Экспериментальное исследование методов выявления нечетких дубликатов научных публикаций

Н. В. Дербенев, Д. А. Козлюк, В. В. Никитин, В. О. Толчеев
nicvic@mail.ru, {dmitry.kozliuk,vadimirtoej}@gmail.com, tolcheevvo@mail.ru
НИУ «МЭИ», Россия, Москва, ул. Красноказарменная, 14

Рассмотрены известные методы, в том числе авторский — обобщенный коэффициент ассоциативности (ОКА), для выявления нечетких дубликатов в научных публикациях и критерии эффективности их работы. Сформулирован целевой критерий работы методов, сочетающий требования к полноте и к точности. Составлена выборка пар документов, представленных библиографическими описаниями (заголовками и аннотациями), получены экспертные оценки схожести документов в парах. Проведены эксперименты по установлению наибольшей точности результатов различных методов при ограничении на полноту. Для коэффициента ассоциативности Джаккарда и ОКА, продемонстрировавших наилучшие результаты, предложены и апробированы способы повышения точности до 74 % при полноте 90 %. Результаты проверены путем анализа полнотекстовых описаний части документов исследуемой выборки, доступных публично.

Ключевые слова: анализ текстовой информации; нечеткие дубликаты; наукометрия

Experimental Research of Near-Duplicate Detection Methods for Scientific Papers

N. V. Derbenev, D. A. Kozliuk, V. V. Nikitin, V. O. Tolcheev
Moscow Power Engineering Institute, Krasnokazarmennaya 14, Moscow, Russia

Near-duplicate detection problem focuses on determining pairs of semantically equivalent documents which differ syntactically. For scientific papers, the case in question corresponds either to plagiarism among different authors or to a single author publishing "cloned" papers to achieve higher citation rank. Given a set of document pairs with *a priori* expert opinions, efficiency of any near-duplicate detection method can be measured by the number of correct near-duplicate detections (recall) and false-positive detection count (precision). These metrics cannot be maximized simultaneously — complex criterion are used.

Improvements can be made by choosing a method with the highest precision for a given recall and then preprocessing documents in favor of specifics of the selected algorithm. An efficiency criteria limiting both recall and precision is proposed. Then, a sample set of title and abstract pairs (publically available bibliographic descriptions) is formed and expert assessments are acquired for them. After that, the sample set is used to evaluate performance of various known near-duplicate detection methods subjected to the criteria proposed.

Precision of 74% at 90% recall for Jaccard and generalized similarity coefficients appeared to be reachable by removing frequent words of authors' vocabulary from documents' abstracts. Generalized Similarity Coefficient (GSC) method was introduced in the authors' previous work. Along with Winnowing, GSC scored best in method comparison without preprocessing. The results were checked by examining a subset of documents with full texts available (about 20 % of sample set). Verification confirmed high sustained precision by revealing documents with near-duplicate titles and abstracts to have identical content.

Keywords: information retrieval; near-duplicates detection; scientometrics

Введение

В настоящее время программно-алгоритмические средства, ориентированные на обнаружение неоригинальных публикаций, способны решать важные практические проблемы: устранять одинаковые тексты из выдачи поисковых систем, идентифицировать совпадающие web-страницы, исключать повторяющуюся информацию из баз данных, предотвращать публикацию идентичных статей в научных журналах, обнаруживать заимствование чужих результатов и идей.

Применение таких средств представляется эффективным для анализа массивов научной информации (статей, докладов на конференциях, диссертаций). В основе современного порядка финансирования научных работ лежат наукометрические показатели. Стремление достичь их закономерно приводит к росту публикационной активности ученых, сопровождающемуся снижением качества научных статей, распространением неоригинальных публикаций и заимствованием фрагментов. Данная работа посвящена проблеме обнаружения дубликатов и нечетких дубликатов (очень похожих по терминологическому составу документов) в массивах научных публикаций.

Отметим, что полные дубликаты могут быть достаточно просто обнаружены с помощью специализированного программного обеспечения. Основные сложности возникают при идентификации нечетких дубликатов. Эта задача лишь частично решается применением автоматизированных средств. Для выработки окончательного суждения о статусе документа (уникальный или неуникальный) необходимо задействовать экспертов.

