Метод обнаружения дубликатов в потоке текстовых документов

© А.М. Андреев © Д.В. Березкин © И.А. Козлов © К.В. Симаков
МГТУ им. Н.Э. Баумана,
Москва
arkandreev@gmail.com dmitryb2007@yandex.ru kozlovilya89@gmail.com skv@ixlab.ru

Аннотация

Работа посвящена решению задачи устранения дублирующихся документов из потока текстовых сообщений. Приведена многокритериальная модель документа, предложен метод обнаружения дубликатов на основе бинарной классификации с помощью метода опорных векторов. Основной акцент сделан на обеспечении применимости метода для обработки документов из разных предметных областей. Предложен способ снижения вычислительной сложности метода посредством предварительной фильтрации кандидатов.

1 Введение

В настоящее время во многих предметных областях существует потребность в формировании больших текстовых коллекций. При этом производится сбор текстовой информации из открытых Интернет-источников, а также специализированных ресурсов. Основной областью использования создаваемых таким образом хранилищ документов является интеллектуальная обработка текстов, которую, как правило, можно отнести к классу Text Mining.

С ростом количества разнообразных источников данных в сети Интернет (новостные сайты, блоги, социальные сети) всё более серьезной проблемой становится дублирование информации. Сообщения, публикуемые одним источником, зачастую многократно перепечатываются другими (в исходном виде или с небольшими изменениями). В результате, при выполнении автоматического сбора документов из многочисленных источников в формируемой текстовой коллекции накапливаются идентичные или близкие по содержанию документы – дубликаты. В некоторых задачах наличие дубликатов должно учитываться – например, при определении значимости сообщений [14]. Но в большинстве случаев попадание таких документов в коллекцию снижает её качество [12, 16].

В данной статье рассматривается решение задачи обнаружения дубликатов в потоке текстовых сообщений. Особое внимание уделяется обеспечению возможности использования разработанного метода для обработки документов из различных предметных областей.

2 Постановка задачи

2.1 Функционирование системы сбора и обработки новостной информации

В работе [9] авторами было предложено решение задачи качественного автоматического сбора новостных данных из Интернет-источников, предполагающего извлечение с веб-страницы текста новости, а также сопутствующих метаданных, включающих название, дату публикации, автора новости и др. При этом осуществляется контроль корректности извлекаемой информации, то есть проверка соответствия текстов загружаемых документов исходным текстам новостей на сайте.

Текстовые данные, извлеченные с веб-сайтов, подвергаются обработке различными методами интеллектуального анализа [7-8], такими как автоматическая классификация и кластеризация документов, извлечение знаний и фактов из естественно-языковых текстов, выявление трендов и прогноз развития ситуаций.

Для эффективного применения перечисленных методов обработки текстовых данных необходимо обеспечить качество анализируемой коллекции документов. Помимо вышеупомянутой корректности каждого конкретного текста, качество коллекции подразумевает требование оригинальности составляющих её новостей. Присутствие в обрабатываемом наборе одинаковых или очень близких по содержанию документов может отрицательно сказаться на качестве обработки. Это касается работы модулей, выполняющих статистический анализ документов, например, модуля анализа трендов. Его работа основана на выявлении в коллекции новостей, относящихся к анализируемой ситуации, и определении зависимости частоты встречаемости таких документов от времени. Появление дублей приведет

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

к многократному учету модулем идентичных новостей, что повлечет за собой некорректный вид построенной зависимости.

Для обеспечения оригинальности документов, составляющих текстовую коллекцию, в систему сбора необходимо встроить подсистему, задачей которой является оперативное обнаружение и удаление из коллекции нечетких дубликатов (рис. 1). Она должна анализировать каждый загружаемый документ и принимать решение о его оригинальности. Для этого необходимо сравнить его с загруженными ранее новостями и определить, является ли он нечетким дубликатом одной из них. При обнаружении дубля он должен быть удален до этапа загрузки данных в базу данных.

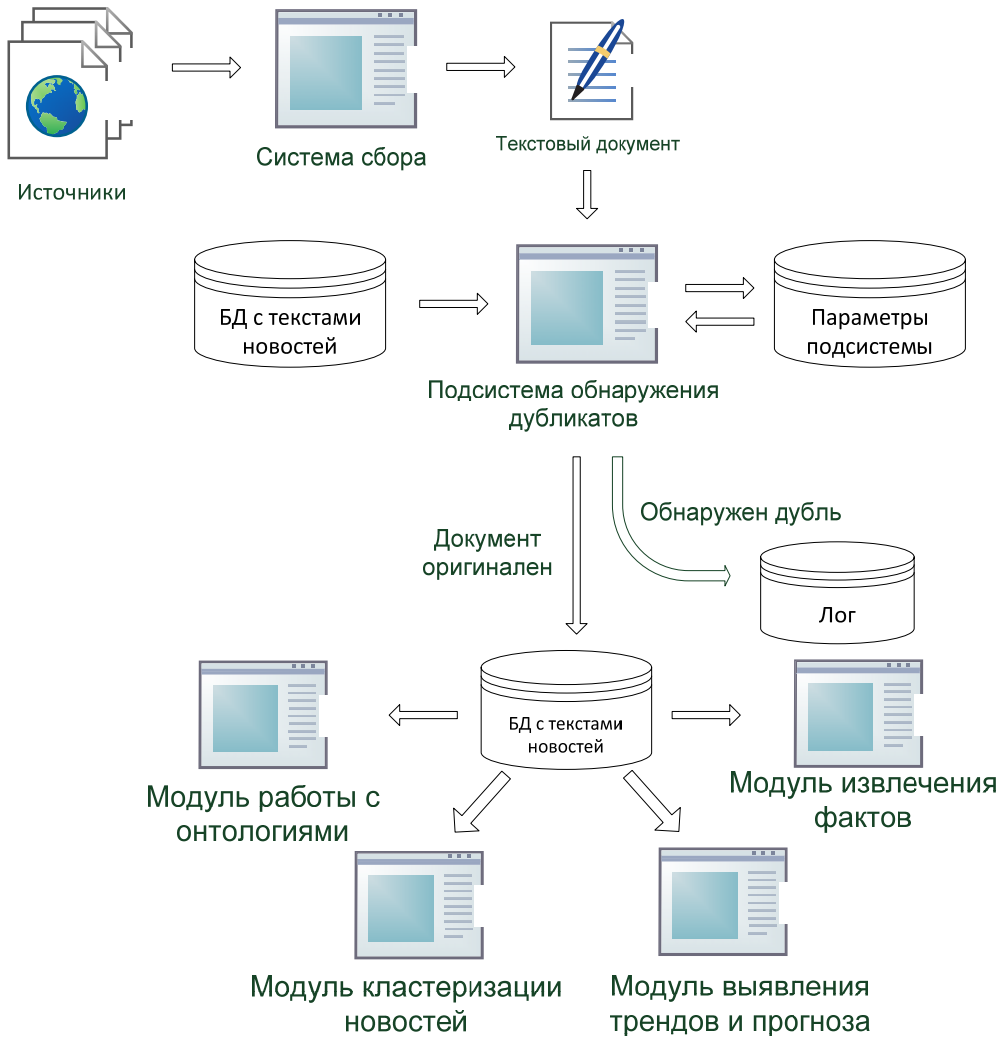

Рис. 1. Место подсистемы обнаружения дубликатов в системе автоматизированного сбора и анализа новостной информации

