформления, а также списка литературы (см. раздел 4).

После предварительной обработки, возможно проведение более глубокого анализа и проверка гипотез (см. раздел 5).

2 Гипотезы и цели исследования

В ходе исследования предполагалось проверить следующие гипотезы и дать ответы на вопросы:

− определить возможность проведения глубокого анализа заимствований в объемных текстовых коллекциях на наличие некорректных заимствований;

− оценить долю работ с существенными заимствованиями текста из других диссертаций;

− понять, является ли подготовка таких работ частью процессов систематической компиляции, либо это единичные не связанные случаи.

3 Выделение корректно оформленных цитат

В тексте диссертации автор может дословно цитировать фрагменты других произведений. Цитаты оформляются в соответствии с правилами русского языка [15], библиографические ссылки к ним – согласно стандарту [16]. Так как цитата дословно повторяет часть другого текста, она может быть распознана поисковыми модулями системы «Антиплагиат» как заимствованный блок, поэтому нужно выделять корректно оформленные цитаты и исключать их из блоков заимствований.

Для выделения цитат предлагается подход, основанный на применении методов машинного обучения и состоящий из трех этапов:

1. Выделение текстовых блоков-кандидатов при помощи эвристик.

2. Расчет значений признаков для блоков-кандидатов.

3. Бинарная классификация блоков-кандидатов по принадлежности к классу корректно оформленных цитат.

На первом этапе текстовые блоки выделяются согласно правилам русского языка [15]. Практически во всех случаях цитируемый текст должен быть заключен в кавычки. Исключением из этого правила являются стихотворения, которые можно цитировать без кавычек в случае сохранения авторских переносов строк. Так как цитирование стихов не свойственно диссертациям по историческим наукам, то для повышения точности распознавания и снижения сложности системы в качестве блока-кандидата выделяется текст, заключенный в кавычки. При этом учитывается, что одни блоки могут быть вложены в другие.

На втором этапе происходит расчет значений признаков блоков-кандидатов. Признаки построены на основе правил оформления цитат и библиографических ссылок. Например, реализован признак, что если после текста цитаты в пределах одного предложения встретилось слово, написанное слитно с числом, или число следует сразу после закрывающей кавычки в блоке-кандидате, то значение признака равно 1, иначе 0.

Таких признаков было построено более 60, однако в результате отбора, о котором будет рассказано ниже, было оставлено только 23.

На третьем этапе к рассчитанным значениям признаков блоков применяется обученная модель дерева решений, выполняющая бинарную классификацию, является ли блок корректно оформленной цитатой или нет.

Для построения и настройки модели были вручную размечены тексты диссертаций по историческим наукам. Для этого была разработана программа разметки корректно оформленных цитат среди блоков текстов с графическим интерфейсом. Всего исходные данные составили 24479 блоков, в которых 4277 корректно оформленных цитат. Набор данных был разделен на обучающие данные из 16320 блоков (из которых 2848 корректно оформленных цитат) и тестовые из 8159 блоков (из которых 1429 цитат).

Далее, на обучающих данных с помощью программы Weka [17] были проанализированы признаки и с применением критерия «Gain Ratio» [18] отобрано 23 признака для классификации блоков.

Для построения дерева решений был использован алгоритм C4.5 [18]. Модель дерева решений использована потому, что ее можно интерпретировать в виде правил «если – то», понятных даже не специалисту в области машинного обучения. Глубина дерева была ограничена значением 7. Оценка качества проводилась по двум критериям: точность и полнота.

Точность – это доля верно выделенных моделью корректно оформленных цитат среди всех выделенных моделью текстовых блоков.

Полнота – это доля верно выделенных моделью корректно оформленных цитат среди всех корректно оформленных цитат.

В результате для использованной в работе модели на обучающей выборке точность составила 96,8%, полнота – 73,5%, на тестовой выборке точность составила 95,8%, полнота – 43,8%.