В системах обработки и анализа информации используется два типа текстов: полнотекстовые документы и библиографические описания, которые содержат сведения об авторах, названия, аннотации, ключевые слова и другие вспомогательные данные. Чаще всего в открытом доступе имеются только библиографические описания. Применение полнотекстовых версий для обнаружения нечетких дубликатов зачастую затрудняется их отсутствием в открытых источниках. Некоторые издания практикуют платное предоставление полных текстов статей. Поэтому, несмотря на сильную усеченность и меньшую информативность библиографических описаний по сравнению с полнотекстовыми версиями, именно по первым целесообразно проводить выявление нечетких дубликатов. Вместе с тем большинство методов, которые применяются для анализа библиографических описаний, могут применяться также для выявления нечетких дубликатов среди полнотекстовых публикаций.

Формирование целевого критерия

В зависимости от специфики используемых методов в работе для описания документа применяется два вида моделей: 1) символьная последовательность, задающая порядок следования слов (символов) и определяющая их позицию в тексте; 2) «мешок слов», в котором не учитываются связи между терминами (словами) и их местоположение:

$$\boldsymbol{X}_j = [x_j^{(1)}, \ldots, x_j^{(M)}]^{\mathrm{T}}, \tag{1}$$

где i — номер термина ($i = 1, \ldots, M$; M — количество терминов после удаления служебных слов), j — номер документа в выборке ($j = 1, \ldots, N$; N — количество документов в выборке). Значение $x_j^{(i)}$ будет существенно зависеть от используемого метода взвешивания термина. В некоторых алгоритмах $x_j^{(i)}$ обозначает не только вес слова, но и позицию i слова в документе.

Для формализации понятия схожести двух статей введем меру близости $\rho(\boldsymbol{X}_j, \boldsymbol{X}_l)$, значения которой изменяются в интервале $[0; 1]$. Мера близости должна равняться единице в случае, если документы $\boldsymbol{X}_j$ и $\boldsymbol{X}_l$ — дубликаты, и стремиться к нулю, если $\boldsymbol{X}_j$ и $\boldsymbol{X}_l$ —

уникальные публикации. Два документа $\boldsymbol{X}_j$ и $\boldsymbol{X}_l$ являются *полными дубликатами,* если мера их близости равна единице, и два документа $\boldsymbol{X}_j$ и $\boldsymbol{X}_l$ являются *нечеткими дубликатами (дублями),* если мера их близости превосходит экспериментально (или экспертно) установленный порог θ ($\rho(\boldsymbol{X}_j, \boldsymbol{X}_l) \geqslant \theta$).

Выбор конкретного значения порога осуществляется, прежде всего, в зависимости от критерия качества (целевого критерия) обнаружения нечетких дубликатов. В задачах выявления нечетких дубликатов целевой критерий обычно основывается на расчете показателя «полнота-точность».

Коэффициент полноты (recall) характеризует долю найденных (с помощью метода автоматического выявления дубликатов) неуникальных публикаций среди их общего количества в выборке (т. е. по сравнению с числом нечетких дубликатов, определенных экспертами):

$$R = \frac{a}{a+c}. \tag{2}$$

Здесь a — количество выявленных (методом) документов выборки, являющихся нечеткими дубликатами; $a+c$ — общее число нечетких дубликатов в выборке (c — количество нечетких дубликатов в выборке, которые не найдены методом).

Коэффициент точности (precision) характеризует долю нечетких дубликатов среди документов, которые определены методом в качестве неуникальных:

$$R = \frac{a}{a+b}. \tag{3}$$

Здесь $a+b$ — общее количество документов, которые идентифицированы методом в качестве неуникальных (b — количество документов, не являющихся, по мнению экспертов, нечеткими дубликатами, но отнесенных к ним методом).

Показателем, объединяющим полноту и точность, является F_1-мера: $F_1 = \frac{2PR}{P+R}$.