2.2 Особенности решаемой задачи

Если проблема обнаружения и удаления полных дублей тривиальна, то при необходимости распознавать нечеткие дубликаты (то есть, документы, имеющие различный текст, но близкие по содержанию) возникают значительные сложности.

В связи с явлением частичного дублирования в работе [10] предлагается понятие информативной необходимости элемента высказывания: «Всякий частично-дублетный элемент, если он отвечает коммуникативной задаче высказывания, признается информативно-необходимым в той же мере, что и прочие, недублетные элементы речи: такой элемент информативно необходим постольку, поскольку вносит свой уникальный, неповторимый в других элементах, вклад в суммарную информацию, передаваемую высказыванием». Таким образом, задача обнаружения нечетких дубликатов состоит в распознавании и удалении сообщений, не являющихся информативно-необходимыми.

Установить информативную необходимость и ценность некоторого высказывания возможно только с учетом соответствующего ситуативного контекста. Так, при ручной проверке документов эксперт, определяя наличие или отсутствие дублирования, принимает решение с учетом предметной области и характера документов, составляющих ана-

лизируемую коллекцию. Например, при работе с юридическими документами особое внимание должно уделяться метаданным – для текстов такого типа два документа с практически идентичным содержанием, но различающимися названиями и датами публикации не могут считаться дублями. Другой подход используется при анализе экономических новостей, например, сводок о состоянии фондового рынка – дубликатами не должны признаваться документы, имеющие одно и то же название и одинаковый текст, но различающиеся числовыми данными (значениями курсов валют).

Таким образом, в разных случаях эксперт сравнивает документы с точки зрения различных критериев (близость текстового содержания, сходство названий, разница во времени публикации), то есть использует различные модели распознавания дубликатов в зависимости от предметной области и контекста коммуникационных сообщений. Поэтому не представляется возможным выработать единую систему правил для обнаружения нечеткого дублирования сразу для всех случаев. Следовательно, разрабатываемая подсистема должна иметь возможность гибкой настройки на разные предметные области, что позволит ей моделировать деятельность эксперта по распознаванию дублей с использованием различных моделей. Поскольку система сбора выполняет извлечение документов из множества источников, которые относятся к различным предметным областям, необходимо обеспечить возможность работы с несколькими моделями распознавания дубликатов одновременно.

Еще одной проблемой, с которой приходится столкнуться при решении задачи устранения дубликатов, является большой объем обрабатываемых данных. Наличие в коллекции сотен тысяч документов делает весьма трудоемким анализ каждого нового сообщения путем сравнения его с каждым из ранее загруженных. Эта проблема может быть решена с помощью приближенных методов, но их применение ведет к снижению качества обнаружения дубликатов, то есть уменьшению точности и полноты [12]. При разработке предлагаемого подхода решалась задача совмещения высокого качества и низкой вычислительной сложности проверки документов.

3 Обзор методов обнаружения дублей

3.1 Методы, основанные на использовании шинглов

Одним из наиболее популярных методов, используемых при поиске нечетких дубликатов вебдокументов, является алгоритм шинглов [2, 5]. Он основан на представлении документа в виде множества всевозможных последовательностей фиксированной длины *k*, состоящих из соседних слов. Такие последовательности называются «шинглами». Два документа считаются похожими, если их множества шинглов значительно пересекаются. Количество шинглов примерно равно длине документа в словах, поэтому в целях повышения эффективности авторы оригинального алгоритма предложили несколько способов сэмплирования множества.

Дальнейшим развитием этого метода стал алгоритм «супершинглов» [4]. Его идея состоит в применении к элементам множества шинглов различных хэш-функций и выборе для каждой из них шингла, минимизирующего её значение. Из выбранных шинглов формируются группы, именуемые «супершинглами». Два документа считаются похожими, если мера сходства их наборов «супершинглов» не меньше заданного значения.

3.2 Сигнатурные методы

Другим распространенным классом приближенных подходов к поиску нечетких дубликатов является класс сигнатурных методов. Подробный обзор алгоритмов этого класса выполнен в [12]. Общей идеей является представление документа с помощью одного числового значения – «сигнатуры», что сводит проверку схожести документов к сравнению их сигнатур. Совпадение этих значений означает, что документы являются нечеткими дубликатами. Существует множество способов вычисления сигнатур документов:

- использование хэш-функции, вычисленной для всего документа (это позволяет обнаруживать лишь точные дубликаты);
- использование хэш-функции, вычисленной для строки, полученной из сцепленных в алфавитном порядке нескольких слов документа с наибольшими значениями весов, рассчитанных различными методами (например, TF, TF-IDF и OptFreq);
- использование хэш-функции, вычисленной для строки, полученной из сцепленных в алфавитном порядке нескольких наиболее длинных или «тяжелых» (то есть, состоящих из слов с наибольшим суммарным значением весов) предложений документа.

Несколько иной подход предложен в работе [14]: здесь сигнатура представляет собой не хэш-сумму цепочки слов, а саму цепочку. При этом документы признаются дубликатами при совпадении заданного числа элементов их цепочек.

3.3 Методы, использующие векторные модели

В задачах интеллектуальной обработки текстов (Text Mining) широко используются векторные модели текстовых документов. При этом каждое сообщение представляется в виде вектора в многомерном признаковом пространстве $D = (D^1, D^2, ..., D^N)$, каждый элемент которого отражает некоторую характеристику документа.