4 Предварительная обработка данных

Система «Антиплагиат» анализирует тексты документов, строит по ним инвертированный индекс групп последовательно идущих слов (n-грамм) [19] и сравнивает документы попарно после нахождения потенциально совпадающих блоков в индексе.

На вход были поданы тексты диссертаций коллекции ЭБД РГБ по историческим наукам 07.xx.xx, всего более 14 тыс. кандидатских и

докторских диссертаций, защищенных преимущественно в 1999–2012 гг. (рис. 1). Атрибуты библиографического описания диссертаций также получены из ЭБД РГБ. Были исключены 51 документ с ошибками выделения текста и 114 документов размером менее 15 тысяч символов. Бимодальное распределение документов по годам соответствует содержанию ЭБД РГБ и, по видимому, является следствием порядка оцифровки документов в РГБ.

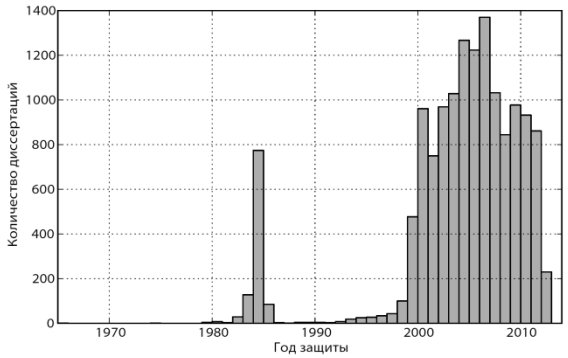

Рис. 1. Количество диссертаций по годам защиты

При поиске заимствований между документами одной коллекции возникает проблема установления направления заимствования и формирования набора источников. В данном исследовании проблема была решена следующим образом. Для каждой диссертации отбиралось 100 источников с наибольшим количеством заимствований из них в данной диссертации. Минимальный размер блока заимствования варьировался от трёх до семи слов в зависимости от контекста. Направление заимствования устанавливалось эвристически по году защиты диссертации. Полагалось, что источником заимствования является диссертация, год защиты который предшествует году зашиты рассматриваемой диссертации.

Вычисления блоков заимствований проводились на сервере с восемью виртуальными ядрами Xeon 1,6 ГГц, 6 ГБ ОЗУ в течение четырех дней. Было проведено три итерации вычислений блоков с различными параметрами. Полное время проведения вычисления блоков с учетом пауз между итерации составило две недели. Общий несжатый объем блоков заимствований в XML формате составил около 4 ГБ.

Полученные блоки заимствования были дополнительно обработаны: выполнено объединение блоков, исключение корректных цитирований, повторное объединение, фильтрация по размеру блока.

Алгоритм объединения блоков составлял из двух блоков, разделенных менее чем 30 символами, один блок, включающий оригинальные блоки и символы между ними (рис. 4).

После объединения блоков из них были исключены корректно оформленные цитаты, сформированы новые блоки, которые были повторно объединены тем же алгоритмом.

Предварительный анализ расположения и размера блоков заимствований (рис. 2) показал, что большая часть совпадающих блоков находится в титульном листе и, по-видимому, области библиографии диссертации. Предполагая, что эти блоки связаны с общим форматом титульного листа и сходными источниками в списке литературы, исключены блоки, находящиеся в первых 1000 символов и последних 10% текста диссертации.

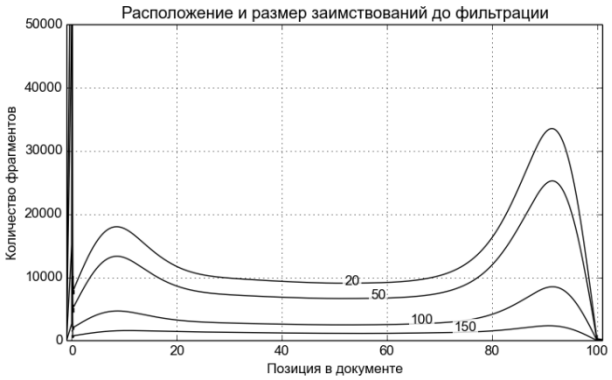

Рис. 2. Размер и позиция блоков до предварительной обработки. Изоденсы обозначают размер блоков, значения выбраны экспертно