Наиболее действенным способом влияния на R и P является изменение порога θ, после которого публикации считаются дубликатами. Чем ближе его значение к нулю, тем больше полнота и меньше точность. С увеличением порога наблюдается обратная зависимость: полнота уменьшается, но при этом увеличивается точность.

Рассмотрим основные варианты целевых критериев, используемых при поиске нечетких дубликатов.

1) Максимальное значение одного из показателей (*полноты* или *точности*) при заданных ограничениях на другой показатель:

$$\exists \theta_* \in [0;1] : P(\theta_*)_* = \max_{\theta \in [0;1]} P(\theta), R(\theta_*) \geqslant C$$

или

$$\exists \theta_* \in [0;1] : R(\theta_*)_* = \max_{\theta \in [0;1]} R(\theta), P(\theta_*) \geqslant C\,.$$

2) Максимальное суммарное значение *полноты* и *точности:*

$$\exists \theta_* \in [0;1] : \forall \theta \in [0;1](R(\theta) + P(\theta)) \leqslant (R(\theta_*) + P(\theta_*))\,.$$

3) Баланс между *полнотой* и *точностью* (близость значений этих параметров):

$$\exists \theta_* \in [0;1] : \forall \theta \in [0;1] R(\theta_*) \approx P(\theta_*)\,.$$

В публикациях по проблеме обнаружения нечетких дубликатов исследователи основное внимание уделяют достижению высокой полноты поиска, которая обычно должна быть не менее 90 % [3, 5]. Это позволяет идентифицировать в анализируемом массиве практически все имеющиеся дубликаты. Вторым по важности показателем является точность поиска, которая должна быть максимально возможной при введенном ограничении на полноту. Таким образом, целевой критерий в данной работе имеет вид: достичь наибольшей точности обнаружения нечетких дубликатов при условии, что полнота составляет не менее 90 %.

Учитывая взаимообратный характер изменения полноты и точности, достижение выше сформулированного критерия возможно, прежде всего, за счет использования дополнительной лексической информации об используемых терминах, частоте их появления, местоположении и т. п.

Краткий обзор известных методов выявления нечетких дубликатов

Несмотря на интенсивную разработку методов обнаружения нечетких дубликатов, на настоящий момент не удалось создать универсальный подход, обеспечивающий наилучшие показатели полноты и точности на различных выборках. Сложности разработки обусловлены спецификой обработки текстовой информации: отсутствием механизмов логико-математического описания смысла излагаемого материала, большой размерностью и трудоемкостью задачи, малым размером выборок и т. п.

К настоящему времени опубликовано большое число теоретических и экспериментальных исследований, направленных на изучение преимуществ и ограничений различных процедур выявления нечетких дубликатов [1, 2, 3, 4]. Имеются достаточно полные обзоры, в которых дается описание алгоритмов и приводятся результаты их сравнительного анализа [5, 6].

В большинстве публикаций принято разделять известные процедуры на два класса: синтаксические методы (анализ последовательностей, состоящих из символов, слов или предложений) и лексические методы (анализ информативных терминов). Вместе с тем, представляется целесообразным использовать более детализированные систематизации. Одна из таких систематизаций объединяет методы по принципу общего подхода к принятию решений при выявлении нечетких дубликатов: специализированные расстояния *(расстояние Хэмминга, Левенштейна, Дамерау-Левенштейна, Джаро, Джаро-Винклера)*, шинглы и их модификации *(супершинглы, мегашинглы, Winnowing, SpotSigs)*, процедуры на основе расчета весов терминов *(метод опорных слов, I-match)*, меры близости *(коэффициенты ассоциативности, косинусоидальная мера, двухуровневая мера совпадения Монге-Элкана (Monge-Elkan matching scheme), двухуровневая функция сравнения на основе «мягкого» tfidf-взвешивания)* [3, 5, 7, 8, 9, 11].

