УДК 004.912

Ю. А. Загорулько, **Н. В. Саломатина**, **А. С. Серый**
Е. А. Сидорова, **В. К. Шестаков**

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

E-mail: zagor@iis.nsk.su; nataly@math.nsc.ru
Alexey.Seryj@iis.nsk.su; lena@iis.nsk.su
shestakov@iis.nsk.su

ВЫЯВЛЕНИЕ НЕЧЕТКИХ ДУБЛИКАТОВ ПРИ АВТОМАТИЧЕСКОМ ФОРМИРОВАНИИ ТЕМАТИЧЕСКИХ КОЛЛЕКЦИЙ ДОКУМЕНТОВ НА ОСНОВЕ WEB-ПУБЛИКАЦИЙ [*]

Рассматриваются методы выявления нечетких дубликатов в тематических коллекциях документов, формируемых в автоматическом режиме на основе публикаций, полученных из сети Интернет. Основное внимание уделяется различным модификациям метода шинглов, который позволяет достаточно быстро выполнить сравнение большого количества текстов без их предварительной обработки, что особенно важно при первичном отборе текстов для коллекции.

Ключевые слова: текстовые коллекции, методы сравнения текстов, метод шинглов, поиск нечетких дубликатов, веб-документы, веб-ресурсы.

Введение

В настоящее время накоплено огромное количество текстовой информации, и чтобы помочь пользователю ориентироваться в ее многообразии, необходимо выделять и структурировать интересующие его фрагменты. Один из способов добиться этого состоит в формировании электронных коллекций, объединяющих тексты на заданную тематику и смежные с ней. Такая организация текстовых данных помогает получать информацию, касающуюся конкретных предметных областей. Тематические текстовые коллекции формируются в рамках составления электронных библиотек, коммерческих полнотекстовых баз данных, фондов электронных документов, чтобы затем использоваться для поиска информации, необходимой пользователю или клиенту. Старейшей информационно-поисковой системой с собственными полнотекстовыми базами является запущенная в 1972 г. система Dialog [1], предоставляющая информацию специалистам. В российском сегменте среди коммерческих проектов следует

[*] Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (договор № 02.G25.31.0054).

[1] Информационно-поисковая система Dialog. URL: http://www.dialog.com/.

Загорулько Ю. А., Саломатина Н. В., Серый А. С., Сидорова Е. А., Шестаков В. К. Выявление нечетких дубликатов при автоматическом формировании тематических коллекций документов на основе Web-публикаций // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 4. С. 59–70.

© Ю. А. Загорулько, Н. В. Саломатина, А. С. Серый, Е. А. Сидорова, В. К. Шестаков, 2013

упомянуть публичную библиотеку public.ru [2] с ее электронными архивами публикаций русскоязычных СМИ и компанию «Интегрум» [3], являющуюся поставщиком деловой информации для бизнес-клиентов.

Задача формирования тематической коллекции может быть частью других задач, таких как поиск плагиата, построение рубрикатора, составление реферата и др. Методы комплектования при этом зависят от изначально поставленных задач. Тексты могут копироваться с других сайтов, присылаться владельцами, сканироваться и распознаваться с бумажного носителя. По возможности тексты, включаемые в коллекцию, проходят ручную очистку и приводятся к читаемому виду (в основном это касается коммерческих электронных библиотек и полнотекстовых баз данных).

Задача автоматического формирования текстовых коллекций. В том случае, когда требуется собрать множество коллекций на различные темы и без непосредственного участия людей, то главным методом формирования коллекции становится поиск текстов на других сайтах. При такой сборке большое количество текстов извлекается из сетевых ресурсов, возможность ручной корректировки в этом случае отсутствует. Именно этот способ отбора текстов предполагают условия решаемой нами задачи. Тип электронного ресурса и его контента может быть как известен нам заранее, если ресурсом является некий заранее известный сайт из заданного списка, так и неопределенным, как в случае анализа результатов выдачи поисковых систем. В целом процесс формирования коллекции состоит из следующих этапов.

1. Формулирование тематики коллекции. Тематика коллекции задается набором ключевых слов.

2. Сбор текстовой информации с использованием доступных поисковых машин Интернета (таких как Google, Yandex, Bing и др.) и / или сайтов из предварительно отобранного экспертами списка. Работа с каждым типом источников текстовой информации осуществляется отдельным программным модулем, что позволяет как распараллелить процесс сбора данных, так и при необходимости расширить список участвующих в нем поисковых машин и других источников.

3. Собранные таким образом тексты и веб-страницы подвергаются очистке. Из них по возможности удаляются элементы разметки, изображения и все, что не является первоначальным авторским текстом, соответствующим тематике коллекции. Тем не менее в текстах могут оставаться множественные шумы в виде элементов интерфейса и меню веб-страниц, подписей к изображениям и другие не относящиеся к теме коллекции фрагменты, затрудняющие дальнейшую работу. На этапе очистки все элементы будущей коллекции, независимо от источника, приводятся к общему текстовому виду.

