УДК 004.42

АНАЛИЗ ПОДХОДОВ К ОБНАРУЖЕНИЮ ЗАИМСТВОВАННЫХ ТЕКСТОВ

Шарапов Р.В.

Муромский институт ГОУ ВПО «Владимирский государственный университет», Муром, e-mail: mivlgu@mail.ru

В работе проводится анализ существующих подходов к обнаружению заимствования текстов. Даются краткие характеристики существующих систем проверки на плагиат и приводится сравнение их функциональности.

Ключевые слова: проверка на плагиат, плагиат, заимствование, текст

ANALYSIS OF THE APPROACHES TO THE DETECTION OF PLAGIARISM

Sharapov R.V.

Murom Institute of Vladimir State University, Murom, e-mail: mivlgu@mail.ru

In this paper we analyze existing approaches to the detection of plagiarism. We give a brief description of existing systems for plagiarism detection and compare their functionality.

Keywords: detection of plagiarism, plagiarism, text

Современное развитие информационных технологий и сети Интернет предоставило широким кругам пользователей доступ к огромным массивам информации. Это вызвало бурный рост количества дублированной и заимствованной информации. Особенно явно это заметно в самой сети Интернет, в сфере образования и средствах массовой информации. Иногда заимствованный контент встречается и в научных кругах. По этой причине задача обнаружения подобных заимствований (фактов плагиата) приобретает повышенную актуальность.

Существует несколько подходов к обнаружению заимствований (которые иногда называют нечеткими дублями). Наибольшую известность получил метод «шинглов» [2]. 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 [9].

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 [3]. 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 [9].

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.

Рассмотрим теперь практическое использование описанных подходов в задачах обнаружения плагиата. В настоящее время существует достаточно большое количество сервисов, позволяющих, так или иначе, выявить заимствованный контент. Большую известность получила система «Антиплагиат», разработанная компанией «Форексис» [8]. Система осуществляет поиск по большому количеству коллекций рефератов, контрольных работ и учебников, хранящихся в собственной базе системы. Тем не менее, система имеет ряд недостатков. Во-первых, система не осуществляет поиск по всем документам, доступным в сети Интернет. Особенно это касается тематических сайтов и новостных порталов – большое число заимствований осуществляется именно с таких источников. Соответственно даже при полном дублировании подобной информации система «Антиплагиат» соответствий не обнаружит. Во-вторых, присутствует ограничение размера проверяемого текста 3000 или 5000 символами (доступно после регистрации). В-третьих, ограничен просмотр документов, частично соответствующих проверяемому тексту. Кроме того, система ограничивает возможность проверки по базе имеющихся работ.

Программа Advego Plagiatus осуществляет проверку с использованием поисковых систем [1]. Использует разные поисковые системы и проверяет их доступность. В отличие от аналогичных систем, Advego Plagiatus не использует Яндекс.XML. Качество обнаружения плагиата – достаточно высокое. Программа выдает процент совпадения текста и выводит найденные источники. Недостатком является отсутствие преобразования букв, отсутствие поддержки поиска по собственной базе. Из-за особенностей работы программы возникают ситуации, когда результаты проверки отличаются от раза к разу.

Сервис www.miratools.ru позволяет осуществлять On-line проверку текста на плагиат [6]. Система использует результаты выдачи поисковых систем. К достоинствам можно отнести возможность замены английских букв на русские. Имеются возможности изменять длину и шаг шинглов (используемых для проверки). По результатам проверки выдается процент совпадений и найденные источники. Система не работает с собственной базой. Присутствует ограничение на длину текста в 3000 символов и на число проверок в течение суток (10 проверок).

Сервис www.istio.com осуществляет проверку текста на наличие заимствованного контента с использованием поисковых систем [7]. 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. Преобразование букв и поддержка поиска по собственной базе отсутствуют. Сервис предоставляет дополнительные средства для анализа текстов, например, проверку орфографии, анализ наиболее частотных слов и т.д.

Программа Praide Unique Content Analyser II [11] имеет широкие возможности по проверке текстов с использованием поисковых систем. Имеется возможность выбора используемых поисковых систем, содержатся средства добавления новых поисковых систем. Проверка осуществляется пассажами и шинглами, длину которых можно изменять. Можно задавать количества слов перекрытия шинглов. Выводится подробный отчет по проверке в каждой поисковой системе. К недостаткам можно отнести отсутствие замены букв и обработки стоп-слов. Нет поддержки работы с собственной базой.

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. Система умеет обрабатывать документы, скомпонованные из перемешенных кусков текста нескольких источников. Проверка может осуществляться с использованием быстрого или углубленного поиска. Результаты проверки выдаются в виде наглядного отчета. Авторы не предоставляют возможности свободного использования или тестирования системы, и оценить качества ее работы невозможно.

Результаты сравнения функциональности рассмотренных сервисов проверки на плагиат приведены в таблице. 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).

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.

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 [Электронный ресурс]. – Режим доступа: http://advego.ru/plagiatus/ (дата обращения: 23.03.2011).

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. Ilyinsky S., Kuzmin M., Melkov A., Segalovich I. An efficient method to detect duplicates of Web documents with the use of inverted index // WWW Conference 2002.

4. Kolcz A., Chowdhury A., Alspector J. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization // KDD 2004, 22-25 August, 2004, Seattle, Washington, USA.

5. SearchInform Плагиат-Информ – система для определения плагиата в документах [Электронный ресурс]. – Режим доступа: http://www.searchinform.ru/main/full-text-search-plagiarism-search-plagiatinform.html (дата обращения: 23.03.2011).

6. www.miratools.ru – Сервис проверки уникальности контента [Электронный ресурс]. – Режим доступа: http://www.miratools.ru/ (дата обращения: 23.03.2011).

7. Анализировать текст, поиск плагиата | istio.com [Электронный ресурс]. – Режим доступа: http://istio.com/rus/text/analyz/ (дата обращения: 23.03.2011).

8. Антиплагиат [Электронный ресурс]. – Режим доступа: http://www.antiplagiat.ru/ (дата обращения: 23.03.2011).

9. 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.

10. 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.

11. Проверка уникальности текста в Интернете – очень полезная программа для качественной раскрутки сайтов [Электронный ресурс]. – Режим доступа: http://www.nado.su/downloads.html (дата обращения: 23.03.2011)

12. Шарапов Р.В., Шарапова Е.В. Пути расширения булевой модели поиска // Информационные системы и технологии // Известия Орел ГТУ. – Орел: ОрелГТУ, 2009. – №6(56) . – С. 74–78.

13. Ширяев М.А., Мустакимов В. Plagiatinform избавит от плагиата в научных работах // Educational Technology & Society. – 2008. – №11(1). – С. 367–374.