Как справедливо указывается в литературе, выбор конкретного метода в существенной степени зависит от цели разработки, особенностей предметной области, исходной информации и имеющихся ограничений. Применительно к проблеме, рассматриваемой в данной работе, при выборе методов для экспериментальных исследований использовались следующие критерии: эффективность обработки коротких документов, чувствительность к операциям редактирования (вставка, замена, удаление терминов), трудоемкость. С этих позиций были отобраны следующие хорошо известные в литературе процедуры: метод шинглов, метод Winnowing, коэффициент ассоциативности Джаккарда, расстояние Джаро-Винклера, нормированное расстояние Левенштейна, а также обобщенный коэф-

фициент ассоциативности (ОКА), предложенный и исследованный в нашей предыдущей публикации [10].

Учитывая, что большинство выбранных методов широко известны и описаны в специализированной литературе [1, 3, 4, 11], остановимся более подробно на процедуре ОКА.

Обобщенный коэффициент ассоциативности документов рассчитывается как линейная комбинация коэффициентов ассоциативности между названиями и между аннотациями. Формулы для расчета выбираются отдельно из нескольких вариантов в зависимости от схожести терминологического состава названий или аннотаций соответственно. Набор формул расчета для выбора обоснован экспериментами, в которых проводился сравнительный анализ известных коэффициентов ассоциативности и исследовались их возможные комбинации.

Введем следующие обозначения: A_{title} и A_{abstract} — число совпавших терминов соответственно в названиях и аннотациях двух анализируемых документов; B_{title} и B_{abstract} — число терминов в названиях и аннотациях, имеющихся в первом документе и отсутствующих во втором; C_{title} и C_{abstract} — число терминов в названиях и аннотациях, имеющихся во втором документе и отсутствующих в первом. Для расчета ОКА предложена следующая формула [10]:

$$\text{ОКА} = \frac{1}{2}(K_1 + K_2) = \\ = \frac{1}{2}\{\frac{A_{\text{title}}}{\max(A_{\text{title}}; B_{\text{title}}; C_{\text{title}})} + \min(\frac{A_{\text{abstract}}}{A_{\text{abstract}} + B_{\text{abstract}}}; \frac{A_{\text{abstract}}}{A_{\text{abstract}} + C_{\text{abstract}}})\} \quad (4)$$

Коэффициент K_1 вычисляется только по терминам, встречающимся в названиях статей ($K_1 \in [0; 1]$); коэффициент K_2 рассчитывается только по терминам из аннотаций ($K_2 \in [0; 1]$). Усреднение суммы коэффициентов и приводит значения ОКА в диапазон $[0; 1]$.

В данной работе процедура выявления нечетких дубликатов осуществлялась в несколько этапов:

1) Составление выборок, состоящих из пар документов одного автора.
2) Экспертная оценка пар на наличие нечетких дубликатов.
3) Выявление из выборки нечетких дубликатов с помощью вышеуказанных методов.
4) Оценка полноты и точности методов. Определение соответствия требованию целевого критерия.
5) Улучшение показателя «полнота-точность» путем учета дополнительных характеристик текстовых документов.

Экспериментальные исследования методов выявления нечетких дубликатов

Для создания специализированного документального массива, в котором осуществляется поиск нечетких дубликатов, использовалась российская научная электронная библиотека eLibrary.ru. С помощью авторского указателя было найдено 1070 авторов, которые имеют не менее пяти публикаций (включая соавторство) в области автоматики и вычислительной техники с аннотациями на русском языке. В итоге был создан документальный массив, состоящий из 7257 библиографических описаний. Для анализа выбирались наиболее тематически близкие авторские работы (для оценки близости использовался коэффициент ассоциативности Джаккарда больше $0,3$ [12]). Полные дубликаты, содержащиеся в eLibrary.ru (например, из-за опечаток при указании страниц,

Таблица 1. Сравнительные результаты работы методов выявления нечетких дубликатов

Метод	Пороговое значение	Полнота	Точность
Коэффициент ассоциативности Джаккарда	0,54	0,908	0,712
Шинглы (длина 2 слова)	0,38	0,908	0,718
Жаро	0,74	0,908	0,701
Жаро-Винклер	0,74	0,908	0,701
Расстояние Левенштейна	0,51	0,939	0,697
Winnowing ($k = 2, t = 4$)	0,35	0,917	0,720
OKA	0,57	0,908	0,724

фамилий авторов и т. п.), из рассмотрения удалялись. Далее случайным образом было отобрано 150 пар библиографических документов (общий размер выборки — 300 документов). Библиографические описания, вошедшие в выборку, были проанализированы тремя экспертами. Статьи относились к нечетким дубликатам, если все эксперты принимали согласованное решение (в противном случае тексты считались оригинальными). Всего нечеткими дубликатами эксперты сочли 98 пар (65 %). Выборка размещена по адресу: `http://uii.mpei.ru?term=49`.