4. Вследствие того, что при формировании коллекции одновременно используется несколько различных информационно-поисковых систем, велика вероятность несколько раз включить в ее состав один и тот же текст либо несколько очень похожих друг на друга. Последнее случается особенно часто в связи с распространением технологий размножения контента, таких как рерайт и автоматическая генерация текста. Таким образом, выявление нечетких дубликатов среди элементов коллекции необходимо для устранения избыточности информации.

В данной статье речь пойдет о четвертом этапе, а именно о выявлении схожих, но не совпадающих полностью документов в текстовых коллекциях, собранных автоматически с использованием информационно-поисковых систем и вики-сайтов.

Методы сравнения текстов

Метод обнаружения дубликатов в составе текстовой коллекции предусматривает наличие возможности оценить, насколько те или иные тексты похожи друг на друга. В [1] приведены подробный обзор и сравнительный анализ популярных методов оценки сходства текстов. Одна из первых работ в области сравнения документов и файлов принадлежит Уди Ман-

[2] Публичная библиотека. URL: http://www.public.ru/.
[3] Компания «Интегрум». URL: http://www.integrum.ru/.

беру [2]. Для сравнения файлов предлагалось представлять их в виде дактилограмм – контрольных сумм, вычисляемых для всех подстрок одинаковой длины.

В 1997 г. Андрей Бродер [3] развил эту идею, предложив метод, основанный на представлении документа в виде множества последовательностей фиксированной длины, только теперь последовательности, названные шинглами, состояли не из отдельных символов, а из идущих подряд слов. Этот метод завоевал наибольшую известность и теперь в различных модификациях применяется повсеместно, в частности для поиска плагиата и оценки оригинальности контента веб-сайтов. Если мощность пересечения множеств шинглов двух документов достаточно велика, то такие документы можно считать похожими. Тот факт, что количество шинглов практически совпадает с количеством слов в документе, заставил авторов также предложить два метода составления репрезентативной выборки.

Дальнейшим развитием метода шинглов стал предложенный в 2003 г. [4] метод «мегашинглов». При этом отпала необходимость в теоретико-множественных операциях. Документ представлялся теперь в виде сигнатуры – вектора, чья длина фиксирована и не зависит от количества слов в тексте документа. Таким образом, сравнение документов было сведено к сравнению координат соответствующих векторов. Двум последним алгоритмам мы уделили особое внимание. Рассмотрим их более детально.

И в первом, и во втором методах шинглы представляются своими контрольными суммами (*fingerprints*), вычисляемыми по алгоритму Карпа – Рабина [5]. Метод А. Бродера для сравнения документов использует теоретико-множественные операции. Сравнить множества шинглов, соответствующие документам, можно двумя способами. При фиксированной длине шингла сходство (*resemblance*) между документами T_1 и T_2 вычисляется как мера сходства

$$res(T_1, T_2) = \frac{|S(T_1) \cap S(T_2)|}{|S(T_1) \cup S(T_2)|}, \tag{1}$$

где $S(T)$ – множество шинглов документа T. Данное выражение, очевидно, коммутативно, и $res(T, T) = 1$.

Также можно вычислить и степень вхождения одного документа в другой (англ. containment) как меру вхождения

$$cont(T_1, T_2) = \frac{|S(T_1) \cap S(T_2)|}{|S(T_1)|}. \tag{2}$$

Выражение (2) вычисляет степень вхождения текста документа T_1 в текст документа T_2 и не является коммутативным. Схема данного алгоритма, с дополнительным шагом приведения текстов к каноническому виду, приведена на рис. 1. Что понимается под приведением текста T_i к каноническому виду TC_i, зависит от задачи и от исследователей, которые ее ставят. Мы понимаем под этим очистку текста от стоп-символов и стоп-слов, в основном предлогов, союзов и вводных слов, а также приведение всех символов текста к нижнему регистру. В других модификациях алгоритма шинглов при канонизации текста из него также удаляются все прилагательные, а слова других изменяемых частей речи приводятся к нормальной форме. Сигнатуры FP_i строятся на основе текстов, уже приведенных к каноническому виду.

Как было сказано, число шинглов сопоставимо с числом слов в документе. При увеличении объемов текстов число сравнений растет как $O(LK)$, где L и K – количество слов в соответствующих текстах. Данный факт заставил авторов метода шинглов задуматься над проблемой уменьшения мощности соответствующих множеств в формулах (1) и (2) путем замены их репрезентативной выборкой. Первый способ (метод 1 на рис. 1) построения такой выборки оставлял во множестве $S(T_i)$ только шинглы, кратные некоторому положительному целому параметру $m > 1$, второй (метод 2 на рис. 1) – n наименьших шинглов (под шинглами здесь понимаются уже не цепочки слов, а соответствующие им контрольные суммы). Выборка, построенная по методу 1, растет вместе с размером документа и может быть подставлена как в формулу (1), так и в формулу (2). Одним из основных недостатков метода 1 является

невозможность сравнения документов малого размера; в рамках же нашей задачи подобные документы отсеиваются на ранних этапах и не участвуют в составлении коллекции, это позволяет нам не опасаться, что выборка того или иного документа окажется слишком бедной или пустой. Выборка, построенная по методу 2, имеет фиксированный размер, но, как показано в [3], может использоваться только в вычислениях по формуле (1). На рис. 1 соответствующие репрезентативные выборки обозначены V_1 и V_2.