По результатам анализа распределения блоков по размеру в разных частях документа, были исключены блоки размером менее 250 символов как незначительные заимствования, по большей части относящиеся к введению и библиографии. В дальнейшем при построении графа заимствований были исключены блоки размером менее 750 символов, в результате пропадает зависимость между размером блока и его положением в документе.

В результате были построены распределение блоков по размеру и положению в документе (рис. 3), направленный граф заимствований, составлен список диссертаций с наибольшей долей заимствованного текста.

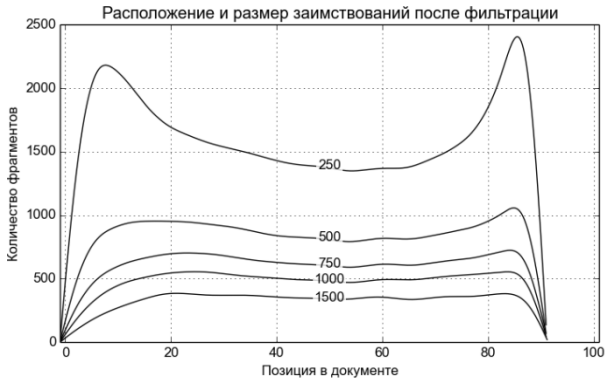

Рис. 3. Размер и позиция блоков после фильтрации, исключения цитат и объединения блоков. Изоденсы обозначают размер блоков, значения выбраны экспертно

В текстах диссертаций были замечены и исследованы аномалии – чаще всего связанные с ошибками оцифровки или обработки документов.

В частности, около 50 документов состояло из склеенных в одном тексте нескольких диссертаций, которые также встречались отдельно.

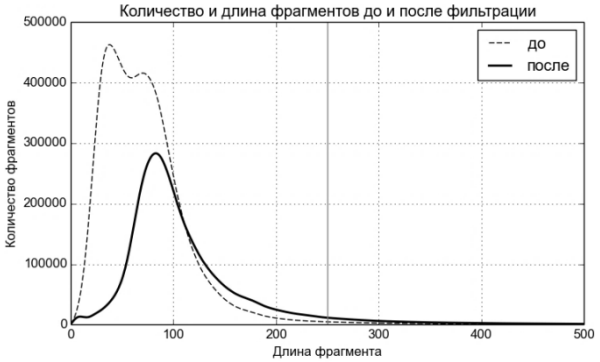

Рис. 4. К описанию алгоритма слияния блоков

5 Выделение групп диссертаций

Анализ групп и сообществ диссертаций позволяет установить «контекст» заимствований между ними, выделить скрытые внутренние структуры заимствований. Для проведения такого анализа заимствования между диссертациями в данной работе был построен граф, в котором в качестве вершин были диссертации, а ребра определялись заимствованиями из этих работ. Вес ребра рассчитывался как количество совпадающего текста в символах.

Для анализа графов и сетей используются специализированные алгоритмы объединения вершин графа в кластеры, называемые сообществами (community). В работе [2] предложен быстрый алгоритм поиска сообществ в графах, основанный на максимизации внутреннего критерия качества – модульности (modularity):

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

где A_{ij} – вес дуги между i и j, $k_i = \sum_j A_{ij}$ – сумма весов дуг, связанных с вершиной i, c_i – сообщество, к которому принадлежит вершина i, δ-функция $\delta(u,v)$ равна 1, если $u = v$, и 0 иначе, и $m = \frac{1}{2}\sum_{ij} A_{ij}$.

Алгоритм выделения сообществ [2] состоит из итеративно повторяющихся двух шагов.