Задача экспериментальных исследований заключалась в определении с помощью автоматизированных средств тех нечетких дубликатов, которые ранее были выявлены экспертами (с учетом требований целевого критерия).

Для проведения автоматизированного анализа и выявления нечетких дубликатов был разработан специализированный программный комплекс, в котором предусмотрена обработка документов, заданных как библиографическими описаниями, так и полными текстами. В комплексе реализованы все вышеуказанные методы обнаружения дублей и основные процедуры предварительной обработки текстовых данных.

Экспериментальные исследования были организованы следующим образом: на первом этапе рассчитывались показатели полнота-точность анализируемых методов (метод шинглов, метод Winnowing, коэффициент ассоциативности Джаккарда, расстояние Джаро-Винклера, расстояние Левенштейна, OKA) и определялось их соответствие целевому критерию, на втором этапе изучалось влияние различных характеристик текстовых документов на значения полноты и точности. Отметим, что в наших исследованиях использовались стандартные методы Левенштейна и Джаро-Винклера, не предусматривающие учета местоположения термина (в названии или в аннотации). Структура текста использовалась только при расчете OKA (именно эта идея была положена в основу разработки данной процедуры). Вместе с тем, в перспективе, планируется провести разработку модификаций методов Левенштейна и Джаро-Винклера, учитывающих структуру текста и исследовать влияние этой информации на целевой критерий.

В таблице представлены значения полноты и точности различных методов на исследуемой выборке для случая предварительного удаления стоп-слов и выполнения стемминга. Для метода шинглов и Winnowing результаты приведены для значений настраиваемых параметров, которым соответствуют наилучшие полнота и точность (размер шингла равен 2 слова; размер к-граммы $k = 2$, значение шумового порога $t = 4$). На основе проведенного исследования можно сделать вывод, что все анализируемые методы удовлетворяют требованию целевого критерия. При этом наилучших результатов достигают два метода: Winnowing и OKA.

