ОСОБЕННОСТИ АНАЛИЗА СХОДСТВА ДОКУМЕНТОВ В РАЗЛИЧНЫХ КОНТЕКСТАХ ЗАИМСТВОВАНИЯ ПРИ ПОДГОТОВКЕ ТЕКСТОВЫХ МАТЕРИАЛОВ

Иванов В.К., Миронов В.И.

Введение

Согласно [1] плагиат (от лат. plagio - похищаю) – это умышленное присвоение авторства на чужое произведение литературы, науки, искусства, изобретение или рационализаторское предложение (полностью или частично). В этом и в других похожих определениях так или иначе основное значение придается присвоению авторства материала. Постоянно и довольно часто поднимается вопрос о противодействии плагиату и плагиаторам в научных публикациях и квалификационных работах. Пример одной из многочисленных дискуссий можно увидеть в [2]. В целом считается, что наличие признаков плагиата, незаконных или неправильно оформленных заимствований контента негативно сказывается на качестве публикуемых научных, учебных, художественных и других материалов.

В настоящей статье мы попытались сформулировать собственное видение этой проблемы применительно к образовательным технологиям, а точнее – к заимствованиям в студенческих работах и обеспечению их уникальности. В частности, рассматриваются основные контексты заимствования материалов или их частей: типология работ, объекты авторских прав, методологический аспект, среда студенческого коллектива. Также в статье кратко описывается инструментарий, используемый авторами для предотвращения подобных нежелательных явлений, и подводятся некоторые предварительные итоги его применения.

Плагиат и заимствование контента

Плагиат надо отличать от легитимного и правильно оформленного заимствования контента. Так, без согласия автора или правообладателя допускается использование произведения в пределах законодательно установленных исключений. Например, Статья 1274 ГК РФ [3] прямо говорит об этом: "1. Допускается без согласия автора или иного правообладателя и без выплаты вознаграждения, но с обязательным указанием имени автора, произведение которого используется, и источника заимствования: цитирование в оригинале и в переводе в научных, полемических, критических или информационных целях правомерно обнародованных произведений в объеме, оправданном целью цитирования, включая воспроизведение отрывков из газетных и журнальных статей в форме обзоров печати". Подробнее о правилах и особенностях цитирования можно прочитать в обзоре [4]. Но нам важно отметить здесь, что научное цитирование **(**использование в статьях и других материалах фрагментов работ других авторов) - это обязательное требование, без выполнения которого любая научная работа не может считаться таковой.

Различие между плагиатом и правильно оформленным заимствованием контента безусловно осознается и принимается специалистами в рассматриваемой области. Так, разработчики известного Интернет-сервиса "Антиплагиат" [5] четко позиционируют свою технологию, как "Система анализа текстов на наличие заимствований" (но не как систему для обнаружения плагиата).

Является ли заимствование текстов, рисунков, диаграмм, графиков и т.п. безусловным пороком при написании студенческих письменных работ? Ответ на этот вопрос не должен даваться без контекста (совокупности различных факторов,

необходимых для правильного понимания границ разумного использования заимствований).

Заимствование в контексте типологии работ

Типология письменных студенческих работ довольно разнообразна. Основные их виды следующие: отчет о лабораторной работе, письменная контрольная работа, реферат, доклад, обзор, эссе, расчетно-графическое задание, курсовая работа, курсовой проект, научно-исследовательская работа, отчет по практике, выпускная квалификационная работа (выпускная работа бакалавра, дипломная работа или проект специалиста; магистерская диссертация). Выполнение письменной работы осуществляется в ходе аудиторных занятий или студентом самостоятельно. Структура и содержание подобных работ трактуется большинством российских университетов одинаково. Например так, как предложено в [6].

Здесь важно отметить следующее. Некоторые из видов работ (реферат, обзор) прямо предполагают использование заимствований, как части их содержания. Написание работ других видов предусматривает то же самое в их отдельных частях (например, в разделах с анализом научных достижений в какой-либо сфере).