На первом шаге каждая вершина графа приписывается к своему уникальному сообществу. Затем для каждой вершины i рассматривается возможность её переноса в сообщество вершины j, до которой из i есть ребро, при условии, что модульность увеличивается. Процесс повторяется, пока модульность не достигнет локального максимума.

На втором шаге из полученных сообществ получают вершины для нового графа, веса ребер которого определяются суммой весов ребер вершин, входящих в сообщество. Таким образом, первый шаг можно заново выполнить для нового графа.

Итерации продолжаются до тех пор, пока с новой итерацией не перестанет изменяться состав сообществ.

Всего в исходном графе получилось порядка 13 000 вершин и 164 000 ребер. В исходном графе, при отсутствии фильтрации, присутствовала гигантская компонента (giant component) размером в 12000 вершин, что указывало на наличие большого числа «шумовых» ребер. Предполагая, что шумовые ребра имеют небольшой вес, можно подобрать пороговое значение, отсекающее большинство таких ребер. С другой стороны, завышение порога отсечения могло привести к удалению значимых связей между вершинами, образующих сообщества и искажении структуры сообществ в графе. Поэтому необходимо было подобрать порог минимального допустимого веса ребра для выделения сообществ.

В эксперименте были проанализированы зависимости следующих параметров от порога отсечения: количество выделяемых сообществ, количество слабо связанных компонент в графе, максимальный размер связанного компонента (рис. 5–6).

При увеличении порога количество сообществ и связанных компонент возрастало за счет «развала» гигантской связанной компоненты (см. рис. 5), достигло максимума, а затем начало убывать. Эта точка максимума и определила искомый порог отсечения, так как дальнейшее его увеличение приводило к удалению значимых связей между вершинами и уменьшению количества сообществ.

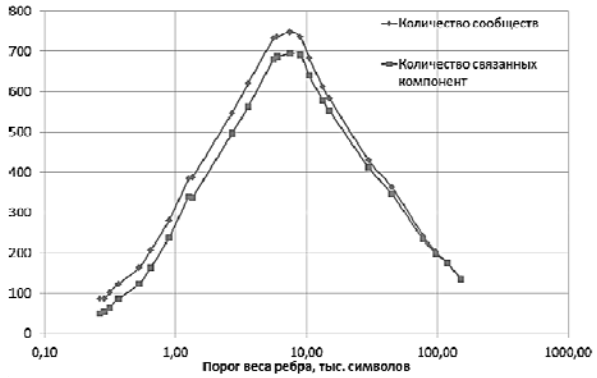

Рис. 5. Зависимость количества связанных компонент и количества сообществ от порога веса ребра

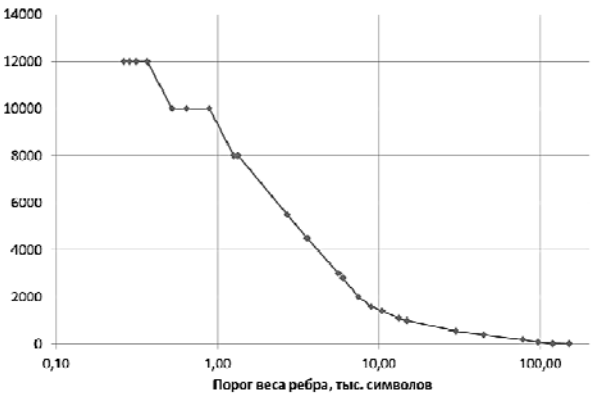

Рис. 6. Зависимость максимального размера связанного компонента от порога веса ребра

В результате порог веса ребра выбран равным 0,05, что соответствует суммарному заимствованию в 7500 символов между диссертациями. При данном пороге в графе выделяется 748 сообществ.

Полученные сообщества характеризуются более высоким уровнем заимствования среди диссертаций сообщества, чем из диссертаций вне сообщества. Пример сообщества и заимствований между диссертациями показан на рис. 7.