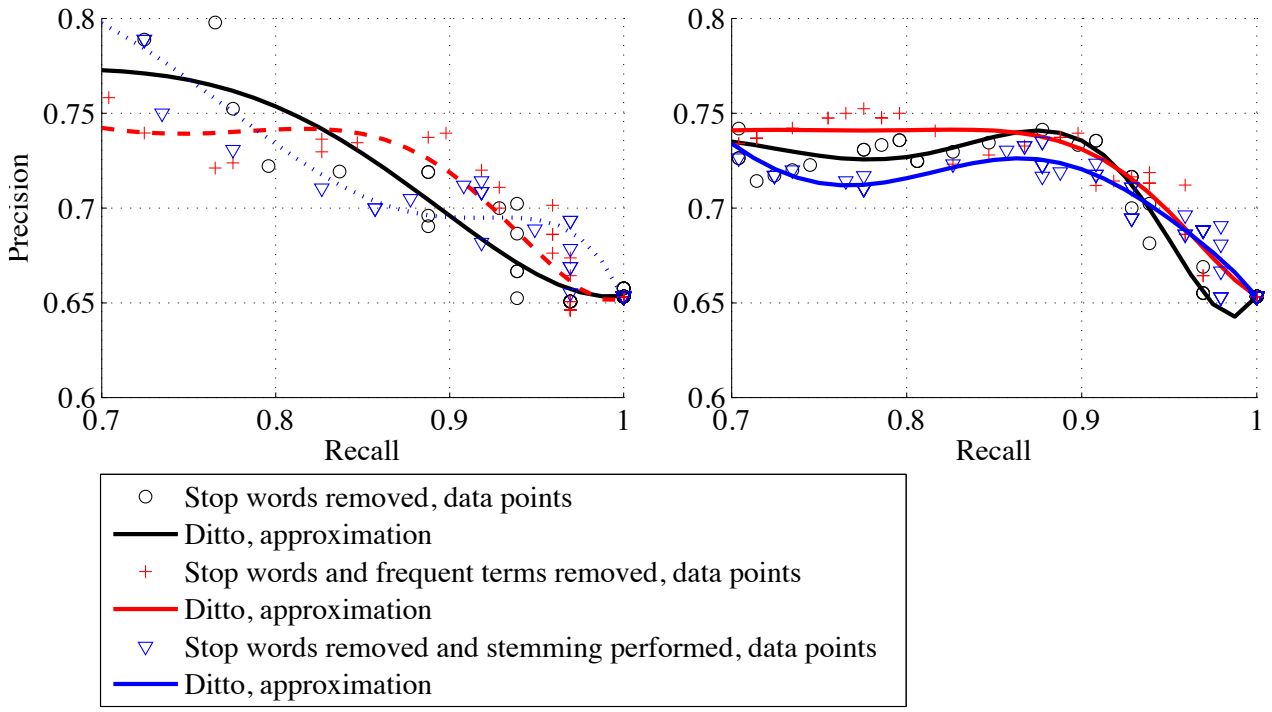

Рис. 1. Графики «полнота–точность» для коэффициента ассоциативности Джаккарда (слева) и для обобщенного коэффициента ассоциативности (справа)

Графики «полнота–точность» методов выявления нечетких дубликатов при различных видах предварительной обработки представлены на рис. 1. Конкретные значения полноты и точности зависят от выборки, поэтому важен качественный характер апроксимаций, главным образом, радиус их кривизны на интересующем участке. Рационально выбирать такой метод и сочетание параметров, чтобы точность не только была достаточной при заданной полноте, но и при изменении полноты не менялась бы резко (то есть, достигнутая точность должна быть устойчива к возмущениям, вносимым особенностями выборки).

На втором этапе проводилась оценка целесообразности использования для более качественного определения нечетких дубликатов следующих характеристик текста:

— стоп-слов, встречающихся в двух сравниваемых документах с одинаковой частотой (наличие одних и тех же служебных слов: союзов, предлогов и т. п.);
— служебных символов, включая знаки препинания и сокращения.

Проведенные исследования подтвердили известные в литературе результаты, которые свидетельствуют о том, что учет вышеуказанных характеристик не позволяет значимо улучшить показатель «полнота-точность» выявления нечетких дубликатов [4].

Нами также изучалась возможность повышения полноты и точности за счет создания специального терминологического словаря автора. Такой словарь составляется из всех информативных слов, которые содержатся в пяти авторских публикациях, включенных в сформированный документальный массив. Аналогичные работы проводились ранее группой авторов при решении задач выявления идентичных строк в базах данных при сравнении наименований и адресов юридических лиц [8, 9], а также в [4] для анализа web-документов.

Анализ частоты встречаемости терминов из словаря позволяет определить степень «общности–специфичности» различных слов в статьях автора. Наше предположение за-

ключается в том, что исключение высокочастотных слов из расчета коэффициента ассоциативности Джаккарда и ОКА, способно увеличить точность выявления нечетких дубликатов без ухудшения значения полноты. Эксперименты показали, что без учета высокочастотных общих слов точность коэффициента ассоциативности Джаккарда и ОКА возросли и составили соответственно $0,74$ и $0,745$. У остальных методов, кроме расстояния Левенштейна, было отмечено снижение точности, значения которой вышли из пределов, допустимых по целевому критерию. Отметим, что при расчете ОКА высокочастотные слова исключались только из текста аннотации, т. е., при вычислении коэффициента K_2 в формуле (4). В работах [8, 9] при использовании более сложных двухуровневых функций сравнения на основе «мягкого» tfidf-взвешивания также указывается на улучшение показателей полнота-точность.