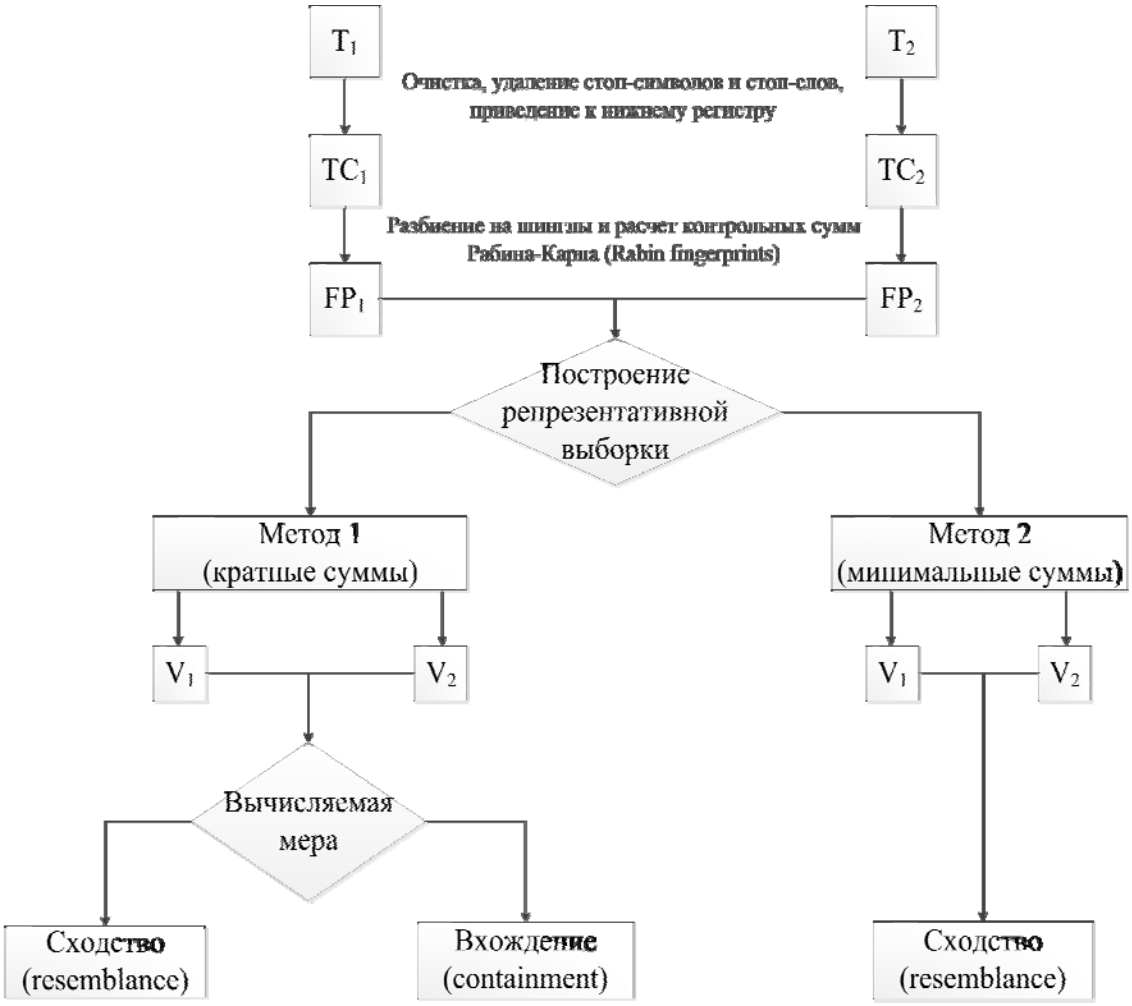

Рис. 1. Схема алгоритма шинглов

Метод «мегашинглов» использовался его авторами для решения задачи быстрого сравнения большого числа веб-страниц [4]. Каждый документ разбивался на шинглы длиной 5 слов, для каждого шингла по алгоритму Рабина [5] вычислялись 64-битные контрольные суммы, называемые прообразами. Далее, к полученному множеству прообразов применялись 84 независимых инъективных отображения из заранее подобранного семейства. Отображения данного семейства выбирались случайным образом, но после этого семейство оставалось неизменным. Для каждого отображения из множества прообразов выбирался элемент, дающий наименьший образ. В результате каждому документу ставился в соответствие сигнатурный вектор длиной 84. Сравнение документов, таким образом, сводилось к сравнению их сигнатур, причем длина сигнатуры документа, а следовательно, и число операций сравнения фиксированы.