В качестве элементов могут использоваться слова, встречающиеся в текстах коллекции [11]. При этом значениями элементов вектора (1), представляющего некоторый документ, являются веса соответствующих слов, отражающие их значимость для этого документа:

$$d_i = (w_i^1, w_i^2, ..., w_i^n), \quad (1)$$

где N – общее количество различных слов во всех документах, w_i^j – вес j-ого слова в i-ом документе. Хотя такой выбор признакового пространства является наиболее распространенным, могут применяться и другие характеристики текстов, например, частота появления различных пар символов или частота появления тех или иных частей речи [17].

В работе [1] векторная модель использована при решении задачи обнаружения дубликатов. При этом вектором представляется не отдельный документ, а пара документов из обучающей выборки. В качестве значения элемента вектора здесь используется произведение весов соответствующего слова в первом и втором документе пары. Полученный вектор подвергается классификации с помощью метода опорных векторов (support vector machine, SVM) [15] для принятия решения о наличии или отсутствии дублирования.

Несколько иной подход предложен в статье [13]. Он также использует векторное представление пары документов, но вектор в целом здесь характеризует схожесть элементов пары, а его отдельные компоненты – близость документов с точки зрения различных критериев. Пары, отмеченные в обучающей выборке как «дубликаты» или «не дубликаты», представлены двумя кластерами точек многомерного пространства. Таким образом, задача обнаружения дублей сводится к классификации новых пар документов, то есть отнесению их к одному из этих кластеров. Классификация основана на выборе кластера, центроид которого находится ближе к точке, представляющей новую пару документов.

3.4 Метод, основанный на выявлении близких по смыслу частей текстов

В [16] решается несколько иная, но близкая задача: формирование из группы документов, описывающих некоторое событие, одного сообщения, содержащего только оригинальную информацию о событии. Для этого выполняется поиск и исключение из текстов документов фрагментов, содержащих идентичную информацию. С этой целью выполняется представление документов цепочками значимых слов, сравнение этих цепочек и обнаружение их схожих участков.

3.5 Анализ рассмотренных методов обнаружения дубликатов

У сигнатурных методов и алгоритмов, использующих шинглы, есть некоторые схожие черты: эти методы минимизируют вычислительную сложность операции сравнения документов. Поэтому они находят широкое применение в системах, работающих с гигантскими объемами данных (например, в поисковых системах). Обратной стороной медали является их узкая направленность – эти методы и модели представления текстов, которыми они оперируют (шинглы, сигнатуры), пригодны лишь для устранения дублей и не могут быть использованы для других задач. Однако, как было показано выше, очищенные от дубликатов данные впоследствии подвергаются разнообразной обработке и анализу. Поэтому представляется целесообразным применять для обнаружения дублей алгоритмы и модели, которые могут быть использованы для задач интеллектуальной обработки текстов.

Модель и метод, представленные в [16], также предназначены исключительно для решения конкретной задачи, а именно устранения идентичных фрагментов сообщений. Кроме того, для эффективного использования этого метода документы должны быть предварительно распределены по кластерам, соответствующим событиям. В нашей же ситуации устранение дублей, напротив, выполняется на этапе предварительной обработки данных перед использованием интеллектуальных аналитических методов, таких как обнаружение событий.

4 Предложенный подход к обнаружению дубликатов

В целях обеспечения возможности применения разрабатываемой модели документов для решения различных задач из области Text Mining, было решено использовать в качестве её основы векторное представление текстов. Однако использование слов документов в качестве признаков (1) позволяет сравнить сообщения лишь с точки зрения состава слов, что недостаточно для принятия правильного решения. Во многих предметных областях существенную роль играют и другие критерии (см. 2.2), и эти критерии должны быть включены в модель.

Таким образом, модель должна предусматривать возможность сравнения документов по различным признакам. Окончательно же решение должно приниматься на основе анализа пары документов с точки зрения всех критериев. Исходя из этого, удобно представить пару документов (d_i, d_j) вектором (2), элементами которого являются результаты сравнения документов по соответствующим признакам:

$$\rho_{i,j} = (\rho_{i,j}^1, \rho_{i,j}^2, ..., \rho_{i,j}^k), \qquad (2)$$

где $\rho_{i,j}^k$ характеризует сходство документов d_i и d_j по k-му критерию. На основе этого вектора принимается решение о наличии или отсутствии дублирования. Для этого необходимо задать функцию, выполняющую интерпретацию вектора $\rho_{i,j}$, то есть определяющую вектор в один из двух классов, один из которых означает наличие дублирования (обозначим его M_+), другой – отсутствие (M_-):

$$D(\rho) = \begin{cases} 1, & \rho_{i,j} \in M_+; \\ 0, & \rho_{i,j} \in M_-. \end{cases} \qquad (3)$$

С учетом выбранного подхода, процесс обнаружения дублей можно разбить на следующие этапы (рис. 2):

1. Построение модели документа, отражающей характеристики новостного сообщения с точки зрения каждого из выбранных критериев.

2. Сравнение моделей двух документов и получение результирующего вектора $\rho_{i,j}$.

3. Интерпретация вектора с помощью решающей функции $D(\rho_{i,j})$.

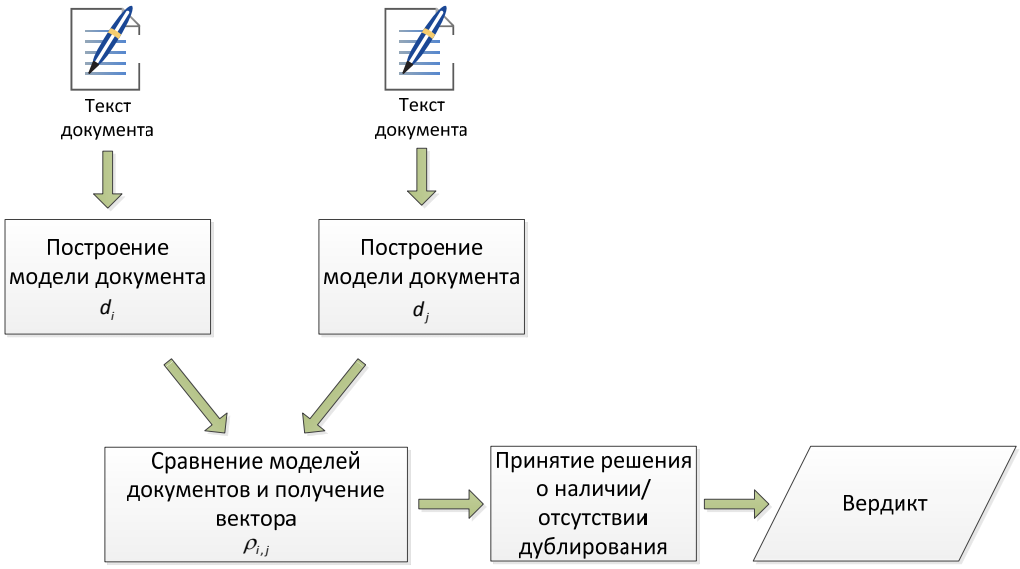

Рис. 2. Этапы обнаружения дубликатов

5 Модель документа

Важным этапом решения задачи является выбор критериев для сравнения документов. Набор критериев должен быть достаточно выразительным, чтобы обеспечивать возможность гибкой настройки модели в соответствии со спецификой различных предметных областей. Для определения набора критериев было проведено исследование выборки текстовых документов, автоматически собираемых из различных Интернет-источников.