Таким образом, использование специализированного документального массива, содержащего не только две статьи, которые рассматриваются в качестве кандидатов в нечеткие дубликаты, но и другие работы автора, позволяет улучшить точность выявления неоригинальных публикаций с помощью методов, основанных на коэффициентах ассоциативности. Это достигается за счет предварительной оценки частоты встречаемости терминов в авторском словаре.

Обсуждение результатов экспериментальных исследований

Главный вопрос — насколько можно доверять полученным результатам, полезны ли они при выявлении не гипотетических нечетких дубликатов (по библиографическим описаниям), а конкретных полнотекстовых публикаций, зеркально повторяющих друг друга? Другими словами, можно ли утверждать, что в выборке, по которой проводились исследования, действительно встречаются случаи «недобросовестных» действий автора (авторов) по клонированию в различных журналах совершенно одинакового материала.

Для ответа на этот вопрос в интернете был проведен поиск полнотекстовых версий 300 статей из выборки (при этом необходимо было найти полнотекстовые варианты сразу двух статей автора). Из всех 150-и пар документов, в открытом доступе нам удалось обнаружить только 31 пару полнотекстовых статей (около 20 % от всего объема выборки). С помощью разработанного программного комплекса [13] было проведено сравнение документов методом шинглов. В результате было обнаружено 3 пары полных дубликатов, в которых авторы лишь незначительно изменили название и аннотацию. При этом основные тексты публикаций оказались абсолютно идентичными (эти выводы подтверждены также экспертно). Следует отметить, что библиографические описания для этих пар публикаций были признаны нечеткими дубликатами всеми вышеперечисленными методами.

Процент самоплагиата, который мы получили, достаточно низкий, что может говорить о хорошем качестве статей в области автоматики и вычислительной техники. Следует отметить, что мы рассматривали только статьи с близкими библиографическими описаниями.

Авторы, конечно же, понимают, что делать обобщения и заключения на весьма небольшом статистическом материале крайне не разумно и планируют в ближайшее время подтвердить или опровергнуть полученные оценки на новых выборках.

Отметим также, что разработанные программно-алгоритмические средства, на наш взгляд, уже в настоящее время могут быть эффективно использованы редколлегиями журналов (и, возможно, диссертационными советами) при анализе статей авторов и соискателей. Такой анализ разумно проводить на предварительном этапе по библиографическим описаниям. В случае выявления с помощью автоматизированных методов потенциальных

нечетких дубликатов необходимо реализовывать полнотекстовую проверку и экспертное изучение публикаций.

Литература

[1] *Рубцов Д. Н., Барахнин В. Б.* О возможности борьбы с дубликатами при запросах к разнородным библиографическим источникам // *Тр. 11-й Всеросс. научной конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»*. — Петрозаводск: изд-во ПетрГУ, 2007. С. 293–298.

[2] *Зеленков Ю. Г., Сегалович И. В.* Сравнительный анализ методов определения нечетких дубликатов для Web–документов // *Тр. 9-й Всеросс. научной конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»*. — Переславль-Залесский: Изд-во ИПС РАН, 2007. С. 166–174.

[3] *Chowdhury A., Frieder O., Grossman D., Mccabe C.* Collection statistics for fast duplicate document detection // *ACM Trans. Inform. Syst.*, 2002. Vol. 20, No. 2. P. 171–191.

[4] *Heintze N.* Scalable document fingerprinting // *2nd USENIX Electronic Commerce Workshop Proceedings* , 1996. P. 191–200.

[5] *Косинов Д. И.* Использование статистической информации при выявлении схожих документов // *Сборник «Интернет-математика»*. — Екатеринбург: Изд-во Уральского университета, 2007. С. 84–90.

[6] *Kumar J., Govindarajulu P.* Duplicate and near duplicate documents: A review // *European J. Scientific Research*, 2009. Vol. 32, No. 4. P. 514–527.

[7] *Дербенев Н. В., Толчеев В. О.* Выявление нечетких дубликатов в наукометрическом анализе // *Информационные технологии*, 2011. № 12. С. 24–29.