Заимствование в контексте объектов авторских прав

Относительно объектов авторских прав в Статье 1259 ГК РФ [3] говорится следующее: "5. Авторские права не распространяются на идеи, концепции, принципы, методы, процессы, системы, способы, решения технических, организационных или иных задач, открытия, факты, языки программирования". Кроме этого, являются объектами авторских прав официальные документы государственных органов и органов местного самоуправления муниципальных образований, а также сообщения о событиях и фактах, имеющие исключительно информационный характер.

Очевидный вывод, который следует из вышесказанного – многое из того, что является собственно предметом изучения и составляет естественное содержание студенческих работ, не должно считаться не только возможным признаком плагиата, но также не должно рассматриваться и как объект разрешенного заимствования (цитирования). В самом деле, не будет же нарушением законодательства изложение решения задачи кластеризации объектов популярным методом k-средних без цитирования фрагментов работ Гуго Штейнгауза (изобретателя метода).

Заимствование в контексте методологии обучения

Допустим или недопустим тот факт, что студент при подготовке текста курсовой работы использовал абзац из учебника или методического пособия без указания на источник? Авторы склоняются к тому, что допустим, хотя полностью однозначного ответа нет и, более того, возникает множество сопутствующих вопросов.

Каково содержание заимствованного абзаца? Является ли оно описанием идеи, концепции, принципа, метода, процесса, системы, способа и т.п. (см. предыдущий контекст)? Является ли заимствованный абзац в источнике цитатой или это авторский текст? Надо ли менять изложение абзаца, не затрагивая его смысла (например, используя синонимы) для того, чтобы не спровоцировать анализ текста на признаки плагиата? А откуда еще, как не из учебника, брать студенту знания и применять их, например подтверждая полученную квалификацию курсовой работой?

Методологический аспект обучения, касающийся способов доставки студенту образовательного контента, его усвоения, дальнейшего использования, приобретения умений и навыков обязательно должен учитываться в ходе анализа текстов студенческих работ на предмет заимствований.

Заимствование в контексте студенческого коллектива

К сожалению, в студенческой среде существует крайне негативное явление - использование текстов работ сокурсников для сдачи работы под своим именем (мягкий термин - "списывание"). Это можно квалифицировать даже не как плагиат, а как подлог (преступление, заключающееся в подделке подлинных или в составлении фальшивых документов [3]). В этом контексте налицо полное отсутствие позитивного результата для всех участников учебного процесса плюс неизбежные репутационные потери.

Представляется, что с подобным видом заимствования следует бороться в первую очередь.

Предотвращение плагиата

Можно предложить два основных направления борьбы с плагиатом:

1. Профилактика плагиата.

Главное здесь - предотвращение самой возможности нелегитимного заимствования контента, включая заимствования из других студенческих работ. Положительный результат может быть достигнут как формулировкой заданий (например, уникальным содержанием заданий для каждого студента) так и использованием специфических требований к выполнению работы, снижающих вероятность плагиата. В качестве примера приведем часть требований к списку использованной литературы для реферата, которые применяются одним из авторов в своей преподавательской деятельности:

- список должен содержать не менее 5 наименований источников; все источники должны быть изданы за последние 5 лет;
- все использованные источники должны быть доступны либо через фонд научной библиотеки ТвГТУ либо их электронные варианты находятся в свободном доступе в Интернет;
- в случае если использован источник из Интернет, например статья или книга, его описание должно включать прямой адрес на текст этой статьи или книги. Не допускается описание источника вида: www.microsoft.com;
- в списке не должно быть источников, которые сами являются рефератами;
- указание автора обязательно, анонимные материалы не допускаются;
- описания источников, включая источники в Интернет, должны быть оформлены в соответствии с действующими стандартами;
- ссылки на использованную литературу в тексте реферата обязательны.

2. Идентификация плагиата.

Выполняется на множестве (корпусе) уже готовых работ, представленных к проверке и оцениванию. При этом обеспечивается контроль уникальности письменных работ. В качестве примера реализации такого подхода ниже в статье рассматривается технология анализа сходства текстов документов DTA.