В качестве источников были выбраны 35 сайтов по различной тематике: основные новостные сайты, публикующие материалы общественно-политической и экономической тематики, официальные сайты органов государственной власти РФ, некоторые сайты органов законодательной и исполнительной власти субъектов РФ. Такой набор сайтов позволил охватить значительное число тем и типов информационных сообщений (ленты новостей, аналитические статьи и обзоры, документы правового характера, документы, содержащие финансово-экономические показатели).

1200 пар документов были проанализированы экспертами вручную. В результате выполненного анализа были выявлены следующие критерии, характеризующие модель распознавания дубликатов.

5.1 Содержание текста

В большинстве случаев определяющее значение имеет близость текстов. Прежде всего, это характерно для общественно-политических новостей, вклад которых в суммарную информацию, передаваемую новостным потоком, определяется оригинальностью их текстового содержания.

Для обеспечения возможности сравнения составов слов документов, модель должна включать векторное представление текста сообщения (1). Если слово не встречается в документе, его вес равен нулю. Для остальных слов вес рассчитывается по методу TF-IDF с использованием алгоритма Okapi BM25 [6]. Рассчитанный таким образом вес слова $wt = tf * idf$ пропорционален частоте его употребления в документе tf и обратно пропорционален частоте употребления слова в других документах коллекции idf. В результате, модель текста документа d_i представляет собой вектор:

$$d_i^w = (w_i^1, w_i^2, ..., w_i^{N^w}), \quad (4)$$

где N^w – общее количество различных слов во всех документах, w_i^j – вес j-го слова в i-ом документе.

Векторное представление позволяет использовать для сравнения текстов простые алгебраические методы. В качестве меры сходства часто используют евклидову метрику – расстояние между двумя точками в многомерном пространстве, вычисляемое по теореме Пифагора. Однако она плохо подходит для сравнения документов, похожих по содержанию, но значительно различающихся по размеру. Поэтому при решении задач информационного поиска более распространен другой способ сравнения векторов, называемый косинусной мерой. Близость векторов оценивается на основании значения косинуса угла θ между ними, что позволяет не учитывать при сравнении длину векторов:

$$sim_{\cos}(d_i^w, d_j^w) = \frac{\sum_{n=1}^{N} w_i^n w_j^n}{\sqrt{\sum_{n=1}^{N} (w_i^n)^2} \sqrt{\sum_{n=1}^{N} (w_j^n)^2}}. \quad (5)$$

На первый взгляд, в целях обнаружения дубликатов лучше использовать евклидово расстояние, поскольку размер должен играть роль при сравнении документов: тексты существенно различающейся длины с очень низкой вероятностью являются нечеткими дубликатами. Однако для обеспечения большей гибкости системы было решено разделить оценку документов с содержательной и структурной точки зрения. Чтобы сделать оценку содержательной близости документов независимой от других характеристик, решено использовать для сравнения косинусную меру близости. При использовании неотрицательных весов слов косинусная мера принимает значения в интервале [0, 1], поэтому в качестве оценки различия векторов используется значение

$$\rho^w_{i,j} = 1 - sim_{\cos}(d^w_i, d^w_j). \quad (6)$$

5.2 Содержание заголовка

Отдельно при принятии решения учитываются заголовки сообщений, причем их роль значительно варьируется в зависимости от предметной области. При перепечатке новостей общественно-политической тематики с одного сайта на другой нередко изменяется только заголовок, причем иногда – весьма существенно. В таком случае наличие дублирования может быть обнаружено на основе близости остального текста новостей. Однако для сообщений другого типа (например, правового характера), различие заголовков имеет решающее значение, и такие документы не должны признаваться дубликатами, несмотря на близкое содержание.

В модели заголовок текста представляется аналогично основному тексту (4), с той лишь разницей, что в качестве элементов вектора используются слова, встречающиеся в заголовках документов коллекции: $d^t_i = (t^1_i, t^2_i, ..., t^{N^t}_i)$, где N^t – общее количество различных слов в заголовках всех документов, t^j_i – вес *j*-го слова в заголовке *i*-го документа.

Сравнение составляющих моделей, отражающих содержание заголовков, также выполняется аналогично сравнению содержания документов (6):

$$\rho^t_{i,j} = 1 - sim_{\cos}(d^t_i, d^t_j).$$

5.3 Предложения и абзацы

Помимо оценки близости текстового содержания документов в целом, эксперт обращает особое внимание на наличие в сообщениях идентичных структурных элементов текстов – предложений и абзацев. Это связано с тем, что при перепечатке некоторым источником ранее опубликованного документа многие из этих элементов переносятся в текст-дубликат без изменений.

Для сравнения структурных элементов текста документ представляется множествами своих предложений $d^c_i = (c^1_i, c^2_i, ..., c^{N^c_i}_i)$ и абзацев $d^p_i = (p^1_i, p^2_i, ..., p^{N^p_i}_i)$, где N^c_i и N^p_i – количество предложений и абзацев в *i*-ом документе. В свою очередь, каждое предложение представляет собой последовательность слов, а потому может быть представлено векторной моделью $c^j_i = (c^{j,1}_i, c^{j,2}_i, ..., c^{j,N^w}_i)$, где $c^{j,k}_i$ – вес *k*-го слова в *j*-ом предложении *i*-ого документа. Аналогичным образом представлен каждый абзац документа: $p^j_i = (p^{j,1}_i, p^{j,2}_i, ..., p^{j,N^w}_i)$.

