УДК 004.78:025.4.036

ЛОКАЛЬНЫЕ ПАРАМЕТРЫ ТЕКСТОВ И ПРОБЛЕМА ОПРЕДЕЛЕНИЯ ПОЧТИ-ДУБЛИКАТОВ*

Д. И. Косинов

Воронежский государственный университет

Рассматривается метод построения сигнатуры документа на основании исключительно локальных количественных параметров его содержимого. Набор параметров подбирается исходя из соображений устойчивости к различным видам модификаций документа. Проведен ряд экспериментов, использующих некоторые из этих параметров. Показана возможность использования данного подхода в условиях больших объемов документов.

ВВЕДЕНИЕ

Одной из задач, с которой сталкивается любая информационно-поисковая машина, является задача определения схожести различных документов между собой. Выявление дубликатов позволяет устранять повторяющиеся документы в списках-результатах запросов, уменьшать размеры индекса путем устранения избыточности, обнаруживать плагиат и распознавать массовые почтовые рассылки (спам).

Причины модификаций документов в сети могут быть самыми различными, например: создание зеркал сайтов, преобразования документов в другой формат или его редактирование. Отдельным пунктом идут намеренные искажения текстов, применяемые спамерами в массовых рассылках.

Прямое решение путем попарного сравнения текстов документов, в условиях гигантских объемов данных в Интернете, не представляется возможным. Применяются различные методы снижения вычислительной сложности за счет выбора различных эвристик: хеширования фиксированного набора значимых слов, сэмплирование цепочек элементов текста, использование дактилограмм и т.д. Затем полученные хеши (отпечатки документов) сравниваются, и документы считаются схожими (также употребляется термин «почти-дубликаты»), если доля совпавших отпечатков превышает некий порог.

Перед созданием отпечатка документ обычно проходит через ряд упрощающих преобразований: удаляются HTML-разметка, лишние пробелы, пунктуация, стоп-слова, символы приводятся к единому регистру, производится стемминг (выделение значащей части слов).

Логичным способом ускорения сравнения становится переход от множества отпечатков, идентифицирующих документ, к единому. Это приводит к заметному «огрублению» результатов, т. к. степень схожести определить уже не удается. Однако высокая производительность, особенно вместе с использованием исключительно «локальной» информации о документе (без опоры на глобальные коллекции значимых термов, например), в ряде задач дает преимущества.

1. МЕТОД ПОСТРОЕНИЯ СИГНАТУРЫ ДОКУМЕНТА НА ОСНОВАНИИ ЛОКАЛЬНЫХ КОЛИЧЕСТВЕННЫХ ПАРАМЕТРОВ ЕГО СОДЕРЖИМОГО

В работе [1] автором был предложен метод построения сигнатуры документа, строящийся на основе определенного набора статистических параметров документа, подобранных исходя из соображений устойчивости к определенным формам модификаций документа. Например, количество точек, запятых и заглавных букв в тексте позволяет примерно зафиксировать количество предложений в тексте; общая длина текста в символах с исключением пробелов и стоп-слов дает общую оценку объема; оценка количества и отношения разночастотных слов может служить некой заменой семантическому анализу документа и т.д.

Частично вопросы построения локальной статистической сигнатуры документа рассматривались в [2, 3]. Данное исследование проводилось с целью сравнения результативности предложенных подходов, а также для рассмотрения возможности использования набо-

* Данная работа поддержана исследовательским грантом компании Яндекс в рамках конкурса «Интернет-Математика 2007»

ра спецсимволов документа при создании сигнатуры.

При проведении эксперимента использовался набор данных, включающий порядка 750 тысяч страниц, содержащихся более чем в 23 000 сайтов домена narod.ru. Данные о дубликатах этого набора данных составлялись на основе функции Левенштайна. «Релевантными» считались пары документов, состоящие не менее, чем из 20 слов, и коэффициент сходства которых, рассчитанный функцией Perl String:: Similarity, составлял не менее 85 % после удаления тэгов.