Реализация технология анализа сходства текстов документов

С 2010 г. в Центре eScience&Learning Тверского государственного технического университета ведутся работы по реализации программной платформы для анализа документов Document Text Analyzer (DTA) [7, 8, 9]. В настоящее время работы находятся на этапе опытной эксплуатации бета-версий компонентов, выполняющих следующие функции:

- Очистка текста. Выполняет ликвидацию невидимых символов, элементов XML (например, тегов HTML), знаков препинания, цифр, однобуквенных слов.
- Морфологический анализ. Сопоставляет отдельные слова и словоформы в лексиконе, определяет все формы слова и его грамматические характеристики.

- Лемматизация (не стемминг). Анализ лексем и преобразование их в основную форму (лемму). Используются соответствующие словари.

- Фильтрация данных. Применение стоп-листов - списков частых неспецифичных слов, не несущих информации о смысле текста, для сокращения размерности векторного пространства.

- Построение и поддержка модели векторного пространства документов (VSM), включая расчеты метрики TF*IDF, используемой для оценки важности терма документа корпуса. Используются оригинальные алгоритмы авторов.

- Расчет степени сходства документов. Используется классическая косинусная мера близости между векторами.

- Генерация отчетов. Формируется расширяемый набор предопределенных отчетов из базы данных.

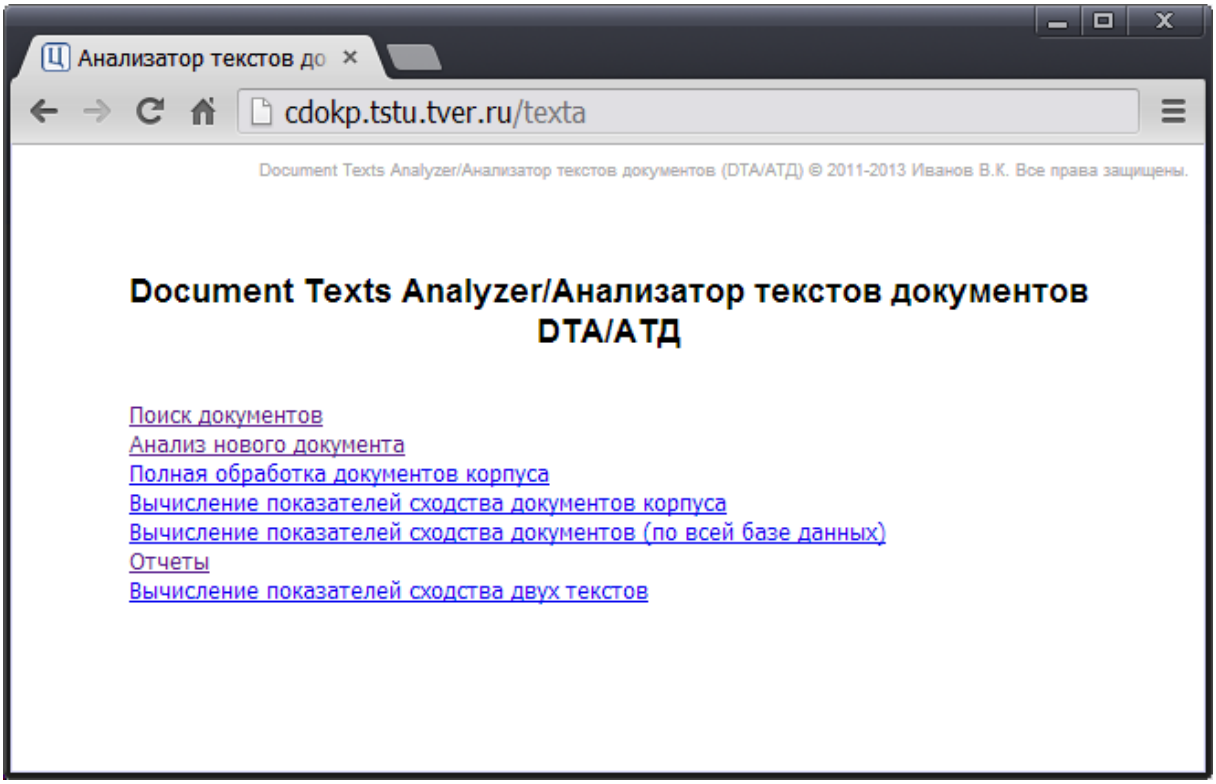

Рис. 1. Главное меню DTA

Весь комплекс программного обеспечения функционирует на портале http://cdokp.tstu.tver.ru в авторизованном режиме.

