электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.051

Анализ подходов к обнаружению плагиата

***Гаврилова М.А.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»

Научный руководитель: Ерёмин О.Ю., к.т.н., ассистент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
v.suzev@bmstu.ru

Введение

Плагиат (лат. plagiatus - похищенный) - присвоение чужого авторства, выдача чужого произведения или изобретения за свое(словарь иностранных слов). В современной интерпретации плагиат приобрел немного другое значение. И больше подходит определение - это присвоение чужого авторства, идеи или изобретения без ведома автора или по договоренности с ним. В последнее десятилетие плагиат развивается в усиленном темпе. На сегодняшний день это широкая область массовой культуры, ставящая свои "нормы и правила". Как мы знаем, из средств массовой информации, что плагиаторство замечено со стороны общественных, культурных и научных деятелей. Это в свою очередь ставит российскую науку в не очень выгодное положение среди других стран.

Задача поиска плагиата, отлова и наказания пользователей, занимающихся данной деятельностью является первоочередной задачей для российской науки.

Как написано в статье А.Абрамова "Плагиат и антиплагиат", "Плагиат– это растущая раковая опухоль, разъедающая общество и запускающая метастазы во все сферы жизни. Это удар по отечественной науке: дискредитируется звание ученого, разрушаются традиции Служения науке".

Но, к сожалению "лекарства" от этой болезни еще не создали. Но это нужно делать, и чем скорее, тем лучше. Первым шагом в этом направлении это проработка существующих систем, не на российском, так на зарубежном пространстве.

По данным социсследования ГУ-ВШЭ, чаще всего скачивают рефераты, эссе и курсовые студенты четвертых курсов вузов - 52 процента. Реже первокурсники - их 47 процентов. Покупают готовые работы от 3 до 7 процентов студентов[14].

На одной из конференций, посвященной проблеме плагиата были представлены данные результатов американского исследования [17]. Оказалось, что 80% студентов

колледжей признаются, что хотя бы раз в жизни списывали. 36% студентов отмечают, что они списывают регулярно, 90% учащихся уверены, что их плагиат никогда и никем не будет обнаружен. Интересно, что в 1969 году 58% американских школьников давали свои работы для списывания своим соученикам. В 1989 году таковых было 97,5%. Две трети студентов (74%) признаются, что они достаточно регулярно списывали и 47% американских студентов уверены, что преподавателям лучше не замечать факты плагиата в их работах [17]. Это свидетельствует о том, что проблема академической недобросовестности и распространения плагиата в студенческих работах весьма актуальна и требует всестороннего рассмотрения.

Проблема восприятия систем «антиплагиата в ВУЗах»

Сейчас Высшие учебные заведения России переживают очередную волну кампании по борьбе с плагиатом на всех уровнях: от студенческих рефератов до докторских диссертаций.

Еще с 2005 года начал разрабатываться масштабный российский интернет-проект «Антиплагиат»; с 2007 года указанная система была рекомендована для использования в российских вузах. С этого же времени проект используется ВАК Министерства образования и науки РФ [15].

У преподавателей, ученых, руководителей образовательных и научных учреждений появилась возможность при сравнительно небольших временных затратах и минимуме компьютерных знаний осуществлять проверку электронных документов на предмет незаконных заимствований у других авторов (плагиата).

Научные работы, особенно кандидатские и докторские диссертации, сегодня в обязательном порядке должны проходить экспертизу на наличие в них плагиата. Также необходимо отметить, что сегодня только отдельные, самые крупные библиотеки (например, Российская государственная библиотека) оказывают услуги по проведению экспертизы на предмет корректности заимствований. Однако такие экспертные заключения являются платной услугой и поэтому не могут решить проблемы контроля над плагиатом и заимствованиями в науке и образовании[15].

Экспертиза — это то, что должны осуществлять эксперты. В рассматриваемом контексте экспертами являются библиографы. Появление в вузах информационно-библиографических структур, осуществляющих экспертную оценку и поддержку научной деятельности, позволит решить целый комплекс проблем. А со временем такие структуры должны будут стать частью общероссийской экспертной системы.

Но проблема в том, что педагоги боятся, что проверка выявит не только переписанные рефераты и дипломы, но и кандидатские, а то и докторские диссертации.

Не заинтересованы в установке проверочных систем, таких, как "антиплагиат" и "аура", и интернет-торговцы работами на заказ. Если бы они действительно писали для каждого конкретного заказчика!

