Болотин Е.И.

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ ИДЕНТИФИКАЦИИ ДЛЯ ВЫЯВЛЕНИЯ ОБЩИХ ФРАГМЕНТОВ ТЕКСТОВЫХ МАТЕРИАЛОВ

nv-maks@yandex.ru
Национальный исследовательский ядерный университет «МИФИ» (НИЯУ МИФИ)
г. Москва

Формирование распределенной информационно-образовательной среды исследовательского университета обуславливает интерес к теории и практике разработки технологий агрегирования, консолидации материалов различных форматов, технологий поиска, репозитарного хранения и поддержки регламентов редактирования, хранения версий, анализа текстов, в том числе фрагментов технических и технологических материалов [1].

Поэтому применение методов обнаружения документов, содержащих заданные фрагменты текста, обнаружения дубликатов текстов в документальных базах данных, выявления заимствований и плагиата в различных видах научных текстов является актуальным, и представляет собой частный случай общей задачи идентификации документальной информации [2].

Методы идентификации позволяют сопоставить документу идентификатор, отражающий свойства информации, содержащейся в документе. Поскольку централизованное хранилище данных масштаба университета содержит подсистемы различных размеров, на начальном этапе разработки технологий поиска необходим сравнительный анализ уровня сложности применяемого алгоритма в зависимости от размера базы данных.

В данной работе сделан вывод о перспективах применения метода идентификации содержания документов в виде совокупности сверток частей текста [4]. Показано, что использование алгоритма нормализации необходимо на начальных стадиях анализа, этот алгоритм позволяет уменьшить влияние таких показателей, как форматирование текста, синтаксис, разбиение на фрагменты, страницы. На следующем этапе необходимо выполнить подпроцесс разбиения текста на шинглы (с использованием CRC32), а также подсчет общей контрольной суммы текста (с использованием MD5). На следующем этапе рекомендуется выполнить проверку данного фрагмента с помощью алгоритма выявления полных дубликатов. После определение наиболее похожего документа по шинглам и выявления общих шинглов и связанных с ними блоков текста принимается решение о занесении нового документа в базу.

Для выявления точных дубликатов в некотором множестве документов используется алгоритм, основанный на вычислении контрольных сумм документов при помощи хеширования. При использовании предварительной нормализации текстов данный метод может выявлять не только точные дубликаты, но и варианты документов, нормализованный текст которых идентичен. Этот подход особенно актуален для справочной, технологической информации, а также для управляющих и нормативных документов, например, версий бюджета, табличных данных, моделей процессов деятельности университета.

Проведен эксперимент по поиску нечетких дубликатов документа, то есть тех вариантов, которые представляют собой документы, частично измененные в содержательной части и/или в части форматирования, в котором использовались последовательности из пяти слов. Выполнен анализ метода отбора шинглов по признаку делимости.

Проведенный в работе анализ показывает, что высокая производительность алгоритма позволяет использовать его в Web-ориентированной распределенной информационной среде исследовательского университета для фильтрации спама, выявления плагиата, поисковых машинах для оптимизации поисковой выборки.

Литература

1. Максимов Н.В. Информационная среда науки и образования: от информационного обслуживания к распределенной системе управления знаниями. Информационное общество. № 6 – 2009. – С. 58-67.
2. Болотин Е.И. Разработка методики мультиагентной идентификации содержания документов. Научная сессия МИФИ-2010. Сборник научных трудов. - М.: МИФИ, 2010. - XIV Выставка-конференция "Телекоммуникации и новые информационные технологии в образовании". - С. 127-129.
3. A. Broder, S. Glassman, M. Manasse and G. Zweig. Syntactic clustering of the Web. Proc. of the 6th International World Wide Web Conference, April 1997.