Был проведен ряд экспериментов, исследующих возможность использования параметров текста, подобных вышеописанным, при генерации подписей документов. Во всех тестах предварительно проводилось удаление HTML-тэгов.

Тест № 1: спецсимволы (количество спецсимволов и знаков препинания).

Подсчитывалось число вхождений в текст документа каждого символа из следующего набора: . , — _ : ; ! ? ( ) и *символ пробела*. Документ представлялся в виде вектора размерностью 11 элементов, компонентами которого являлось число вхождений каждого символа в документ.

Тест № 2: последовательности спецсимволов.

Из текста документа удалялись буквенно-цифровые символы, оставляя спецсимволы, пробелы и переводы строк.

Тест № 3: длины элементов (количество и средняя длина слов и предложений).

Подсчитывалась средняя длина слова и предложения в документе, а также общее количество тех и других. Сигнатура составлялась путем конкатенации полученных значений строкой вида: «[средняя длина слова] — [средняя длина предложения] — [число слов] — [общее число предложений]».

Тест № 4: два самых длинных предложения.

Из текста документа выделялись два самых длинных предложения и сцеплялись друг с другом. Тест проводился для сравнения результатов с аналогичным тестом из работы Ю. Г. Зеленкова и И. В. Сегаловича [2], показавшим там наилучшие результаты.

Тест № 5: хеш текста.

Над текстом документа вычислялась хеш-функция MD5. Тест служит для предоставления базовых результатов определения дубликатов.

В качестве метрик использован следующий, во многом стандартный набор:

— Полнота — отношение общего числа найденных «релевантных» пар дубликатов a к общему числу «релевантных» пар: $P = \frac{a}{a+b}$, где b — количество найденных пар дубликатов, не совпадающих с «релевантными» парами.

— Точность — отношение общего числа найденных пар дубликатов a к общему числу найденных пар: $R = \frac{a}{a+c}$, где c — количество не найденных пар дубликатов, совпадающих с «релевантными» парами.

— F-мера — гармоническое среднее полноты и точности: $F = \frac{2PR}{P+R}$.

Результаты тестов приведены на рис. 1—3.