Поскольку структурные элементы представлены множествами, для определения их сходства можно использовать коэффициент Жаккара. Так, для предложений близость документов будет равна

$$sim^c_j(d^c_i, d^c_j) = \frac{\left| d^c_i \cap d^c_i \right|}{\left| d^c_i \cup d^c_i \right|}. \quad (7)$$

Такая мера близости принимает во внимание лишь количество совпадающих и различающихся предложений, но не учитывает, какие именно предложения совпадают и различаются. Однако совпадение значимых, содержательных предложений должно иметь больший вес, чем одновременное появление в обоих документах одинаковых коротких и незначительных фраз. В связи с этим вместо количества предложений используется их суммарный вес. Вес каждого предложения рассчитывается как сумма весов составляющих его слов.

Кроме того, представленная мера является симметричной, и потому она плохо подходит для сравнения документов, один из которых получен из другого путем удаления нескольких предложений: в этом случае первое сообщение содержит дополнительную информацию относительно второго, но второе не имеет оригинальных данных относительно первого. Чтобы учесть требуемую несимметричность, было решено использовать меру включения вместо меры сходства:

$$sim^c_{inc}(d^c_i, d^c_j) = \frac{\sum\limits_{c \in \left| d^c_i \cap d^c_i \right|} [\sum\limits_{k=1}^{N^w} c^k]}{\sum\limits_{c \in d^c_i} [\sum\limits_{k=1}^{N^w} c^k]}. \quad (8)$$

Для оценки различия документов с точки зрения предложений используются значения $\rho^c_{i,j} = 1 - sim^c_{inc}(d^c_i, d^c_j)$ и $\rho^c_{j,i} = 1 - sim^c_{inc}(d^c_j, d^c_i)$.

Аналогичным образом выполняется сравнение абзацев, результатом которого являются значения меры различия $\rho^p_{i,j}$ и $\rho^p_{j,i}$.

5.4 Числовые данные

Кроме документов, содержащих только текстовую информацию, часто встречаются и те, которые включают числовые данные. В ряде случаев даже незначительное изменение этих данных может существенно повлиять на содержание сообщения. Примером таких документов являются новостные сообщения из области экономики (новости о со-

стоянии фондового рынка) или спорта (сообщения о результатах соревнований). Такие документы не должны признаваться дубликатами даже при полном совпадении их текста.

Для сравнения документов с точки зрения числовых значений каждое сообщение представляется набором чисел, извлеченных из его текста: $d_i^n = \{n_i^1, n_i^2, ..., n_i^{N_i^{n_d}}\}$, где $N_i^{n_d}$ – количество различных чисел в *i*-ом документе. Для выполнения оценки сходства таких наборов также используется мера включения, однако элементы сравниваемых множеств не являются взвешенными, а потому учитывается лишь количество одинаковых и различающихся числовых значений:

$$sim_{inc}^n(d_i^n, d_j^n) = \frac{\left|d_i^n \cap d_i^n\right|}{\left|d_i^n\right|}. \quad (9)$$

Расстояние между документами по данному критерию равно $\rho_{i,j}^n = 1 - sim_{inc}^n(d_i^n, d_j^n)$.

Для некоторых предметных областей важен не только состав набора чисел, но и порядок их следования в тексте документа. Это относится, в частности, к спортивным новостям, где разные последовательности одних и тех же чисел могут соответствовать различным результатам соревнований (например, два сета в теннисном матче, завершившиеся со счетом «6:4» и «4:6»). В таком случае для представления сообщения используется кортеж $d_i^n = \{n_i^1, n_i^2, ..., n_i^{N_i^{n_a}}\}$, где $N_i^{n_d}$ – общее количество чисел в *i*-ом документе.

В этом случае для сравнения документов необходимо выбрать меру различия, учитывающую порядок следования элементов. Такой мерой является расстояние Дамерау–Левенштейна [3], равное количеству операций вставки, удаления, замены и перестановки элементов, необходимых для преобразования одной последовательности символов (в данном случае – чисел) в другую. Эта мера является модификацией расстояния Левенштейна, отличающаяся наличием операции перестановки двух соседних символов (транспозиции). Это важно для нашей задачи, поскольку при перепечатке документа иногда изменяется порядок следования его абзацев и предложений, что приводит к появлению перестановок в последовательности чисел.

Расстояние между сообщениями при использовании этой меры равно $\rho_{i,j}^n = dist_{DL}(d_i^n, d_j^n)$ и является симметричным.

5.5 Фотографии и ссылки

Информация в сообщении может быть представлена не только текстом или числовыми данными, но и различными объектами, включенными в текст документа - фотографиями, видеороликами, ссылками на сторонние источники. Присутствие в документе дополнительных фото- и видеоматериалов значительно повышает вероятность его оригинальности

При сравнении фотоматериалов, включенных в сообщение, возникают сложности: определить идентичность фотографий в двух документах проблематично, поскольку одинаковые с точки зрения эксперта фотографии могут иметь различные URL и разный размер. Поэтому было принято решение учитывать не сами фотографии, а их количество в документе: d_i^{im}. Также в модель включается компонент, отражающий количество ссылок, присутствующих в тексте сообщения: d_i^h. Различие документов с точки зрения этих критериев определяется как разность соответствующих значений: $\rho_{i,j}^{im} = d_i^{im} - d_j^{im}$, $\rho_{i,j}^h = d_i^h - d_j^h$.

5.6 Дата и время публикации

Существенным фактором, влияющим на принятие решения о наличии или отсутствии дублирования, является разница во времени публикации сообщений. Так, при дублировании новостных статей перепечатыванию обычно подвергаются свежие новости, недавно опубликованные на сайте первоисточника. С увеличением интервала между моментами появления документов в сети вероятность дублирования быстро убывает

В модели эта характеристика сообщения представлена посредством POSIX-времени момента публикации (которое определяется как количество секунд, прошедших с полуночи 1 января 1970 года до момента, когда документ был опубликован источником): d_i^{dt}. Различие между документами определяется как разность между моментами публикации сообщений в секундах: $\rho_{i,j}^{dt} = d_i^{dt} - d_j^{dt}$.

5.7 Авторитетность источника

При анализе документов эксперты обращают внимание на источники сообщений, при этом они руководствуются своими представлениями об авторитетности источников. Статья из авторитетного источника (который обычно публикует оригинальные материалы) имеет существенно меньшую вероятность быть признанной дубликатом, чем документ, полученный из источника, регулярно занимающегося перепечаткой чужих сообщений.

Авторитетность источника *s* представляется значением $aut(s) \in [0, 1]$, отражающим вероятность публикации им оригинального сообщения. Это значение может быть задано экспертом вручную или получено на основе обучающей выборки как соотношение количества оригинальных документов, поступивших от источника, к общему количеству опубликованных им сообщений.