При дальнейшей оптимизации вычислений было замечено, что в силу случайного выбора инъекций в семействе соответствующие координаты сигнатур двух документов совпадают

с вероятностью равной вероятности совпадения самих документов. С учетом этих соображений сигнатура документа разбивалась на 6 «супершинглов», где «супершингл» – это конкатенация четырнадцати идущих подряд прообразов. Допустив, что документы совпадают с вероятностью p, будем иметь вероятность совпадения соответствующих супершинглов равной p^{14}, а вероятность совпадения любой пары супершинглов будет равняться

$$1-(1-p^{14})^6-6\cdot p^{14}\cdot(1-p^{14})^5. \qquad (3)$$

Следовательно, при совпадении документов на 95 % вероятность совпадения любых двух пар супершинглов составляет около 90 %.

Из сказанного выше следовало, что для подтверждения сходства документов достаточно, чтобы у них совпали не менее двух пар супершинглов. Для ускорения проверки таких совпадений векторы супершинглов были заменены на векторы «мегашинглов», координаты которых представляют собой конкатенации всех возможных пар супершинглов. Для 6 супершинглов получается 15 мегашинглов и для сходства документов достаточно совпадения одной пары мегашинглов (рис. 2). По исходной сигнатуре FP_i строятся дополнительные сигнатуры фиксированного размера: S_i, построенная с помощью семейства инъекций, содержит 84 элемента, сигнатура супершинглов SS_i – 6. Сигнатур SS_1 и SS_2 уже достаточно для проверки сходства документов, но в целях оптимизации вводится дополнительный этап построения сигнатур мегашинглов MS_i, состоящих из 15 элементов.

Рис. 2. Схема алгоритма мегашинглов

Среди сигнатурных методов сравнения текстов кроме синтаксических, учитывающих только порядок слов, существуют и такие, которые используют лексическую информацию документов. Метод, предложенный в 2002 г. и детально описанный в [6], предусматривает вычисление специальной сигнатуры I-Match на основе слов со средним значением IDF, т. е. с использованием статистики распределения слов по всей коллекции. Два текста схожи, если их I-Match совпадают, другими словами, как и ранее, сравнение текстов сводится к сравнению их сигнатур. Отличие в данном случае кроется в самих сигнатурах, способах их построения и требуемых для этого данных.

Среди отечественных разработок еще одним лексическим сигнатурным подходом является метод «опорных» слов, предложенный С. Ильинским [1]. Сигнатура строится на основе множества опорных слов, выбираемых по определенным правилам. В целом, при построении сигнатуры предполагается наличие индекса коллекции. В [7] предлагается комплексный метод сравнения текстов, где учитываются как лексика, морфология и синтаксис, так и семантическая информация.

Характеристики исходных текстов

Как было сказано, в нашем случае процесс сбора коллекции должен быть полностью автоматизирован. Более того, в отличие от задачи, решаемой в [4], где отслеживались изменения в одних и тех же веб-страницах, нас интересует в первую очередь контент документов.

Все получаемые текстовые данные можно разделить на две группы по типу источника. В первую группу попадают веб-страницы, получаемые при анализе выдачи поисковых систем. Вторую группу составляют статьи, получаемые с различных вики-сайтов через программный интерфейс, предоставляемый соответствующим вики-движком.

Данные первой группы представляют собой веб-страницы, найденные поисковыми машинами и напрямую загруженные из Интернета. Их структура и размер в общем виде никак не стандартизированы, хотя для повышения эффективности обработки полезной может оказаться классификация источников (например, блоги, каталоги, форумы, интернет-магазины, электронные библиотеки и проч.). Нетекстовую информацию, такую как html-разметка, довольно просто распознать и удалить, однако при этом теряются данные о структуре документа, поэтому, в общем случае, удалять то, что не относится к контенту (посторонние включения и служебная информация, такая как меню, элементы оформления и т. д.), надо до удаления разметки. Может так случиться, что после очистки размер документа окажется очень малым, и он не будет включен в коллекцию.

Данные из второй группы имеют единообразную структуру, в них присутствует вики-разметка, а размер колеблется в относительно небольшом диапазоне (в силу распространенной концепции формирования вики-систем, а также технических ограничений, налагаемых вики-движком). Единообразие вики-разметки облегчает очистку текстов, однако сложность данной разметки не позволяет полностью удалить ее, сохранив при этом весь значимый текст статьи. Таким образом, тексты второй группы характеризуются содержанием остатков вики-разметки и отсутствием некоторых смысловых частей (последнее – результат более глубокой очистки). Наличие стандартов оформления вики-статей налагает отпечаток на структуру всех текстов данной группы: абзацы разделены пустой строкой, все разделы имеют заголовки.

Очевидно, что при независимом использовании нескольких поисковых машин и вики-сайтов в выборку часто попадают тексты, похожие друг на друга, но не совпадающие целиком, вследствие чего на данном этапе и возникла необходимость привлечь методы автоматического выявления дублей. В большей степени это касается текстов, полученных одновременно из источников первого и второго типов. Проверка на сходство контента присутствует как на этапе формирования новой текстовой коллекции, так и при пополнении уже существующей новыми данными. В первом случае все документы будущей коллекции необходимо сравнить между собой, во втором – сравнить вновь поступивший документ со всеми документами в коллекции.