Опытная эксплуатация программных компонентов разрабатываемой платформы проходит в рамках поддержки технологии работы с электронными учебно-методическими комплексами - важнейшими объектами инфраструктуры учебного процесса [10]. В частности, отрабатываются методы оценки тематической близости документов, которые используются при построении баз данных работ студентов, их автоматической каталогизации, оценки уникальности, а также выполнения других видов интеллектуального анализа текстов.

Некоторые предварительные результаты

Ниже представлены некоторые предварительные результаты применения описанной выше технологии при анализе на заимствования студенческих работ.

Для хранения документов и результатов их обработки и анализа использовалась база данных под управлением MS SQL Server Express. Общий объем – около 2000

документов. Вид анализируемых документов (письменных работ студентов) - рефераты по темам из схожих предметных областей. Период выполнения студентами работ - 2012-2013 гг. (соответствует периоду опытной эксплуатации DTA), Определялось количество пар документов P, рассчитанные меры сходства которых входили в один из шести заданных диапазонов. В качестве показателя, характеризующего диапазон, принималась средняя мера сходства документов $\hat{S}$, входящих в этот диапазон. Результаты представлены на рис. 2.

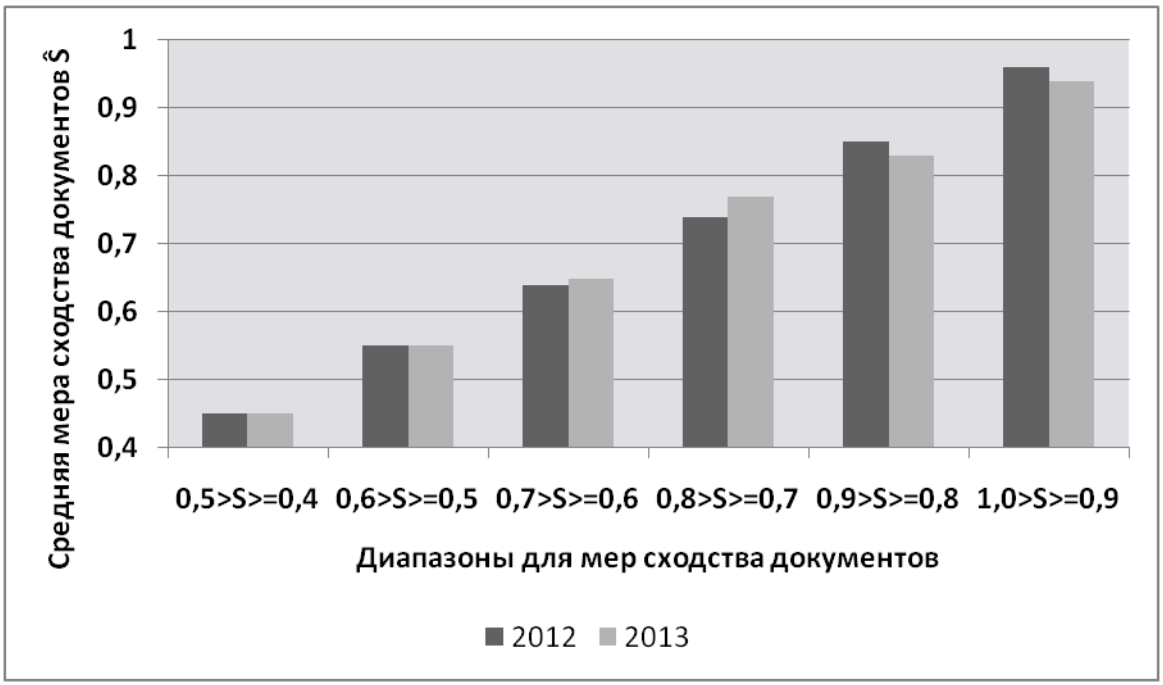

Рис. 2. Показатели сходства документов из экспериментальной выборки

Интересно отметить следующий факт: количество очень похожих документов с потенциально высокой степенью заимствований (мера сходства $S > 0,8$) начала снижаться. То есть, используемые методы контроля могут способствовать повышению уровня уникальности текстов, а следовательно – повышению качества студенческих работ. Конечно, пока еще рано делать окончательные выводы (не накоплена достаточно представительная база анализируемых документов как по количеству, таки по типам), однако предварительные результаты обнадеживают.

Перспективы развития

Предполагается, что технология DTA будет дополнена следующими компонентами:

- Структурный анализ документа. Выявление составных частей и их иерархии, оценка соответствия контента составных частей заданным правилам.
- Поиск ключевых понятий в тексте - именованных семантически значимых сущностей (имена, названия, события, даты, адреса, идентификаторы и т.п.)
