Е.В. Шарапова
Муромский институт (филиал) Владимирского государственного университета
Россия, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: sharapovamivlgu@gmail.com

О проблеме обнаружения нечетких дубликатов текстов*

Задача нахождения полных дубликатов текстов решается достаточно просто. Например, можно использовать контрольные суммы или хеш-коды, подсчитанные по всему тексту и сравнивать их с аналогичными кодами других текстов. Гораздо сложнее обстоит дело с поиском частично совпадающих текстов (которые иногда называют нечеткими дубликатами). Существует несколько подходов к обнаружению нечетких дубликатов. Наибольшую известность получил метод «шинглов» [1]. 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 [2].

Существует ряд методов, использующих сигнатурную лексическую информацию документов. В [3] 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 [4]. 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 [2].

Для обнаружения нечетких дубликатов иногда используются алгоритмы, построенные на классических принципах информационного поиска, таких как TF, TF*IDF и т.д. В [5] предлагается использовать функцию схожести Джаккарда, применение которой позволяет добиться неплохих результатов даже в текстах с использованием синонимов и наличием орфографических ошибок. Для нахождения нечетких дубликатов могут использоваться алгоритмы, построенные на основе суффиксных деревьев, N-грамм и т.д.

Рассмотрим теперь практическое использование описанных подходов в задачах обнаружения нечетких дубликатов. В настоящее время существует достаточно большое количество сервисов, позволяющих, так или иначе, выявить дублированный (заимствованный) контент. Большую известность получила система «Антиплагиат» [http://www.antiplagiat.ru/]. 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 [6]. Из-за особенностей архитектуры даже после проведения модификаций система «Антиплагиат» не сможет обеспечить эффективный поиск по источникам в сети Интернет.

Программа Advego Plagiatus осуществляет проверку с использованием поисковых систем [http://advego.ru/plagiatus/]. 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. Недостатками 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 [7]. Сервис

www.miratools.ru позволяет осуществлять On-line проверку текста на плагиат [http:// www.miratools.ru]. 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.miratools.ru позволяет осуществлять On-line проверку текста на плагиат [http:// www.miratools.ru]. 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).

Несмотря на большое количество существующих решений, ни одно из них не может служить универсальным средством обнаружения нечетких дубликатов. Основные недостатки большинства существующих подходов:

1. 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.

2. большинство систем не способны обходить существующие подходы к сокрытию следов заимствований (обрабатывать замену букв, убирать знаки переносов, изменение окончаний и т.д.).

3. область поиска нечетких дубликатов ограничивается чаще всего небольшими текстами в несколько тысяч слов. Системы не адаптированы для работы с большими текстами. Большинство рассмотренных систем используют в своей работе метод «шинглов».

По исследованиям [2] 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, так как существует большое количество средств синонимизации текстов. Это является значительным недостатком существующих систем.

Литература

1. Broder A. On the resemblance and containment of documents // Compression and Complexity of Sequences (SEQUENCES'97). IEEE Computer Society, 1998. P. 21-29

2. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2007: сб. работ участников конкурса –Переславль-Залесский, 2007. – Т. 1. – С. 166-174]

3. 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

4. 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

5. Неелова Н.В., Сычугов А.А. Сравнение результатов детектирования дублей методом шинглов и методом Джаккарда // Вестник РГРТУ. № 4 (выпуск 34). Рязань, 2010 – с. 72-78

6. Шарапова Е.В. Исследование возможностей системы «Антиплагиат» для обнаружения заимствований // Перспективы науки и образования, №3, 2013, с. 215-219

7. Sharapova E.V. Analysis of methods and systems for fuzzy duplicate detection // 14 International multidisciplinary scientific Geoconference SGEM2014. Informatics, Geoinformatics and Remote Sensing. Conference proceedings. 17-26 June 2014, Albena, Bulgaria, 2014. Vol. 1. P. 27-33

*** Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00692**