[8] *Cohen W. W., Ravikumar P., and Fienberg S. E.* A comparison of string distance metrics for name matching tasks, 2003. P. 73—78.

[9] *Bilenko M., Mooney R., Cohen W., Ravikumar P., and Fienberg S.* Adaptive name matching in information integration. *IEEE Intelligent Systems,* 2003. Vol. 18. P. 16—23.

[10] *Broder A., Glassman S., Manasse M., Zweig G.* Syntactic clustering of the Web. *6th World Wide Web Conference (International)*, 1997. P. 393–404.

[11] *Дербенев Н. В., Толчеев В. О.* Разработка метода выявления нечетких дубликатов по библиографическим описаниям // *Тр. Междунар. конф. «Интеллектуализация обработки информации»*. — Черногория, Будва. М.: Изд-во «ТОРУС ПРЕСС», 2012. С. 613–616.

[12] *Дербенев Н. В., Толчеев В. О.* Сравнительный анализ коэффициентов ассоциативности для выявления нечетких дубликатов текстовых документов // *Тр. 18-й Междунар. научно-технич. конф. «Информационные средства и технологии»*. — М.: Изд-во МЭИ, 2010. С. 266–270.

[13] *Дербенев Н. В., Козлюк Д. А., Никитин В. В., Толчеев В. О.* Разработка программно-алгоритмических средств выявления плагиата в учебных и научных кафедральных работах // *Мат-лы 6-й Всеросс. мультиконф. по проблемам управления — МКПУ-2014*. — Дивноморское, 2014. Т. 1. С. 59–63.

References

[1] *Chowdhury A., Frieder O., Grossman D., Mccabe C.* 2002. Collection statistics for fast duplicate document detection. *ACM Trans. Inform. Syst.* 20(2):171–191.

[2] *Heintze N.* 1996. Scalable document fingerprinting. *2nd USENIX Electronic Commerce Workshop Proceedings* . 191–200.

[3] *Roubtsov D. N., Barakhnin V. B.* 2007. On the possibility of duplicates struggle when performing queries to heterogeneous bibliographic sources. *XI All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" Proceedings.* Petrozavodsk. 293–298.

[4] *Kosinov D.* 2007. Use of statistical parameters in similar documents detection. *"Internet-Mathematics" Digest.* Yekaterinburg. 84–90.

[5] *Zelenkov Yuri. G., Segalovich Ilya V.* 2007. Comparative analysis of near-duplicate detection methods of Web documents . *IX All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" Proceedings.* Pereslavl-Zalessky. 166–174.

[6] *Kumar J., Govindarajulu P.* 2009. Duplicate and near duplicate documents: A review. *European J. Scientific Research* 32(4):514–527.

[7] *Derbenev N. V., Tolcheev V. O.* 2011. Using a method of detecting near duplicates in sciencemetric analysis. *Information Technologies* 12:24–29.

[8] *Cohen W. W., Ravikumar P., and Fienberg S. E.* 2003. A comparison of string distance metrics for name matching tasks. 73–78.

[9] *Bilenko M., Mooney R., Cohen W., Ravikumar P., and Fienberg S.* 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems* 18:16—23.

[10] *Derbenev N. V., Tolcheev V. O.* 2012. Development of a method of near-duplicates detection on the base of bibliographic descriptions. *Conference (International) "Intellectualization of Information Processing" Proceedings.* Montenegro, Budva. 613–616.

[11] *Broder A., Glassman S., Manasse M., Zweig G.* 1997. Syntactic clustering of the Web. *6th World Wide Web Conference (International).* 393–404.

[12] *Derbenev N. V., Tolcheev V. O.* 2010. Comparative analysis of assitiativity coefficients for near-duplicate detection within textual documents. *XVIII Conference (International) "Information Tools and Techniques" Proceedings.* Moscow. 266–270.

[13] *Derbenev N. V., Kozliuk D. A., Nikitin V. V., Tolcheev V. O.* 2014. Development of software for plagiarism detection in academic and scientific papers. *6th All-Russian Multiconference on Control Problems — MCCP-2014 Proceedings.* Divnomorskoye. 1:59–63.