Если плагиат действительно попадет под запрет, что будут делать создатели сайтов с банками рефератов, курсовых и дипломов? В этой сфере крутятся огромные деньги. Все знают, что этот бизнес вне закона, но количество предложений растет с каждой сессией. Этой зимой на запрос "скачать диплом" Интернет предлагал 93 тысячи ссылок!

Подходы и методы к обнаружению плагиата

Существует несколько подходов к обнаружению заимствований. Наибольшую известность получил метод «шинглов». Метод основан на представлении текстов в виде множества последовательностей фиксированной длины, состоящих из соседних слов. При значительном пересечении таких множеств документы будут похожи друг на друга. Одна из модификаций метода, получившая название «супершинглов», используется для быстрого обнаружения подобных документов [11].

Существует ряд методов, использующих сигнатурную лексическую информацию документов. В работе [4] для этих целей используется I-Match сигнатура, вычисляемая для слов со средним значением IDF (инверсной частоты слов в документах). Другим сигнатурным подходом, основанным на лексических принципах, является метод «опорных» слов . В данном случае для документов составляются по определенным правилам наборы опорных слов, для которых строятся сигнатуры документов. Совпадение сигнатур говорит о подобии самих документов. Эта группа методов, несмотря на большую сложность реализации, показывает более хорошие результаты в обнаружении похожих документов [10,11].

Для обнаружения заимствований иногда используются алгоритмы, построенные на классических принципах информационного поиска, таких как TF, TF*IDF и т.д. [12]. В работе [10] предлагается использовать функцию схожести Джаккарда, применение которой позволяет добиться неплохих результатов даже в текстах с использованием синонимов и наличием орфографических ошибок[14].

Структура систем обнаружения плагиата

Большинство компаний разрабатываемых системы «антиплагиата» не раскрывают своей структуры, для снижения уровня уязвимостей к различным способам обмана(атакам на сервер). Однако детально анализируя структуры открытых систем можно выработать типовую структуру данных систем. Эта структура дает описание, таких лидеров рынка, как «Антиплагиат» (РФ), Turnitin, SafeAssign (США). Типовая структура систем отображена ниже на схеме (рис.1). Пользователь (студент или преподаватель) передает на проверку документ в СОП через информационную систему своего университета либо напрямую через web-интерфейс ситем. Затем содержимое документа преобразуется системой, с целью выделения «чистого» текста, т.е. избавления от форматирования документа присущего современным текстовым процессорам. На основе полученного текста строится запрос к базе данных документов СОП, результатом которого является набор документов, вероятных источников плагиата. Далее, после детального сравнения текстов документов, определяются схожие части и формируется отчёт о найденных совпадениях. В результате, пользователь получает отчёт о проведенной проверке, с указанием частей текста и источников «заимствования», если таковые имелись. При этом база данных документов СОП может содержать индексы открытых сегментов сети Интернет (как в случае с системой Turnitin), так и доступ к некоторым библиотекам с ограниченным доступом.