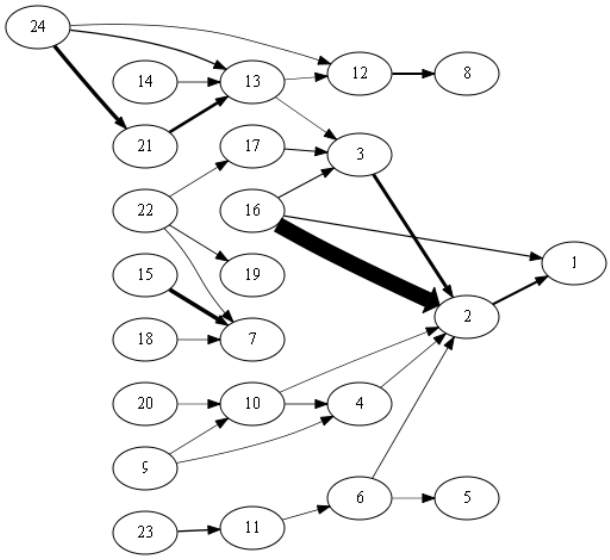

Рис. 7. Пример найденного сообщества. Диссертации представлены вершинами графа и пронумерованы, заимствования показаны ребрами, толщина ребра пропорциональна объему заимствования

В сообществах диссертации могут выполнять две функции: являться источниками для заимствований и получателями заимствований из других источников. На рис. 7 диссертации 24, 16, 22 можно назвать популярными источниками в данном сообществе. Диссертации 2, 3, 7, 13 – получатели заимствований. Заметим, что 2, 3 и 13 при этом так же используются в качестве источников для заимствования другими диссертациями. Жирная стрелка между работами 2 и 16 указывает на большой объем заимствованного текста.

Источники и получатели заимствований можно найти в большинстве сообществ. В таких сообществах существенны заимствования текста между диссертациями, что указывает на наличие коллективов, занимающихся подготовкой диссертаций путем компиляции из других работ. Отнесение источников заимствования к сообществу позволяет увидеть сообщество в целом и не указывает на автора источника как участника коллектива.

Если все сообщества диссертаций расположить на диаграмме с зависимостью полного объема заимствования от среднего их объема по заимствованиям внутри сообщества (рис. 8), то среди них можно выделить три вида. Небольшие сообщества диссертаций с высоким средним объемом заимствований, по-видимому, скомпилированных в индивидуальном порядке из небольшого числа работ назовём «индивидуальными предпринимателями». Большие сообщества с умеренным средним размером заимствований – «фабрики диссертаций», а также «странные сообщества», которые не получается однозначно отнести к предыдущим двум видам. Диссертации из сообществ, не относящихся к указанным, полагаются подготовленными научными группами, не основанными на систематических заимствованиях текстов диссертаций.

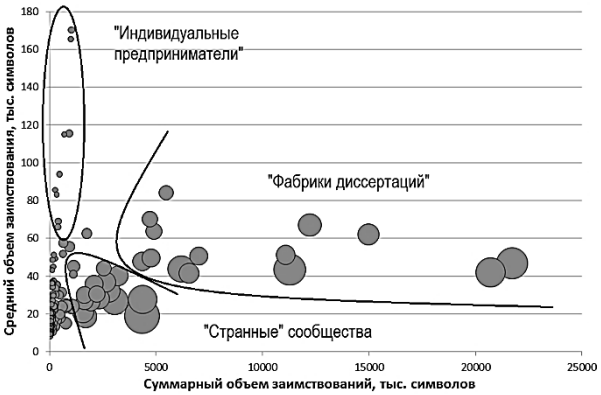

Рис. 8. Сообщества диссертаций по среднему объему заимствования (по вертикали) и суммарному объему (по горизонтали) с условной классификацией по видам. Площадь метки соответствует размеру сообществ, на диаграмме – от 4 до 169 диссертаций

При анализе заимствований в диссертациях, вследствие использования только ЭБД РГБ в качестве источника данных, не учитывались заимствования из других источников, статей, журналов. Такого рода заимствования, в исследуемом графе заимствований, могут косвенно проявляться как заимствования между диссертациями, если в них имеется общий текст из стороннего источника.