Применение метода шинглов в задаче формирования коллекций

При выборе метода выявления дублирующихся текстов в составе коллекции нам было необходимо учесть тот факт, что в общем случае при формировании «с нуля» коллекция

не индексирована, т. е. какая-либо информация о частотных характеристиках входящих в нее слов отсутствует. Кроме того, коллекция существует в пространстве более глобальной задачи и хранится до тех пор, пока эта задача актуальна. Время жизни некоторых коллекций, таким образом, может оказаться небольшим, в отличие от постоянных часто используемых коллекций. На этапе формирования новой коллекции не всегда понятно, будет ли она постоянной или временной. По этой причине использование лексических сигнатурных методов, требующих предварительного индексирования для подсчета показателей TF-IDF и построения словарей (словари опорных слов, слов со средним значением IDF и проч.), а тем более методов, учитывающих морфологические свойства слов, неоправданно на раннем этапе, когда количество дублей еще велико.

При формировании коллекции «с нуля» требуется выполнить быстрое сравнение большого количества документов, не снабженных какой-либо разметкой (метаданными). Здесь могут эффективно применяться методы шинглов и «мегашинглов», которые могут работать с «сырыми документами». В пользу метода шинглов говорит и тот факт, что он позволяет оценить как степень сходства текстов, так и степень их вложенности друг в друга.

Одной из основных задач при построении коллекции является удаление повторяющейся неуникальной информации. В силу распространенности и общедоступности ресурсов, подобных Википедии, нередки цитаты и заимствования в веб-публикациях, на основе которых формируется коллекция. В случае, когда один из сравниваемых документов значительно превосходит по объему другой и при этом содержимое меньшего содержится в большем, возникает ситуация, когда методы оценки сходства документов не дадут достаточно точного представления о степени дублирования информации. С точки зрения оценки, вычисленной по формуле (1), документы не являются похожими, однако в рассмотренном случае меньший документ не представляет ценности для коллекции, так как почти весь его контент уже содержится в другом документе.

Для проверки применимости методов шинглов и их модификаций к решению задачи устранения дублирующей информации в коллекциях была проведена серия экспериментов на четырех тестовых коллекциях на следующие темы: «породы собак», «дизайн интерьеров», «ландшафтный дизайн» и «историческая реконструкция». При этом основное внимание уделялось точности выявления дубликатов. Обусловлено это тем, что ошибки первого рода (ошибочное занесение уникального документа в разряд дубликатов) приводят к утере оригинальной информации в коллекции, обедняя ее контент. Данный факт может в значительной мере повлиять на дальнейшую обработку коллекции, в том числе помешать переходу коллекции в разряд постоянных.

В табл. 1 приведены характеристики тестовых коллекций.

Таблица 1

Параметры тестовых коллекций

№	Тема	Количество текстов		Объем, МБ	
		д/о	п/о	д/о	п/о
1	Дизайн интерьеров	1460	948	65,8	16,8
2	Историческая реконструкция	1478	909	86,3	20,3
3	Ландшафтный дизайн	1341	838	56,2	13,2
4	Породы собак	1378	947	65,9	13,4

Примечание: д/о – до очистки, п/о – после очистки. Имеется в виду очистка от разметки и отбрасывание невалидных текстов.

Для тестирования и калибровки параметров (длина шингла w, числа m и n, пороговое значение t) были отобраны четыре метода оценки сходства и вложенности текстов, основанные на шинглах: метод мегашинглов и три модификации метода шинглов с различными способами построения сигнатур:

- сигнатура формируется из элементов индекса, кратных m;
- сигнатура – это первые n минимальных элементов индекса;
- сигнатура полностью совпадает с индексом.

Пороговое значение t – минимальное значение меры сходства текстов, при котором они считаются дублями. Для построения индексов документов использовался алгоритм Рабина – Карпа [5]. Элементы индекса – 32-битные целые числа.

Ошибки первого рода. В качестве эталона при оценке точности была взята выборка текстов, отнесенных к дубликатам хотя бы одним из методов при пороговом значении 0,5. Была проведена экспертная оценка всех обнаруженных дубликатов и, таким образом, определен процент ошибок первого рода для каждого метода, являющийся, как указано выше, наиболее важной характеристикой. Ввиду того, что невозможно экспертно оценить число необнаруженных дубликатов по всем коллекциям (см. далее), процент ошибок первого рода важен и для установления пороговых значений сходства: если в результат попадают тексты, которые эксперты не сочли дубликатами, возможно, следует увеличить порог, так как в конечном итоге коллекции ориентированы на человеческое восприятие. По этой причине экспертное мнение представляется достаточно существенным.

Обратим внимание на коллекцию 2 (историческая реконструкция), так как она выделяется повышенным содержанием дубликатов. Это связано с тем, что при сборе большое количество текстов содержало заимствования из соответствующей статьи Википедии [4].