В модель документа включается компонент, отражающий авторитетность источника, опубликовавшего этот документ: $d_i^a = aut(src(d_i))$, где $(src(d_i)$ – функция, устанавливающая соответствие между документом и его источником.

Таким образом, модель документа представляет собой совокупность компонентов, характеризующих сообщение с точки зрения различных критериев:

$$d_i = (d_i^w, d_i^t, d_i^n, d_i^c, d_i^p, d_i^{im}, d_i^h, d_i^{dt}, d_i^a). \quad (10)$$

6 Метод обнаружения дубликатов

6.1 Интерпретация результата сравнения

После получения моделей d_i и d_j двух документов необходимо сравнить их и вынести решение о том, являются ли документы нечеткими дубликатами. Для этого выполняется анализ схожести сообщений по каждому из критериев. Результатом сравнения документов с точки зрения некоторого критерия k является значение $\rho_{i,j}^k$. Выполнив попарно сравнение компонентов моделей для каждого из критериев, получим вектор

$$\rho_{i,j} = (\rho_{i,j}^w, \rho_{i,j}^t, \rho_{i,j}^n, \rho_{i,j}^c, \rho_{i,j}^p, \rho_{i,j}^{im}, \rho_{i,j}^h, \rho_{i,j}^{dt}, \rho_{i,j}^a) \quad (11)$$

Ввиду вышеуказанной несимметричности мер сходства, используемых для некоторых критериев, результатом сравнения двух документов являются два различных вектора $\rho_{i,j}$ и $\rho_{j,i}$, характеризующие степень отличия первого и второго сообщения друг от друга. Каждый из этих векторов интерпретируется с помощью функции $D(\rho)$ (3).

Для интерпретации результата сравнения необходимо решить задачу бинарной классификации, то есть отнести вектор к классу M_+ или M_-. Для настройки параметров классификатора используется обучающая выборка – набор векторов, каждый из которых снабжен меткой $m \in \{M_+, M_-\}$, обозначающей класс, к которому принадлежит этот вектор.

Задача бинарной классификации состоит в том, чтобы для вновь поступившего на исследование вектора $\rho = (\rho^1, \rho^2, ..., \rho^K)$ определить класс, к которому он принадлежит, то есть значение m. Для её решения будем использовать метод опорных векторов (SVM). Этот метод основан на построении в K-мерном пространстве $(K-1)$-мерной гиперплоскости, разделяющей объекты классов M_+ и M_-. В зависимости от расположения вектора ρ относительно этой гиперплоскости, выполняется его отнесение к одному из классов.

Возможны следующие результаты интерпретации:

- $D(\rho_{i,j}) = D(\rho_{j,i}) = 0$. В этом случае оба документа признаются оригинальными;
- $D(\rho_{i,j}) \neq D(\rho_{j,i})$. Один из документов является оригиналом, а второй – дублем;
- $D(\rho_{i,j}) = D(\rho_{j,i}) = 1$. Оба документа являются дублями друг относительно друга. То есть, ни один из них не содержит оригинальных данных относительно другого.

На основе полученных результатов принимается решение о дальнейших действиях в отношении документов. Так, в разработанной системе решалась задача проверки документов в момент их поступления от источника, при этом загружаемые сообщения сравнивались на предмет дублирования с документами, уже загруженными в базу. Поэтому интерес представлял лишь один из результатов интерпретации – является ли загружаемый документ дубликатом ранее полученного сообщения. Однако при обработке готовой коллекции документов с целью обнаружения и устранения дубликатов важно выявить все пары сообщений, в которых имеет место дублирование, для чего требуется использовать оба результата.

6.2 Метод предварительного отбора кандидатов

Предложенный метод выявления нечетких дубликатов имеет существенный недостаток – высокую вычислительную сложность. Каждое новое сообщение подвергается сравнению со всеми ранее загруженными, и при каждом сравнении выполняется расчет близости документов по множеству критериев. Однако очевидно, что в большинстве случаев в таком тщательном анализе нет необходимости – сильно различающиеся по тексту документы с высокой вероятностью различны и по содержанию. Следовательно, нужно исключать из рассмотрения те из ранее загруженных новостей, которые слишком сильно отличаются от текущей.

С этой целью вышеописанный метод предваряется процедурой отбора документов, дубликатом которых может быть текущая новость (то есть, отбора кандидатов на роль оригинала этой новости). Эта процедура, по сути, также решает задачу обнаружения дубликатов, причем основными требованиями, предъявляемыми к ней, являются минимальная вычислительная сложность и максимальная полнота (поскольку отброшенные из числа кандидатов документы далее рассматриваться не будут).

В качестве такой процедуры рассматривались представленные в работе [12] приближенные методы обнаружения дубликатов, имеющие высокую производительность. Ввиду наличия набора взвешенных слов, было решено использовать для описания сообщения сигнатуру, представляющую собой строку, состоящую из сцепленных в алфавитном порядке нескольких наиболее «тяжелых» слов документа. При этом процедура отбора кандидатов заключается в выборе из ранее загруженных документов тех, которые имеют такую же сигнатуру, как и текущая новость. Такие пары документов с совпадающими сигнатурами должны быть подвергнуты проверке основным методом, представленным в предыдущем подразделе. В работе [12] предложено использовать сигнатуры из 6 слов, но это приводит к низкому значению полноты (0.54). В целях получения высокой полноты, было решено сократить количество слов, составляющих сигнатуру, до двух.

7 Экспериментальная проверка метода

В рамках данной работы были проведены эксперименты, направленные на анализ качества работы разработанной подсистемы обнаружения дубликатов, реализующей предложенный метод. Все эксперименты проводились на ПЭВМ со следующими основными параметрами: процессор Intel Core 2 Duo 2,2 ГГц, объем ОЗУ 2 Гб.

Целью первого эксперимента была оценка качества метода предварительного отбора потенциальных дубликатов на примере анализа общественно-политических новостей. Тестирование производилось в течение суток. В качестве входных данных использовались 1502 документа, извлеченных с 20 новостных сайтов. Каждый из документов подвергался сравнению с 10293 загруженными ранее сообщениями. В общей сложности было выполнено 16 586 310 сравнений документов, при этом 259 пар были отобраны для проверки основным методом.