Наиболее интересные результаты были получены в тестах № 1; 2. Они оба показывают

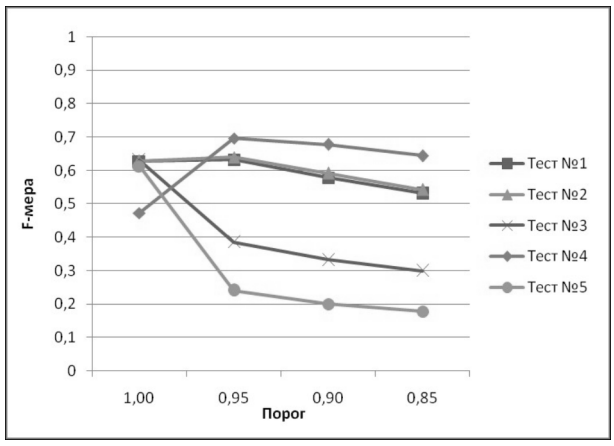

Рис. 1. Показатель F-меры для различных тестов

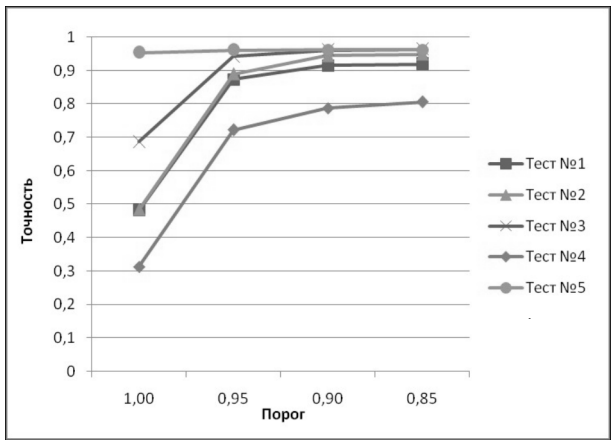

Рис. 2. Показатели точности для различных тестов

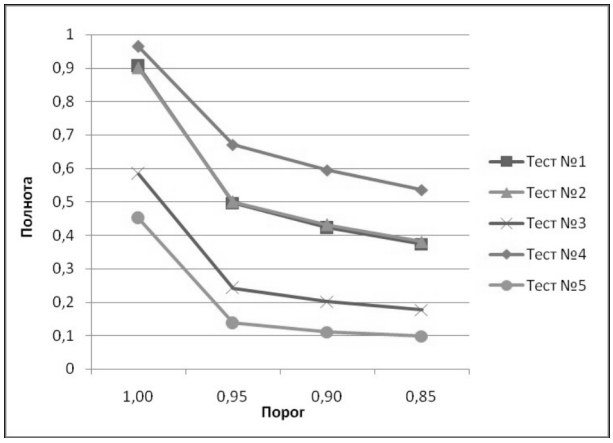

Рис. 3. Показатели полноты для различных тестов

практически одинаковые результаты, давая много «лишних» пар дубликатов при пороге, равном 1, и занимающих свое место при уменьшении порога до 0,9. F-мера всех порогов данного теста оказалась максимальной среди всего набора. В ходе экспериментов с модификациями наборов спецсимволов было установлено, что количество получаемых в результате пар дубликатов можно варьировать в широких пределах, добиваясь нужного баланса полноты и точности. Однако их совокупность, выраженная F-мерой, не превысит 60 % уже при пороге схожести в 0,9.

Модификации теста № 3, основанные на количестве слов различных длин и конкатенации длин всей последовательности слов, показывали практически аналогичные ему результаты. Тесты, использующие по отдельности слова и предложения (среднюю длину и общее количество), показали на 20—25 % худшую точность при сохранении той же полноты. В целом можно сделать вывод, что сигнатуры документа, основанные на длинах различных элементов текста, не позволяют удерживать полноту на должном уровне. Кроме того, было установлено, что ориентация на заглавные буквы (их последовательности и количество), равно как и опора на общую длину текста (с удалением стоп-слов и пунктуации и без) также не дает сколько-нибудь значимых и интересных результатов.

ЗАКЛЮЧЕНИЕ

В целом результаты тестов можно считать удовлетворительными. Невысокие показатели на низких порогах схожести документов (в сравнении, например, с работой Ю. Г. Зеленкова и И. В. Сегаловича [2]) проистекают из «бинарности» суждений, обусловленных единой сигнатурой, и хороших показателях на высоких порогах. «Смягчением» отпечатка можно добиться увеличения количества обнаруженных пар дубликатов (как в тесте № 4), что позволит сместить эффективную границу на более низкий уровень.

Для дальнейших исследований представляет интерес апробация теста, основанного на спецсимволах, к различным классам документов. При условии корректной предварительной классификации документа возможна подстройка набора параметров и весов алгоритма создания отпечатка, что может позволить увеличивать результативность определения дубликатов в конкретном классе.

СПИСОК ЛИТЕРАТУРЫ

1. *Косинов, Д.И.* Использование статистической информации при выявлении схожих документов / Д. И. Косинов // Интернет-математика 2007 : сборник работ участников конкурса. — Екатеринбург: Изд-во Урал. ун-та, 2007. — С. 84—91.

2. *Зеленков, Ю.Г.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов / Ю. Г. Зеленков, И. В. Сегалович // труды 9 Всерос. науч. конф.: Электронные библиотеки: перспективные методы и технологии, электронные коллекции — RCDL'2007. Переславль-Залесский, Россия, 2007. — С. 166—174.

3. *Krovetz R.* Analysis of Lexical Signatures for Finding Lost or Related Documents / S. Park, D. M. Pennock, C. Lee Giles, R. Krovetz // Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. — 2002. — P. 11—18.