6 Сходные исследования

Диссертации, защищаемые в области наук, в целом отражают структуру и состояние исследований в своей области, и представляют отдельный интерес как объект научного исследования. Исследования диссертаций и научных работ, связей между ними проводились ранее в других областях [8–13]. В работах [8, 9] проведено исследование диссертаций и авторефератов с целью выявления научных школ, связей между научными руководителями и диссертантами, использованы методы анализа текстов. В исследовании авторефератов докторских диссертаций [10] проведен анализ качества подготовки диссертаций за 2008–2011 годы по материалам, опубликованным на сайте ВАК.

Проведенное исследование отличается использованием данных ЭБД РГБ [7], полных текстов диссертаций, рассмотрением диссертаций по историческим наукам и механизмом установления связей между диссертациями – по текстовым заимствованиям, и методами анализа

полученного графа. Причем наличие текстовых заимствований, с нашей точки зрения, указывает на общность в подготовке текстов диссертаций.

Определение общности научных работ по текстовым заимствованиям – достаточно распространенный метод [1, 5], однако известны и другие подходы, основанные на методах анализа текстов [13] и рассмотрении совместного библиографического цитирования между документами [14].

7 Заключение

Насколько известно авторам, проведенное исследование по определению структур заимствований в диссертациях является первым в своем роде. Исследованные гипотезы и вопросы ранее не выдвигались. Поэтому так же важно, что были отработаны методы исследования.

Проведенное исследование продемонстрировало техническую возможность проведения анализа заимствований в крупных текстовых коллекциях с применением системы «Антиплагиат» в совокупности с методами анализа данных для фильтрации потока диссертационных работ и выделения документов, для которых необходим последующий экспертный анализ.

Было обнаружено, что большинство проверенных диссертаций не имеют значимых заимствований. Однако не менее 500 работ имеют существенный объем более 33% общих текстовых фрагментов с другими диссертациями, что может указывать либо на наличие общих источников заимствования, либо на прямое заимствование.

В построенном графе заимствований обнаружены коллективы и «сообщества» диссертаций, по-видимому, связанные с процессом их подготовки. Сообщества с большим объемом заимствований между диссертациями отнесены к коллективам, в которых налажен процесс подготовки текстов диссертаций путем компиляции из готовых источников.

Результаты исследований были предоставлены на рассмотрение экспертам РГБ и получили положительную оценку. В дальнейшем планируется проведение подобных исследований и в других областях науки.

Литература

[1] Авдеева Н.В., Ботов П.В., Букаев А.С., Вислый А.И., Груздев И.А., Житлухин Д.А., Романов М.Ю., Чехович Ю.В. Внедрение системы «Антиплагиат» в Российской государственной библиотеке // Интеллектуализация обработки информации: 8-я международная конференция. Республика Кипр, г. Пафос, 17–24 окт. 2010 г.: сб. докл. – М.: МАКС пресс, 2010. – С. 499–503.

[2] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre. Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment 2008(10):P10008 (2008).

[3] R. Lambiotte, J.C. Delvenne, M. Barahona. Laplacian dynamics and multiscale modular structure in networks // Arxiv preprint arXiv:0812.1770 (2008).

[4] ЗАО Анти-Плагиат, Система «Антиплагиат». http://www.antiplagiat.ru

[5] iParadigms, LLC. Turnitin. Plagiarism prevention engine. Available online at: http://www.turnitin.com

[6] Шарапов Р.В., Шарапова Е.В. Система проверки текстов на заимствования из других источников // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2011. –Воронеж, 2011. –С.121–126.