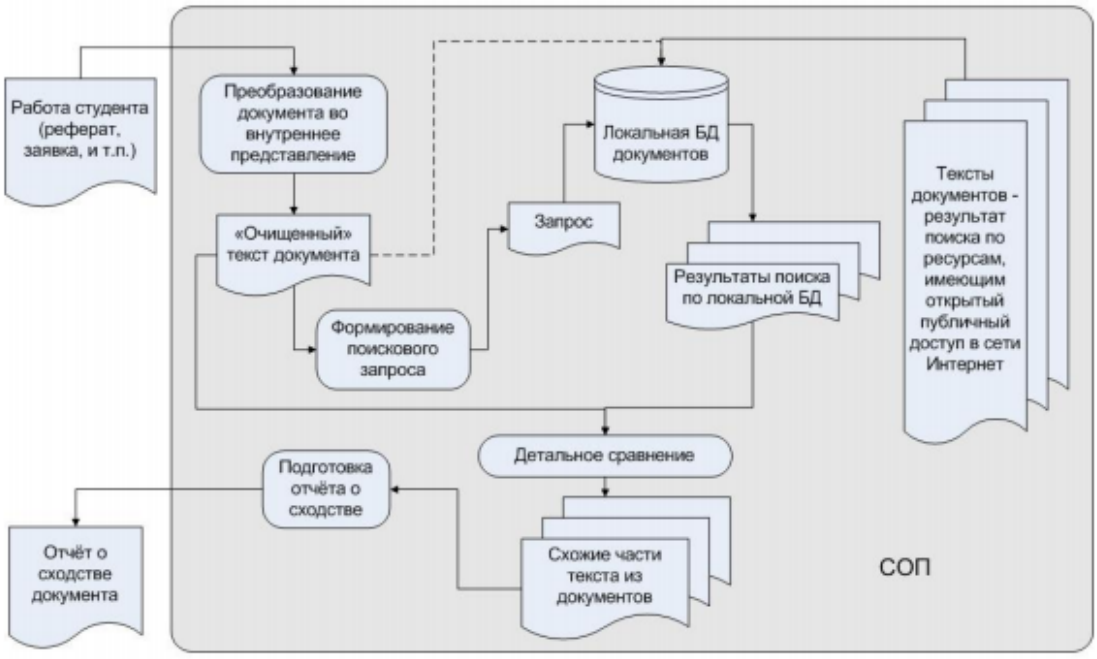

Рис. 1. Типовая структура система обнаружения плагиата

Анализ систем обнаружения плагиата

В настоящее время существует достаточно большое количество сервисов, позволяющих, так или иначе, выявить заимствованный контент.

Достоинства и недостатки рассматриваемых систем приведены в таблице 1.

Большую известность получила система «Антиплагиат», разработанная компанией «Форексис» [8]. Система осуществляет поиск по коллекциям рефератов, контрольных работ и учебников, хранящихся в собственной базе системы.

Программа Advego Plagiatus осуществляет проверку с использованием поисковых систем [1]. Использует разные поисковые системы и проверяет их доступность. Качество обнаружения плагиата - достаточно высокое. Программа выдает процент совпадения текста и выводит найденные источники.

Сервис www.miratools.ru позволяет осуществлять On-line проверку текста на плагиат [6]. Система использует результаты выдачи поисковых систем. По результатам проверки выдается процент совпадений и найденные источники.

Таблица 1

Достоинства и недостатки систем обнаружения плагиата

Система	Достоинства	Недостатки
Advego Plagiatus	Не использует Яндекс.XML	- отсутствие преобразования букв; - отсутствие поддержки поиска по собственной базе; - результаты проверки могут отличатся от раза к разу.
Antiplagiat	-	- система не осуществляет поиск по всем документам, доступным в сети Интернет (особенно на узкоспециализированных ресурсах); - присутствует ограничение размера проверяемого текста 3000 или 5000 символами (доступно после регистрации); - ограничен просмотр документов, частично соответствующих проверяемому тексту; - ограничена возможность проверки по базе имеющихся работ.
Istio	Дополнительные средства для анализа текстов, например, проверку орфографии, анализ наиболее частотных слов и т.д.	- преобразование букв и поддержка поиска по собственной базе отсутствуют.
Miratools	Возможность замены английских букв на русские. Измеряет длину и шаг шинглов (используемых для проверки)	- не работает с собственной базой; - ограничение на длину текста в 3000 символов и на число проверок в течение суток (10 проверок)
Plagiatinform	Возможность обрабатывать документы, скомпонованные из перемешенных кусков текста нескольких источников. Проверка может осуществляться с использованием быстрого или углубленного поиска.	- не предоставляют возможности свободного использования или тестирования системы.
Praide Unique Content Analyser II	Возможность выбора и добавления поисковых систем. Проверка осуществляется пассажами и шинглами, длину которых можно изменять. Возможность задавать количество слов перекрытия шинглов.	- отсутствие замены букв и обработки стоп-слов; - нет поддержки работы с собственной базой.

Сервис www.istio.com осуществляет проверку текста на наличие заимствованного контента с использованием поисковых систем [14]. Для этих целей используют Яндекс.XML и Yahoo.com. Возможности сервиса несколько слабее по сравнению с www.miratools.ru. По результатам проверки выдается сообщение о том, является ли текст уникальным или нет, и выдается список подобных сайтов.

Программа Praide Unique Content Analyser II [14] имеет широкие возможности по проверке текстов с использованием поисковых систем. Выводится подробный отчет по проверке в каждой поисковой системе.

Система Plagiatinform, по заверениям авторов, имеет наиболее широкий функционал [5, 13]. Она умеет проверять документы на наличие заимствований, как в локальной базе, так и в сети Интернет. Результаты проверки выдаются в виде наглядного отчета.