Таблица 1. Оценка качества метода отбора кандидатов

	N_p	N_{or}	N_{dup}
Всего	16 586 310	16 586 121	189
Прошли отбор	259	108	151
Отброшено	16 586 051	16 586 013	38

Где N_p – общее количество пар, N_{or} – число пар, элементы которых не дублируют друг друга, и N_{dup} – количество пар документов-дубликатов.

Из 189 пар дубликатов отбор прошла 151 пара (80%). Таким образом, использование сигнатуры из двух слов позволяет увеличить полноту по сравнению с шестисловными сигнатурами, однако добиться полноты, близкой к 100%, не удалось. Метод часто отбрасывает дубликаты в случаях, когда один из документов является урезанной копией другого – отсутствие нескольких параграфов значительно влияет на веса слов. Анализ результатов экспериментов показывает необходимость доработки метода предварительного отбора.

Отброшенные пары не используются для обучения и тестирования основного метода, однако было обнаружено, что 38 ошибочно отброшенных пар дубликатов по своим характеристикам близки к тем 151, которые прошли отбор. Таким образом, недостаточная полнота метода предварительного отбора ведет к появлению в коллекции большего количества дублирующихся сообщений, но не снижает качество обучения основного метода.

Среди всех 259 пар, прошедших отбор, дубликаты составляют 58%. Столь низкая точность метода доказывает необходимость дополнительного анализа отобранных пар. При этом метод продемонстрировал высокую эффективность (под эффективностью понимается отношение количества пар, отброшенных на этапе предварительного отбора, к общему числу пар): из 16 586 310 пар документов было отброшено 16 586 051 (0,99998%).

В рамках второго эксперимента выполнялась оценка качества основного метода. Для тестирования использовались 26 036 документов, извлеченных с 20 новостных сайтов. С помощью метода фильтрации было отобрано 2650 пар-кандидатов, каждая из которых была проанализирована экспертами на предмет наличия дублирования. Часть пар использовалась для обучения, на остальных выполнялось тестирование метода. Целью эксперимента было определение зависимости показателей качества (точности, полноты и *F*-меры) от мощности обучающей выборки и от учитываемых критериев.

Полученные зависимости представлены на рис. 3 (*а* – при использовании только близости составов слов, *б* – при использовании только схожести параграфов, *в* – при использовании всех критериев, приведенных в разделе 5).

Как видно из рисунка, при использовании 400 обучающих примеров происходит насыщение, и с дальнейшим увеличением обучающей выборки качество работы метода не улучшается. Таким образом, для обучения системы достаточно 400 пар документов, размеченных экспертами.

Проведенный эксперимент доказывает целесообразность многокритериального сравнения документов: при использовании всех критериев достигаются более высокие показатели качества (F-мера в зоне насыщения равна 0,82), чем при анализе документов только с точки зрения слов (0,67) или параграфов (0,64).

При анализе результатов эксперимента было выявлено несколько факторов, негативно сказывающихся на качестве. Одним из них является человеческий фактор: каждый эксперт, принимавший участие в подготовке обучающей и тестовой выборок, имеет свое представление о том, какие документы являются информативно необходимыми, а какие – нет, в результате чего возникают конфликты в суждениях экспертов. Также эксперимент показал, что система не может обнаруживать дублирование в случае переписывания оригинального текста без изменения его содержания (рерайтинга), что говорит о необходимости доработки модели для обнаружения такого рода дублирования. Наконец, при проведении эксперимента была выполнена попытка настройки единой модели распознавания для всех документов, загружаемых с новостных сайтов. Но эти документы принадлежат различным предметным областям – среди общественно-политических новостей попадаются экономические, спортивные, юридические. Для повышения качества работы эти документы должны анализироваться с использованием специализированных моделей распознавания.

Проведенный эксперимент также показал, что среднее время, затрачиваемое на анализ пары документов основным методом, составляет 5 мс. С учетом высокой эффективности метода предварительной фильтрации это означает, что система способна обрабатывать 10 000–50 000 документов в час (в зависимости от количества загруженных ранее сообщений, с которыми требуется сравнивать новые документы).

