Экспертная поддержка формирования цифровой коллекции научных работ удостоверенного качества «Научный архив»

Expert support of building the Scientific Archive digital collection of quality scientific papers

Н. В. Авдеева, Т. А. Блинова, И. В. Сусь
Российская государственная библиотека,
Москва, Россия

Nina Avdeeva, Tatyana Blinova and Irina Sus
Russian State Library,
Moscow, Russia

Экспертное сопровождение специалистами Российской государственной библиотеки разработки автоматизированной интеллектуальной информационной системы оценки качества текстовых научных документов – важный этап формирования единой цифровой коллекции научных работ удостоверенного качества, названной «Научный архив».
Ключевые слова: качество, текст, научный, автоматизированное ранжирование, методика, экспертная проверка, верификация, Российская государственная библиотека, Электронная библиотека диссертаций.

Expert support realized by the specialists of the Russian State Library in the development of automatic intellectual information system for evaluating quality of textual scientific documents turned out to be a prominent step towards formation of a united digital collection of scientific works of quality which was entitled «Scientific Archives».
Keywords: quality, text, scientific, computer-aided ranking, methods, expert examination, verification, the Russian State Library, Digital Dissertation Library.

В нынешний век стремительного развития информационных технологий для научного, образовательного и культурного сообществ актуальной становится возможность быстрой, качественной и независимой проверки авторских произведений на плагиат. Обеспечить такую проверку могут специальные программные средства, которые обнаруживают совпадающие фрагменты в огромных массивах документов. Результаты проверки зависят не только от программного обеспечения, но и от качества, объема и актуальности коллекции, с которой идентифицируется документ. Чем больше будет в коллекции первоисточников, чем уникальнее будут наполняющие её документы, тем качественнее и объективнее будут результаты проверки. Следовательно, референтную коллекцию необходимо формировать из заведомо оригинальных документов и пополнять трудами, свободными от плагиата. Создать эталонную коллекцию документов на базе имеющихся огромных электронных архивов можно только с помощью автоматизированной интеллектуальной информационной системы, обладающей аналитическими и самоконтролируемыми свойствами.

Актуальность данного вопроса сформулирована в государственном задании на «разработку информационной системы с целью формирования и структуризации единой цифровой коллекции первоисточников научных работ удостоверенного качества с обеспечением планового пополнения современными научными произведениями и трудами для создания условий публичного доступа к коллекции», полученном ГПНТБ России, Российской государственной библиотекой и ЗАО «Анти-Плагиат» от Министерства образования и науки Российской Федерации.

В рамках этого проекта авторский коллектив ЗАО «Анти-Плагиат» разработал прототип информационной системы обеспечения публичного доступа к единой цифровой коллекции первоисточников научных работ удостоверенного качества, который проходит в настоящее время апробацию. Специалисты федерального государственного бюджетного учреждения «Российская государственная библиотека» (РГБ) были привлечены к сотрудничеству в качестве соисполнителей в части методического сопровождения и экспертной проверки адекватности работы системы, исходя из имеющегося богатого опыта по проведению оценки текстов научных документов на предмет оригинальности.

Сотрудниками РГБ разработаны «Методика определения первоисточников и качества научных трудов» [3] и дополнение к ней – «Методика подробной экспертной проверки научных трудов с низкими рангами» [4].

Необходимость создания последней обусловлена тем, что при первичном автоматизированном ранжировании научных документов теоретически возможно наделение их ошибочными рангами при определении оригинальности текста документа, его уникальности, цитатности и востребованности.

Апробация системы проходит на текстовых документах, сравниваемых с документами коллекции «Электронная библиотека Российской государственной библиотеки» (не менее 1 млн документов при среднем объеме документа 1 МБ). В короткий промежуток времени (до 24 часов) система обрабатывает большой массив документов (до 100 тыс. документов), ранжируя их по качеству, и выдает первичный корректируемый отчет о качестве каждого документа (рис. 1).

Рис. 1. Пример отчета.

«Электронная библиотека диссертаций Российской государственной библиотеки» (ЭБД РГБ) является основной, самой крупной и востребованной, коллекцией Электронной библиотеки РГБ. Решение о создании коллекции ЭБД РГБ на основе современных информационных технологий принято в 2001 году. В 2003 году был оцифрован стартовый пакет диссертаций по наиболее востребованным в то время специальностям: экономические, юридические, педагогические, психологические и философские науки (всего около 28 тыс. полных текстов). Начиная с 2004 года, ЭБД РГБ пополнялась объемом диссертаций по всем специальностям (кроме медицины и фармации), что составляет около 30 тысяч – включая 20 тысяч кандидатских и 10 тысяч докторских – диссертаций в год. В рамках проекта ретроконверсии в 2006 году были оцифрованы все диссертации за 1985 год. А с 2007 года в ЭБД РГБ поступают диссертации по всем дисциплинам, включая работы по медицине и фармации. На данный момент Электронная библиотека диссертаций Российской государственной библиотеки (http://diss.rsl.ru) содержит более 880 тысяч полных текстов диссертаций, защищенных в СССР и Российской Федерации по всем специальностям Высшей аттестационной комиссии Министерства образования и науки Российской Федерации (ВАК Минобрнауки РФ), а также авторефераты к ним [1, 2].