[7] Лавренова О.А. Развитие проекта библиотеки электронных диссертаций и авторефератов в открытом доступе // Образовательные технологии и общество (Educational Technology & Society). – Казань: Изд-во Казанский государственный технологический университет. – 2006. – Т. 9, № 3. – С. 335–341.

[8] Ю.В. Леонова, А.М. Федотов. Извлечение знаний и фактов из текстов диссертаций и авторефератов для изучения связей научных сообществ // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2013, Ярославль, Россия, 14–17 октября 2013 г. – Ярославль: ЯрГУ, 2013. – С. 135–144.

[9] Леонова Ю.В., Добрынин А.А., Веснин А.Ю. Построение графа диссертаций // XIV Российская конференция с участием иностранных ученых «Распределенные информационные и вычислительные ресурсы» (DICR-2012): программа конференции и тезисы докладов (Новосибирск, Россия, 26–30 нояб. 2012). – Новосибирск: ИВТ СО РАН, 2012. – С. 17. – ISBN 978-5-905569-05-0.

[10] Донецкая С.С. Статистическое исследование структуры и качества подготовки докторских диссертаций в России // Вопросы статистики. – 2012. – № 12. – С. 71–76.

[11] Бескаравайная Е.В., Митрошин И.А. Анализ базы данных диссертаций ПНЦ РАН // Информационное обеспечение науки. Новые технологии: сб. науч. тр. / Н.Е. Каленов (ред.). – М.: Научный Мир, 2011. – С. 124–133.

[12] Ю.Н. Климов. Количественно-информационный анализ потока публикаций по библиотекам и библиотековедению на основе поиска по ключевым словам в базе данных Science-Direct // Межотраслевая информационная служба. – 2011. – № 3. С. 51–58.

[13] В.Н. Захаров, А. А. Хорошилов. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15–18 окт. 2012 г. – С. 189–195.

[14] Bela Gipp and Joeran Beel, 2009 "Citation Proximity Analysis (CPA) – A new approach for identifying related work based on Co-Citation Analysis" in Birger Larsen and Jacqueline Leta, editors, Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), volume 2, pages 571–575, Rio de Janeiro (Brazil), July 2009.

[15] Розенталь Д.Э., Джанджакова Е.В., Кабанова Н.П. Справочник по правописанию, произношению, литературному редактированию. – Издание второе, исправленное. – М.: ЧеРо, 1998. – 400 с.

[16] ГОСТ Р 7.0.5–2008 Библиографическая ссылка, общие требования и правила составления.

[17] University of Waitako. Weka Toolkit. http://www.cs.waikato.ac.nz/~ml/weka/

[18] J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.

[19] К.Д. Маннинг, П. Рагхаван, Х. Шютце. Введение в информационный поиск. : Пер. с англ. – М.: ООО «И.Д. Вильямс», 2011. – 528 с.

[20] Авдеева Н.В., Никулина О.В., Сологубов А.М. Система «Антиплагиат.РГБ» и недобросовестные авторы диссертаций: кто победит? // Научная периодика: проблемы и решения. – 2012. – №5(11). – С. 11–16.

[21] Авдеева Н.В., Лобанова Г.А. Классификация фрагментов текста при экспертизе диссертаций на предмет заимствований (плагиата) // «Информационные ресурсы России»: науч.-практ. журн. – М.: ФГБУ «Российское энергетическое агентство» Минэнерго России. – 2014. – № 11. – С. 2–6.

Structures of Text Paraphrasing and Plagiarism in Dissertations on Historical Sciences

P.V. Botov, Y.V. Chehovich, A.S. Khritankov, N.S. Surovenko, S.V. Tsarkov, D.V. Viuchnov

We report on the research of structures in graphs of text paraphrasing and plagiarism in Ph.D. dissertations on historical sciences in Russia (07.xx.xx, according to HAC classification). Using algorithmic, statistical and network analysis methods we discovered groups of highly related dissertations, which intensely borrowed from each other, which we call “science shops”, found so-called “compiled” works and probable sources of such compilations.