УДК 681.3.07

INTERNET-ПРИЛОЖЕНИЕ ДЛЯ ПРОВЕРКИ ПЛАГИАТА

Завадский М.А., Шейна М.П.
Научный руководитель – Разоренов Н.А., доцент.

Плагиат — это заимствование чужой работы, будь то преднамеренно или непреднамеренно, как своей, ради собственной выгоды. Так же это использование, перефразирование и подведение итогов работы в любой форме без подтверждения ссылками на источники и представление её как своей собственной работы.

Плагиат с появлением Интернета превратился в серьёзную проблему. Попав в Интернет, знание становится достоянием всех, соблюдать авторское право становится всё труднее и даже невозможно. Постепенно становится сложнее определить первоначального автора.

В связи с участившимися случаями выдачи чужих работ за свои исследуется данная тема, целью которой является уменьшение процента присвоения и выдачи чужих идей и материалов за свои.

Сравнительная характеристика

Существует множество систем, занимающихся проверкой и выявлением плагиата в документах, однако большинство из них имеет массу недостатков.

Ниже представлена таблица сравнения функциональных возможностей сервисов контроля текстов на плагиат**.**

Система	Поиск в Интернете	Поиск в локальных базах	Применяемые типы файлов	Количество языков
eTXT Антиплагиат	+	-	.doc, .txt	неограни-ченно
Advego Plagiatus	+	-	.txt	неограни-ченно
Anti-Plagiarism	+	+	.rtf, .doc,*.docx, .pdf	неограни-ченно
Double Content Finder			.txt	русский
Viper			.doc, .docx, .pdf, .html, .odt, .rtf, .text, .s, .cs, .app, .java, .ptt, .pttx.	английский
Плагиата НЕТ	+	-	.doc, .docx, .rtf, .txt	неограни-ченно
Anti-Plagiarism	+	-	.rtf, .doc,*.docx, .pdf	неограни-ченно
Plagiat-inform	+	+	.doc, .txt	неограни-ченно
Praide Unique Content Analyser	+	-	.doc, .txt	неограни-ченно
Автор.NET	+	+	.doc, .txt	неограни-ченно

Главным недостатком большинства систем анти плагиата является отсутствие возможности детального лингвистического анализа текстов на естественном языке, который включает морфологический, синтаксический, семантический и прагматический виды анализа.

Также существенным недостатком большинства систем является отсутствие возможностей совместного проведения проверок в интернете и локальной базе. Большинство сервисов рассчитаны на выполнение проверки путем поиска в одной области, или имеют ограничение в выборе проверяемых документов. Такие сервисы чаще всего используют для проверки метод «шинглов», который в свою очередь можно достаточно легко обмануть путем изменения построения предложений в тексте, заменой букв на аналогичные из других языков и иными возможностями выдачи чужой работы за авторскую.

Алгоритм шинглов

Этапы

Этапы, которые проходит текст, подвергшийся сравнению:

- канонизация текста;
- разбиение на шинглы;
- вычисление хэшей шинглов;
- случайная выборка 84 значений контрольных сумм;
- сравнение, определение результата.

Канонизация текста

Канонизация текста приводит оригинальный текст к единой нормальной форме. Текст очищается от предлогов, союзов, знаков препинания, HTML тегов, и прочего ненужного «мусора», который не должен участвовать в сравнении. В большинстве случаев также предлагается удалять из текста прилагательные, так как они не несут смысловой нагрузки.

Также на этапе канонизации текста можно приводить существительные к именительному падежу, единственному числу, либо оставлять от них только корни.

Разбиение на шинглы

Шинглы — выделенные из статьи подпоследовательности слов. Необходимо из сравниваемых текстов выделить подпоследовательности слов, идущих друг за другом по 10 штук (длина шингла). Выборка происходит внахлест, а не встык. Таким образом, разбивая текст на подпоследовательности, мы получим набор шинглов в количестве равному количеству слов минус длина шингла плюс один (кол_во_слов — длина_шингла + 1).

Вычисление хэшей шинглов

Принцип алгоритма шинглов заключается в сравнении случайной выборки контрольных сумм шинглов (подпоследовательностей) двух текстов между собой.

Проблема алгоритма заключается в количестве сравнений, ведь это напрямую отражается на производительности. Увеличение количества шинглов для сравнения характеризуется экспоненциальным ростом операций, что критически отразится на производительности.

Во избежание приравнивания к плагиату участков документа с указанной ссылкой на первоисточник, будет проводится проверка соответствующего участка с источником и при совпадении этот участок не будет считаться плагиатом.

Заключение

Реализация данных алгоритмов поможет уменьшить количество случаев выдачи чужих материалов за свои. Так же данные идеи помогут расширить функциональные возможности стандартного алгоритма проверки методом «шинглов», повысить эффективность его работы.

Литература

1 https://ru.wikipedia.org/wiki/Выявление_плагиата
2 Философские проблемы информационных технологий и киберпространства. «Метод семантического сравнения нечеткой информации при проверке текстов на наличие плагиата» – Корманицкая О.И, Корманицкая И.И. – декабрь, 2015
3 https://ru.wikipedia.org/wiki/Алгоритм_шинглов