Пополнение коллекции происходит без проверки диссертаций на плагиат, следовательно, для создания эталонной базы (цифровой коллекции первоисточников научных работ удостоверенного качества) требуется проверка и ранжирование всех имеющихся и включаемых текстов и отбор из них оригинальных диссертаций. Таким образом, ЭБД РГБ является идеальной базой цифровых документов для тестирования и обучения разработанной информационной системы.

Для исключения из единой цифровой коллекции научных трудов удостоверенного качества ошибочно попавших документов низких рангов и для дальнейшей адаптации и настройки автоматизированной системы оценки электронных научных документов потребовалась выборочная экспертная проверка сомнительных текстов, а также объективная и доступная статистика о соотношении верно и неверно распознанных программой показателей.

Объектами экспертной проверки являлись документы, содержащие полные тексты научных трудов с разными рангами, присвоенными при первичной автоматизированной проверке, а также

метаданные этих документов. Нетекстовые формы представления содержания (рисунки, графики, таблицы и пр.) экспертами не рассматривались.

При автоматизированном ранжировании данные об уровнях оригинальности, уникальности, цитатности и востребованности (цитируемости) формируются системой на основе автоматического сопоставления текста проверяемого документа с текстами источников, включенными в референтную цифровую коллекцию научных трудов. Под оригинальностью приняли отношение объема незаимствованного из других источников текста и корректно заимствованного к общему объему текста документа; уникальностью – отношение объема текста, не совпадающего ни с какими источниками, к общему объему текста документа; цитатностью – отношение объема корректно заимствованного текста в виде цитат из более ранних произведений того или иного автора к объему всех заимствований; востребованностью (цитируемостью) – отношение объема незаимствованного текста документа, использованного в других (более поздних) произведениях вне зависимости от корректности (правомерности) использования к общему объему текста документа.

В ходе выборочной экспертной проверки научных трудов использовались следующие корректируемые критерии оценки их качества: оригинальность текста, уникальность текста, востребованность текста.

При подробной экспертной проверке происходила верификация данных, полученных системой, путем неавтоматизированного сопоставления частей проверяемого текста с фрагментами текстов источников. Соответствие первичных данных и результатов экспертного анализа оригинальности, уникальности и цитируемости проверенного текста позволило специалисту удостовериться в том, что первичное наделение документа определенным рангом было произведено автоматизированной информационной системой объективно, а при их несоответствии и «ручном» переводе неверно определенных фрагментов в другую категорию происходила автоматическая корректировка отчета системы о ранжировании (рис. 2, 3).

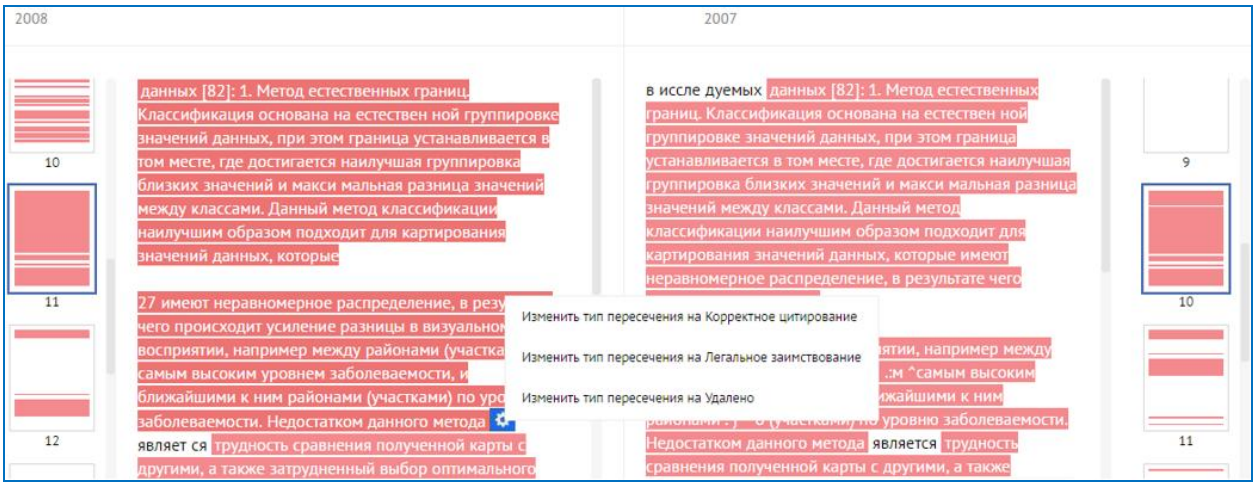

Рис. 2. Сравнение текстовых фрагментов, определенных системой как некорректное заимствование