В табл. 2 приведены результаты применения тестируемых методов с различными пороговыми значениями меры сходства в сравнении с экспертной оценкой. Установленные значения параметров:

- модуль кратности – $m = 25$;
- размер сигнатуры из минимальных контрольных сумм – $n = 160$;
- длина шингла – $w = 4$.

Таблица 2

Результаты тестов с различными значениями t

Название коллекции	Метод шинглов					
	с кратной сигнатурой		с минимальной сигнатурой		с полным сравнением	
	к. д.	о. п. р., %	к. д.	о. п. р., %	к. д.	о. п. р., %
$t = 0{,}5$						
Дизайн интерьеров	49	8	31	0	48	5,5
Историческая реконструкция	136	8,8	79	0	132	6
Ландшафтный дизайн	79	13,5	53	0	74	9,2
Породы собак	72	6,5	52	0	72	6,5
Итого	**356**	**16,4**	**215**	**0**	**326**	**6,8**
$t = 0{,}6$						
Дизайн интерьеров	46	2	31	0	42	0
Историческая реконструкция	119	2,5	73	0	115	0
Ландшафтный дизайн	65	3	52	0	65	1,5
Породы собак	68	0	50	0	65	0
Итого	**298**	**1,3**	**206**	**0**	**287**	**0,4**
$t = 0{,}7$						
Дизайн интерьеров	37	0	30	0	36	0
Историческая реконструкция	102	0	64	0	89	0
Ландшафтный дизайн	59	0	49	0	57	0
Породы собак	63	0	49	0	56	0
Итого	**261**	**0**	**192**	**0**	**238**	**0**

[4] http://ru.wikipedia.org/wiki/историческая_реконструкция

Окончание табл. 2

Название коллекции	Метод шинглов					
	с кратной сигнатурой		с минимальной сигнатурой		с полным сравнением	
	к. д.	о. п. р., %	к. д.	о. п. р., %	к. д.	о. п. р., %
			$t = 0{,}8$			
Дизайн интерьеров	33	0	29	0	32	0
Историческая реконструкция	73	0	63	0	73	0
Ландшафтный дизайн	54	0	47	0	53	0
Породы собак	56	0	49	0	51	0
Итого	**216**	**0**	**188**	**0**	**209**	**0**
			$t = 0{,}9$			
Дизайн интерьеров	30	0	28	0	30	0
Историческая реконструкция	67	0	60	0	64	0
Ландшафтный дизайн	51	0	46	0	50	0
Породы собак	53	0	47	0	49	0
Итого	**191**	**0**	**181**	**0**	**193**	**0**
			$t = 1$			
Дизайн интерьеров	29	0	28	0	28	0
Историческая реконструкция	62	0	55	0	55	0
Ландшафтный дизайн	49	0	43	0	43	0
Породы собак	51	0	45	0	48	0
Итого	**191**	**0**	**171**	**0**	**174**	**0**

Примечание: к. д. – количество дубликатов, найденных соответствующими методами; о. п. р. – процент ошибок первого рода.

Отдельно следует сказать про метод мегашинглов. Он тестировался для порогового значения 0,95. Число найденных дубликатов для всех четырех коллекций составило 198, при этом зафиксировано две ошибки первого рода. Полнота данного метода для заданного порога $t = 0{,}95$ вычисляется по формуле (3) и составляет около 90 (речь идет о полноте обнаружения текстов схожих более чем на 0,95, полнота обнаружения всех дубликатов, найденных экспертами, может быть значительно ниже). Если положить $t = 0{,}75$, значение полноты снижается практически до нуля, следовательно, невозможно изменить порог без существенной переработки сигнатуры, т. е. без изменения формулы (3). Согласно тестам, приведенным в [1], для порогового значения 0,75 удалось добиться полноты 33 %, для 0,8 – 42 %, для 0,85 – 50 %. Таким образом, доля дубликатов, которые не удалось обнаружить, превосходит 50 % даже для выделенного значения t. Точность метода при этом достаточно высока и составляет 91 % [1].

Из табл. 2 видно, что при сравнимом числе ошибок методы шинглов с полным сравнением индексов и кратной сигнатурой находят больше дубликатов, чем метод шинглов с минимальной сигнатурой. Последний хорошо зарекомендовал себя при выявлении текстов с высокой степенью сходства, что также отражено в табл. 2 и следует из рис. 3.

По итогам экспертной проверки лучшие результаты дают методы шинглов с полным сравнением и с кратной сигнатурой при пороговом значении сходства в промежутке от 0,5 до 0,6. На рис. 3 показано, как изменяется количество пар документов в коллекции «дизайн интерьеров», отнесенных разными методами к разряду схожих, в зависимости от t.

Ошибки второго рода проявляются в занесении дубликатов в разряд уникальных документов. Число таких ошибок коррелирует с полнотой. Из формулы (1), а также из описания построения сигнатуры из мегашинглов следует, что тексты, степень совпадения которых близка к 100 %, всегда будут найдены, так как совпадут шинглы и, следовательно, индексы, а в сигнатуры будут отобраны одинаковые элементы. Из формулы (2) можно заключить, что будут найдены и пары, где один текст целиком вложен в другой.