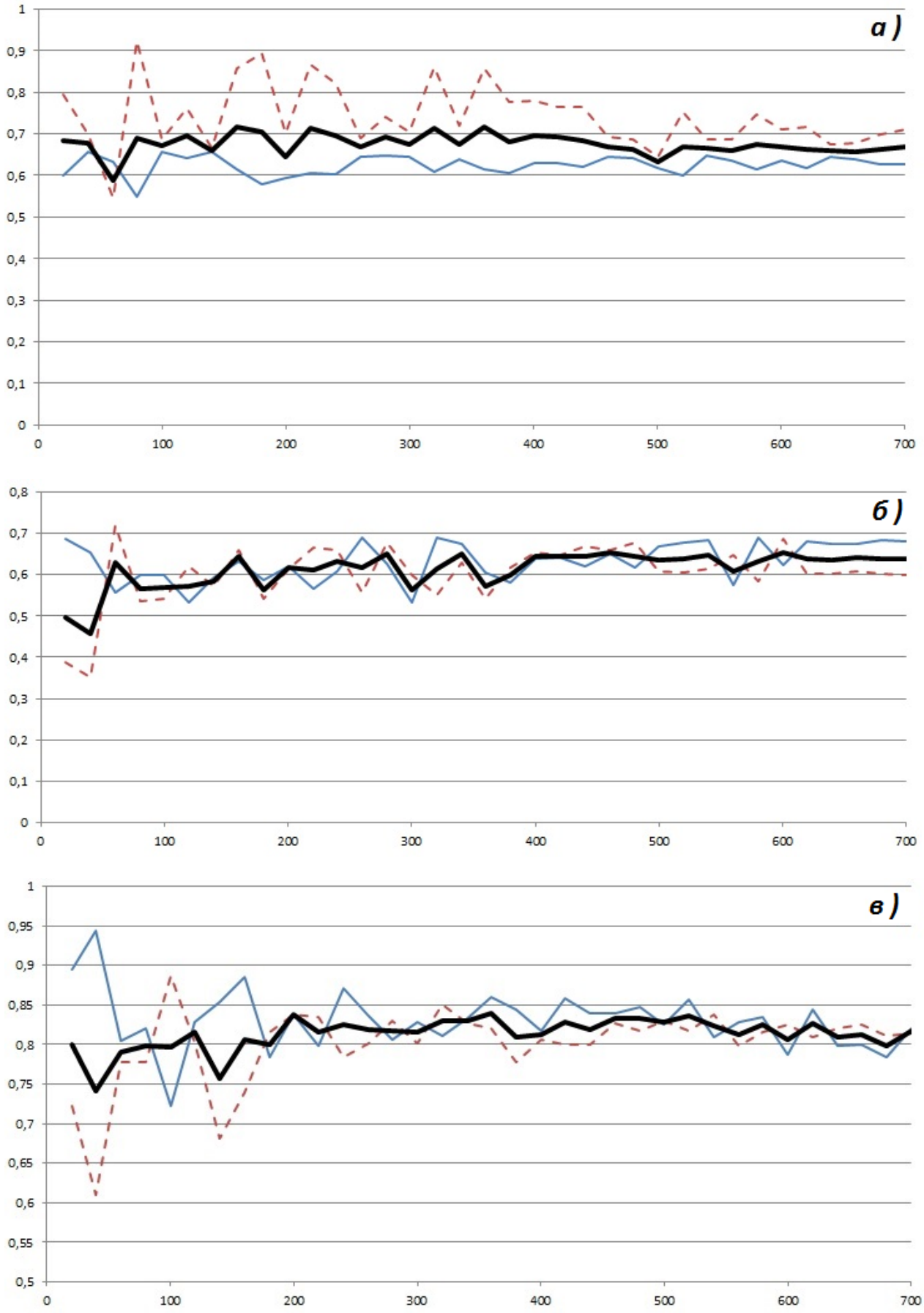

Рис. 3. Зависимость точности (тонкая сплошная линия), полноты (тонкая пунктирная линия) и *F*-меры (жирная линия) от мощности обучающей выборки

8 Направления дальнейших исследований

Помимо обнаружения и устранения дубликатов, предлагаемый метод может быть использован для решения других задач интеллектуального анализа текстов. При соответствующей настройке набора учитываемых критериев и порога близости документов, необходимого для вынесения решения о наличии дублирования, метод может быть применен для решения общей задачи обнаружения документов, близких по содержанию к заданному. Это позволит, в частности, выполнять формирование подборок тематически близких документов, а также сообщений, которые с большой долей вероятности связанны с каким-то общим событием. Таким образом, имеется возможность использования разработанного метода для решения задачи динамической кластеризации коллекции документов.

Еще одним перспективным направлением развития метода является снабжение его возможностью не только обнаружения наличия или отсутствия дублирования, но и выделения в тексте близких по содержанию сообщений фрагментов с оригинальной (недублированной) информацией.

9 Заключение

В работе предложен метод обнаружения и устранения нечеткого дублирования в потоке текстовых сообщений. В его основе лежит отнесение пар документов к классу «дубликатов» или «недубликатов» с помощью метода опорных векторов.

Предлагаемый метод обладает высокой гибкостью благодаря возможности его настройки для обработки сообщений из различных предметных областей. Это достигается посредством включения в модель документа компонентов, отражающих критерии, которыми руководствуются эксперты при анализе текстовых коллекций вручную.

Для обеспечения низкой вычислительной сложности предложена процедура отбора пар-кандидатов на основе сравнения числовых сигнатур документов. Это позволяет применять основной метод лишь к документам, прошедшим отбор.

Представленный метод был апробирован при решении задачи анализа потока текстовых сообщений, загружаемых из открытых интернет-источников, с целью устранения документов, являющихся дубликатами ранее загруженных материалов.

Литература

[1] M. Bilenko, R.J. Mooney. Adaptive Duplicate Detection Using Learnable String Similarity Measures. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003), Washington DC, pp. 39–48, August, 2003.

[2] A. Broder. Algorithms for duplicate documents. http://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/Princeton.pdf

[3] Damerau, F. 1964. A technique for computer detection and correction of spelling errors. Communications of the ACM 7, 3 (1964), 171–176.

[4] D. Fetterly, M. Manasse, M. Najork. A Large-Scale Study of the Evolution of Web Pages, WWW2003, May 20–24, 2003, Budapest, Hungary.

[5] U. Manber. Finding Similar Files in a Large File System. Winter USENIX Technical Conference, 1994.

[6] Robertson S., Walker S., Jones S., Hancock M.-Beaulieu, M. Gatford. Okapi at trec-3. The Third Text REtrieval Conference (TREC-3), 1995.

[7] Андреев А.М., Березкин Д.В., Симаков К.В. Модель извлечения фактов из естественно-языковых текстов и метод ее обучения // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 8-й Всероссийской научной конференции (RCDL'2006). – Суздаль, 2006.

[8] Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 10-й Всероссийской научной конференции (RCDL'2008). – Дубна, 2008. – С. 220–229.

[9] Андреев А.М., Березкин Д.В., Козлов И.А., Симаков К.В. Метод обнаружения изменений структуры веб-сайтов в системе сбора новостной информации // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 14-й Всероссийской научной конференции (RCDL-2012). – Переславль-Залесский, 2012. – С. 124–133.

[10] Блох М.Я. Теоретические основы грамматики : учебник. – 2-е изд., исправл. – М. : Высш. шк., 2000. – 160 с.

[11] Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие. – М. : МИЭМ, 2011. – 272 с.

[12] Зеленков Ю.Г, Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 9-й Всероссийской научной конференции (RCDL'2007). – Переславль-Залесский, 2007. – С. 166–174.

[13] Князева А.А., Турчановский И.Ю., Колобов О.С. Выявление дубликатов в библиографических базах данных // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 15-й Всероссийской научной конференции (RCDL2013). – Ярославль, 2013. – С. 276–282.

[14] Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю. Подход к выявлению дублирования сообщений в новостных информационных потоках // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 8-ой Всероссийской научной конференции (RCDL2006). – Суздаль, 2006.

[15] Лифшиц Ю. Метод опорных векторов. Курс лекций «Алгоритмы для Интернета», 2006.

[16] Никконен А.Ю. Устранение избыточности и дублирования сюжетов новостных сообщений // Сборник работ участников конкурса «Интернет-математика 2007».

[17] Шевелёв О.Г. Методы автоматической классификации текстов на естественном языке : учеб. пособие. – Томск : ТМЛ-Пресс, 2007. – 144 с.

The Method of Detecting Duplicates in a Stream of Text Documents

A. Andreev, D. Berezkin, I. Kozlov, K. Simakov

The problem of duplicate documents elimination from a stream of text messages is considered. A multicriterion model of text document is given. Criteria are chosen to properly represent documents from different domains. An approach for duplicates detection based on binary classification is proposed. A method of candidates preliminary filtration is proposed in order to reduce the computational complexity of the approach.