Результаты сравнения функциональности рассмотренных сервисов проверки на плагиат приведены в таблице 2. Несмотря на большое количество существующих решений, ни одно из них не может служить универсальным средством проверки на плагиат. Основной недостаток большинства существующих систем - это направленность поиска либо на сеть Интернет, либо на собственную базу. Очевидно, что более точная и универсальная проверка будет в случае использования обоих видов источников. Кроме того, большинство систем не способно обрабатывать замену букв, чем часто пользуются недобросовестные авторы (чаще всего студенты) [14].

Таблица 2

Сравнение функциональности сервисов проверки текстов на плагиат

Система	Поиск в Интернет	Поиск в локальной базе	Обработка замены букв	Подробный отчет
Advego Plagiatus	+	-	-	+
Antiplagiat	-	+	-	+/-
Istio	+	-	-	-
Miratools	+	-	+	+
Plagiatinform	+	+	-	+
Praide Unique Content Analyser II	+	-	-	+

Большинство рассмотренных систем использует в своей работе метод «шинглов». По исследованиям [9], этот метод демонстрирует высокую точность обнаружения дублированных текстов. Тем не менее, из-за особенностей реализации результаты проверки в каждой системе сильно отличаются от других. Минусом метода является

отсутствие возможности обработки синонимов [10]. Это является значительным недостатком существующих систем. Существует большое количество средств синонимизации текстов. Использование подобных средств может свести на нет работу систем по проверке текстов на плагиат[14].

Таким образом, для эффективного обнаружения плагиата системы должны уметь обрабатывать стоп-слова, осуществлять замену букв с английских на русские и уметь обрабатывать синонимы. Кроме того, в универсальных системах должна быть поддержка поиска как в сети Интернет, так и во внутренней базе. Отчет о проверке должен быть достаточно подробным и содержать сведения о найденных совпадениях с отображением списка источников.

Список литературы

1. Advego Plagiatus - проверка уникальности текста. Режим доступа: http://advego.ru/plagiatus/ (дата обращения 23.06.2014).
2. Broder A. On the resemblance and containment of documents // Compression and Complexity of Sequences (SEQUENCES´97). IEEE Computer Society. 1998. P. 21-29.
3. Kolcz A., Chowdhury A., Alspector J. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization // KDD 2004 (Seattle, Washington, USA, 22-25 August, 2004): proceedings. Seattle, Washington, USA, 2004. P. 605-609.
4. SearchInform Плагиат-Информ - система для определения плагиата в документах. - Режим доступа: http://www.searchinform.ru/main/full-text-search-plagiarism-search-plagiatinform.html (дата обращения 23.08.2014).
5. Сервис проверки уникальности контента. Режим доступа: http:// www.miratools.ru/ (дата обращения 12.09.2014).
6. Анализировать текст, поиск плагиата. Режим доступа: http://istio.com/rus/text/analyz/ (дата обращения 24.10.2014).
7. Антиплагиат. Режим доступа: http://www.antiplagiat.ru/ (дата обращения 24.10.2014).
8. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды 9-й Всероссийской научной конференции RCDL´2007: сб. работ участников конкурса. Переславль-Залесский, 2007. Т. 1. С. 166-174.

9. Неелова Н.В., Сычугов А.А. Сравнение результатов детектирования дублей методом шинглов и методом Джаккарда // Вестник РГРТУ. Рязань, 2010. № 4 (выпуск 34). С. 72-78.
10. Проверка уникальности текста в Интернете - очень полезная программа для качественной раскрутки сайтов. - Режим доступа: http://www.nado.su/downloads.html (дата обращения 24.10.2014)
11. Шарапов Р.В., Шарапова Е.В. Пути расширения булевой модели поиска // Информационные системы и технологии // Известия Орел ГТУ. 2009. № 6 (56). С. 74-78.
12. Ширяев М.А., Мустакимов В. Plagiatinform избавит от плагиата в научных работах // Educational Technology & Society. 2008. №11(1). С. 367–374.
13. Шарапов Р.В. Анализ подходов к обнаружению заимствованных текстов // Современные наукоемкие технологии. 2011. № 3. С. 47-49.
14. Раицкая Л. Плагиат vs. заимствование. Режим доступа: http://www.mgimo.ru/news/experts/document240689.phtml (дата обращения 10.11.2014).
15. Кичерова И.Н., Кыров Д.Н., Смыкова П.Н., Пилипушка С.А. Плагиат в студенческих работах: анализ сущности проблемы. Режим доступа: http://naukovedenie.ru/PDF/83pvn413.pdf (дата обращения 10.11.2014).