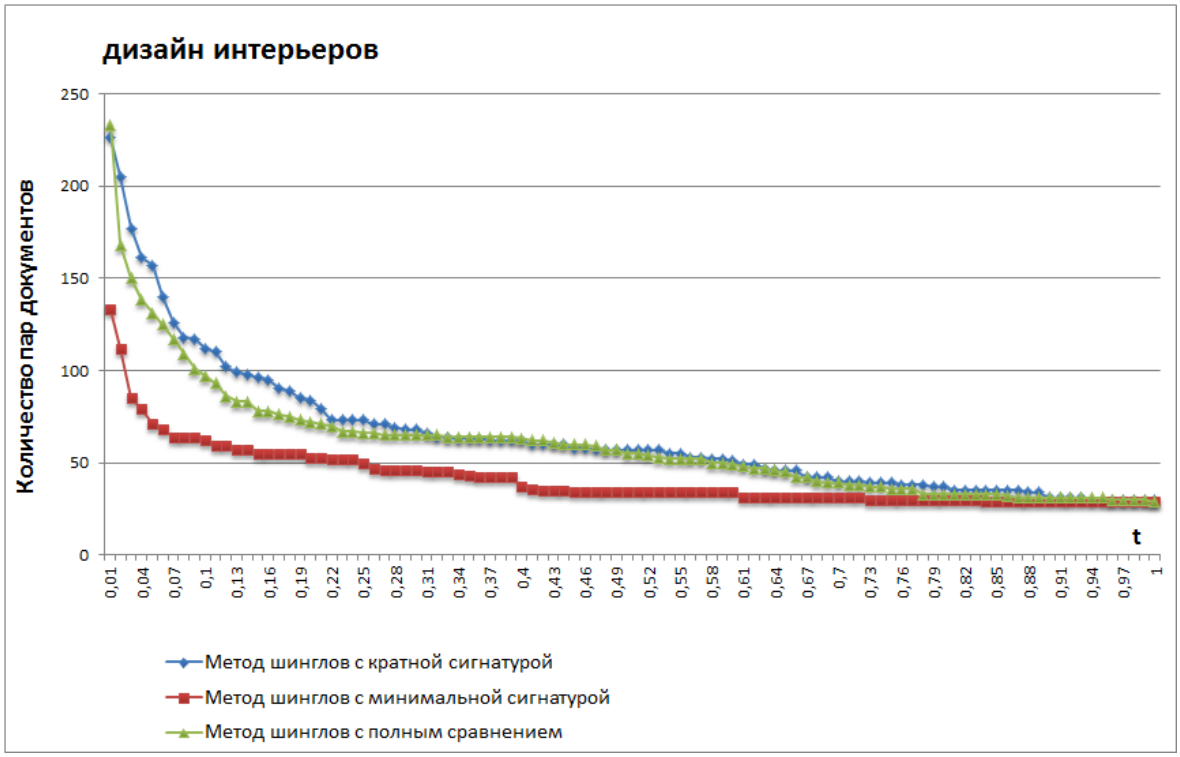

Рис. 3. График зависимости числа пар похожих документов от *t*

Как можно видеть из табл. 1, объемы тестовых коллекций достаточно велики, и для того, чтобы найти все дубликаты, экспертам пришлось бы вручную сопоставить более полутора миллионов пар документов. Очевидно, это невозможно. К тому же экспертная оценка сходства тех или иных текстов субъективна и не имеет точного числового выражения. Более того, не все тексты коллекций ввиду их особенностей, указанных выше, являются легкими для восприятия (и оценки) человеком. По этой причине проверка методов на полноту выполнялась на 100 текстах, отобранных поровну (по 25) из каждой коллекции. Единственным критерием отбора было качество текста с точки зрения его восприятия человеком. В выборку, например, не включались тексты, содержащие остатки разметки, вложенные списки и т. п.

Из ста текстов данной коллекции 24 были экспертно признаны дубликатами. Результаты предыдущих экспериментов позволили предположить, что оптимальное значение *t* находится в промежутке [0,5; 0,6]. Результаты работы методов на тестовой коллекции, приведенные в табл. 3, подтвердили данное предположение. Соответствующие характеристики для метода мегашинглов приведены в описании ошибок первого рода.

Выводы о применимости рассмотренных методов. Метод мегашинглов и метод шинглов с минимальной сигнатурой дают хорошие результаты на коллекциях, содержащих большое количество текстов со степенью сходства, близкой к единице, оба они отличаются малым

Таблица 3

Результаты экспертной проверки исследуемых методов
(ошибки второго рода, %)

t	Метод шинглов		
	с кратной сигнатурой	с мин. сигнатурой	с полным сравнением
0,5	12,5	26,5	12,5
0,6	25	31	16,7

количеством ложных дубликатов. При этом, однако, число не найденных ими схожих текстов намного больше, чем у других методов, что показала экспертная проверка тестовой коллекции.