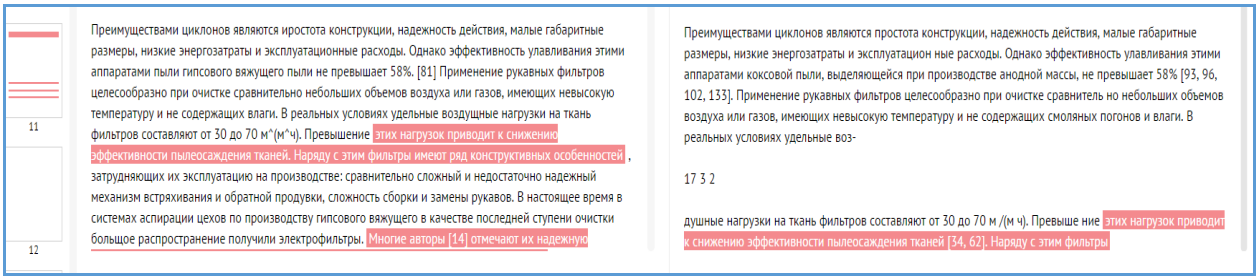

Рис. 3. Сравнение текстовых фрагментов, определенных системой как уникальный текст

В результате первого этапа предварительных испытаний автоматизированной информационной системы ранжирования научных текстовых документов и выборочной проверки документов, наделенных системой определенным рангом, установлено, что система успешно выполняет свою основную функцию – ранжирование научных документов по качеству – и предлагает экспертам возможности для анализа результатов первичного ранжирования, редактирования отчета и сохранения внесенных в него изменений. Несомненным достоинством системы является высокая точность в обнаружении совпадающего текста среди неограниченного массива данных.

В дальнейшем предполагается обучение системы знаниям, накопленным экспертами в процессе работы. Например, при обнаружении дословных совпадений в ранжируемом, более раннем и неопределенном (одного года создания с ранжируемым) документах автоматически переводить неопределенные документы в разряд некорректных источников.

Незначительные недостатки, выявленные при тестировании системы – такие как отсутствие данных об объеме цитатного текста, несоответствие уровней рангов «Методике определения первоисточников и качества научных трудов», идентификация перефразированного текста – в дальнейшем будут скорректированы и настроены. А некоторым моментам (например, сопоставлению цитаты, ссылки на источник с документом из перечня использованной литературы, т.е. выявлению корректно оформленных цитат, отличию корректных заимствований от некорректных) систему еще предстоит научить в процессе дальнейшей работы.

Остается надеяться, что разработанные в рамках Государственного контракта методики и информационная система публичного доступа к единой цифровой коллекции научных работ удостоверенного качества позволят дать комплексную оценку качества создаваемых научных трудов, бороться с плагиатом, повысить результативность научных исследований и исключить их дублирование.

Список использованных источников

1. Авдеева Н.В., Никулина О.В. Независимая экспертиза диссертаций – важный этап на пути повышения качества подготовки научных кадров // Качество образования – № 6, июнь 2014 – С. 16-20. – Режим доступа: http://www.aselibrary.ru/digital_resources/digital_resources69/digital_resources6970/digital_resources69705579/
2. Авдеева Н.В. Исследование российского опыта создания и поддержки полнотекстовых баз данных неопубликованных документов // Информационные системы для научных исследований: Сборник научных статей. Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012). – Санкт-Петербург, октябрь, 2012. – С. 326.
3. Методика определения первоисточников и качества научных трудов [Электронный ресурс] / Государственная публичная научно-техническая библиотека, Российская государственная библиотека, ЗАО «Анти-Плагиат». – Режим доступа: http://www.gpntb.ru/spetsialnye-proekty/metodika-opredeleniya-pervoistochnikov-i-kachestva-nauchnykh-trudov.html
4. Методика подробной экспертной проверки научных трудов с низкими рангами [Электронный ресурс] / Государственная публичная научно-техническая библиотека, Российская государственная библиотека, ЗАО «Анти-Плагиат». – Режим доступа: http://www.gpntb.ru/spetsialnye-proekty/metodika-opredeleniya-pervoistochnikov-i-kachestva-nauchnykh-trudov.html