- Расчет показателей сложности текста документа: легкости чтения, необходимого уровня образования, индекса туманности.
- Модуль расширение (plugin) для среды электронного обучения Moodle. Реализация алгоритмов DTA для объектов документной базы Moodle.

Работы проводились при финансовой поддержке РФФИ (договор № НК13-07-00342\13).

Библиографический список

1. Большой энциклопедический словарь. - 2-е изд., перераб. и доп. - М.: "Большая Российская энциклопедия"; СПб.: "Нопринт", 1998. - 1456 с. ил.
2. Плагиат в научных исследованиях в области социальных наук : (круглый стол в НИУ ВШЭ) / подгот. к публикации С. Винокур // Вопросы экономики. - 2011. - № 5. - С. 146-151.
3. Гражданский кодекс Российской Федерации от 18.12.2006 N 230-ФЗ : Ч. 4. : (в ред. Федерального закона от 08.12.2011 N 422-ФЗ). - М., 2012.
4. Цитирование, правила и способы [Электронный ресурс] : Интернет-портал Копирайт.ру. - Режим доступа: http://www.copyright.ru/ru/documents/practika/avtoramizdatelyam/tsitirovanie Загл. с экрана.
5. Антиплагиат [Электронный ресурс] : Интернет-сервис. - Режим доступа: http://www.antiplagiat.ru. - Загл. с экрана.
6. Стандарт организации. Студенческие работы: виды, требования к структуре и содержанию : СТО-12-2012 : введ. 2012–02–20 / Липецкий государственный технический университет (ЛГТУ). - Липецк, 2012. – 18 с.
7. Иванов, В.К. Критерии интегральной оценки электронных документов в системах подготовки принятия решений [Электронный ресурс];[Текст]: статья / Тверской гос. техн. ун-т // Вестник Тверского государственного технического университета : науч. рецензир. журнал. - Вып.22. - С. 20-26. - Тверь, 2012.
8. Иванов В.К. Критерии и модели для комплексной оценки качества электронных учебных материалов: статья // Международная интернет-конференция «Информационно-технологическое обеспечение образовательного процесса государств-участников СНГ», 27-30 ноября 2012 г., Минск, БГУ. - Минск, 2012 г.
9. Иванов, В.К. Критерии и модели для мониторинга качества электронных учебных материалов в образовательных технологиях: статья // Научно-практическая конференция «Система гарантий качества образования: разработка и внедрение», 30 октября 2012 г., Тверь, ТвГТУ. - Тверь, 2012.
10. Иванов, В.К. Электронные учебно-методические комплексы как объекты инфраструктуры учебного процесса: статья // Вестник Тверского государственного технического университета / Тверской гос. техн. ун-т. - Вып. 15. - С. 207-211. - Тверь, 2009.