По итогам проведенного исследования применимыми для решения поставленной задачи признаны методы шинглов с полным сравнением индексов и с кратной сигнатурой. Данные методы показали близкие результаты во всех экспериментах, оба они позволяют вычислять как сходство текстов, так и степень их вложенности друг в друга. В качестве рабочего метода в задаче формирования текстовых коллекций нами был выбран метод шинглов с кратной сигнатурой, так как он при схожем числе ошибок с методом шинглов с полным сравнением индексов имеет меньшую вычислительную сложность.

Заключение

В данной работе исследована применимость метода шинглов в решении задачи автоматического формирования тематических текстовых коллекций. Метод шинглов дает возможность сравнить большое количество текстов, не снабженных какими-либо метаданными. Это полезно при первичном формировании коллекции. Однако, как было указано ранее, необходимость сравнить тексты и найти среди них похожие возникает не только при построении новой, но и в случае пополнения уже имеющейся постоянной коллекции. При этом сигнатуры документов, хранимых долговременно, также сохраняются вместе с ними, поэтому при добавлении новых элементов достаточно построить для них сигнатуры и сравнить с сигнатурами из соответствующей базы.

Часто используемые коллекции, соответствующие, как правило, популярным тематикам, переходят в разряд постоянных. Такие коллекции хранятся отдельно и периодически пополняются новыми текстами, их можно проиндексировать и снабдить каждый документ метаданными. Естественна мотивация сделать часто используемые коллекции более информативными и «чистыми», тщательнее отбирая контент для их пополнения. Тот факт, что коллекция уже находится в базе, позволяет применять более эффективные методы проверки на совпадение контента, не беспокоясь при этом о времени их обработки, – обновление коллекции может проходить в фоновом режиме. К таким методам относятся упомянутые лексические сигнатурные методы, а также другие методы, учитывающие семантику и морфологию слов, в частности позволяющие сравнивать тексты на уровне словосочетаний и / или с учетом синонимии слов.

Список литературы

1. *Зеленков Ю. Г., Сегалович И. В.* Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. IX Всерос. науч. конф. Переславль-Залесский, 2007. Т. 1. С. 166–174.

2. *Manber U.* Finding Similar Files in a Large File System // Proc. USENIX WINTER Technical Conference. 1994. P. 1–10.

3. *Broder A., Glassman S., Manasse M. Zweig G.* Syntactic Clustering of the Web // Comput. Netw. ISDN Syst. 1997. Vol. 29. P. 1157–1166.

4. *Fetterly D., Manasse M., Najor M. et al.* A Large-Scale Study of the Evolution of Web Pages // ACM. 2003. P. 669–678.

5. *Rabin M.* Fingerprinting by Random Polynomials. Center for Research in Computing Technology. Harvard, 1981. 24 p.

6. *Kołcz A., Chowdhury A.* Lexicon Randomization for near-Duplicate Detection with I-Match // The Journal of Supercomputing. 2008. Vol. 45. Is. 3. P. 255–276.

7. *Соченков И. В.* Метод сравнения текстов для решения поисково-аналитических задач // Искусственный интеллект и принятие решений. 2013. Вып. 2. С. 32–43.

Материал поступил в редколлегию 28.10.2013

Yu. A. Zagorulko, N. V. Salomatina, A. S. Sery, E. A. Sidorova, V. K. Shestakov

DETECTING NEAR-DUPLICATES FOR AUTOMATICALLY FORMING THEMATICAL TEXT COLLECTIONS ON THE BASIS OF WEB DOCUMENTS

Approaches to detecting near-duplicates appearing in thematic text collections accumulated automatically on the basis of text documents obtained from the Internet are discussed. The paper is focused on various modifications of shingle algorithm since it allows comparing a large number of texts quickly and without any preprocessing. The latter is particularly important when forming collections of raw texts.

Keywords: text collection, text comparing, shingle algorithm, near-duplicate, web documents, web resources.

References

1. *Zelenkov Yu., Segalovich I.* Comparative Analysis of Near-Duplicate Detection Methods of Web Documents // Proc. of IX All-Russian Research Conference RCDL'2007. Pereslavl-Zalesskij, 2007. Vol. 1. P. 166–174.

2. *Manber U.* Finding Similar Files in a Large File System // Proc. USENIX WINTER Technical Conference. 1994. P. 1–10.

3. *Broder A., Glassman S., Manasse M. Zweig G.* Syntactic Clustering of the Web // Comput. Netw. ISDN Syst. 1997. Vol. 29. P. 1157–1166.

4. *Fetterly D., Manasse M., Najor M. et al.* A Large-Scale Study of the Evolution of Web Pages // ACM. 2003. P. 669–678.

5. *Rabin M.* Fingerprinting by Random Polynomials. Center for Research in Computing Technology. Harvard, 1981. 24 p.

6. *Kołcz A., Chowdhury A.* Lexicon Randomization for near-Duplicate Detection with I-Match // The Journal of Supercomputing. 2008. Vol. 45. Is. 3. P. 255–276.

7. *Sochenkov I. V.* Text Comparison Method for a Search and Analytical Engine // Artificial Intelligence and Decision Making. 2013. Vol. 2. P. 32–43.