К. Ф. Сафин, М. П. Кузнецов, М. В. Кузнецова, Определение заимствований в тексте без указания источника, *Информ. и её примен.*, 2017, том 11, выпуск 3, 73–79

DOI: https://doi.org/10.14357/19922264170308

ОПРЕДЕЛЕНИЕ ЗАИМСТВОВАНИЙ В ТЕКСТЕ БЕЗ УКАЗАНИЯ ИСТОЧНИКА*

К. Ф. Сафин[1], М. П. Кузнецов[2], М. В. Кузнецова[3]

Аннотация: Для задачи поиска заимствований в тексте существуют два подхода: обнаружение «внешних» и «внутренних» заимствований. При поиске внешних заимствований известен корпус, из которого возможны заимствования. При поиске внутренних заимствований исследуемый текст анализируется изолированно, т. е. возможные источники заимствований неизвестны. Данная работа посвящена поиску внутренних заимствований в тексте. Предполагается, что большая часть текста написана одним автором. Необходимо выделить участки текста, написанные другим автором, если таковые имеются. В работе предлагается алгоритм, строящий статистику сегментов текста, по которой определяется факт зависимости. Эксперимент проводится на коллекции конкурса PAN-2011.

Ключевые слова: обработка естественного языка; детектирование внутренних заимствований; поиск выбросов в статистике

DOI: 10.14357/19922264170308

1 Введение

Текстовые заимствования являются большой проблемой в сфере образования и научных исследований [1]. Развитие сети Интернет, в частности, и информационных технологий, в целом, сделало возможным некорректное заимствование информации.

В задаче обнаружения заимствований существуют два глобальных подхода: выявление «внешних» (external plagiarism detection) и «внутренних» (intrinsic plagiarism detection) заимствований. При поиске внешних заимствований предполагается, что в распоряжении исследователя есть некоторый корпус, из которого возможны заимствования. Таким образом, задача состоит в попарном сравнении участков подозрительного текста и текстов из корпуса заимствований.

Задача поиска «внутренних» заимствований состоит в анализе исключительно подозрительного текста. Алгоритмы должны анализировать стиль письма и выделять характерные признаки, свойственные данному автору.

Алгоритмы разбивают исходный текст на сегменты и сравнивают текст сегмента со всем текстом. Разбиение проводится по предложениям [2, 3], или же определяется окно заданной ширины, согласно которому производится сегментирование текста [4–7]. Выбор меры схожести сегмента со всем текстом или, наоборот, меры различия является ядром алгоритма. Работы [2–4] используют стилистические, синтаксические, лексические характеристики: частотность частей речи, порядок следования частей речи в предложении, пунктуацию, среднюю длину предложения и подобные признаки. Возможно использование символьных n-грамм (чаще других применяют 3-граммы) в качестве признака, а точнее, частот их использования [5, 7].

Метод [8] использует кластеризацию абзацев по частоте встречаемости существительных.

В статье [9] описан алгоритм диаризации текстов, т. е. классификации сегментов текста по авторству, что является обобщением задачи поиска внутренних заимствований.

В 2011 г. был проведен конкурс PAN-2011, посвященный поиску заимствований в текстах. Метод Oberreuter [6], ставший победителем в конкурсе PAN-2011, использует функцию, характеризующую письменный стиль автора. Функция строит вектор частот встречаемости слов во всем документе и в выделенном сегменте. Эти векторы используются для определения величины отклонения сегмента от всего текста. Данный алгоритм показал результат 0,32 по F1-мере. Это демонстрирует тот факт, что алгоритма, решающего данную задачу в большинстве случаев, до сих пор нет.

Предлагаемый алгоритм строит статистическое описание текста, которое используется для нахождения заимствованных сегментов текста. Статисти-

*Работа поддержана РФФИ (проект 16-07-01155).

[1]Московский физико-технический институт; ЗАО «Анти-плагиат», kamil.safin@phystech.edu

[2]ООО «Форексис», mikhail.kuznecov@phystech.edu

[3]Московский физико-технический институт; ЗАО «Анти-плагиат», kuznetsova@ap-team.ru

ка должна удовлетворять следующим условиям: на оригинальных сегментах текста иметь небольшой разброс значений по сравнению со значением на всем тексте, а на заимствованных сегментах иметь значительные отличия.

В работе используются данные конкурса PAN-2011 [10]. Для оценки работы алгоритма определяются микро- и макромеры качества precision и recall (точность и полнота) и затем вычисляется F1-мера как среднее гармоническое для precision и recall. Данная мера представляет качество работы алгоритма.

2 Постановка задачи

Пусть D — коллекция текстовых документов, d — текстовый документ, t_i — сегмент текста ($d = \bigcup t_i$, $d \in D$). Среди сегментов текста t_i необходимо выделить те, значение статистики которых $\sigma(\mathbf{t}_i)$ превосходит некоторый заданный порог значений δ_{susp}.

Описание выборки. В работе используется блок текстовых документов конкурса PAN-2011 [10]. В текстах присутствуют сегменты настоящих, имитированных и искусственных заимствований. Каждый сегмент текста соответственно полностью взят из другого источника, либо заимствованный текст переписан человеком другими словами, либо специально обученный алгоритм строит текст, стараясь повторить стиль автора.

Выборка состоит из 4753 текстов, разделенных на 10 частей, к каждому из текстов прилагается файл с экспертной разметкой заимствованных сегментов.

Для анализа корпуса была собрана подвыборка корпуса, состоящая из 30 документов, которые были просмотрены вручную. Анализ показал, что большая часть документов содержит в себе заимствования, сильно отличающиеся от остального текста по тематике и набору используемых слов. К примеру, в текст по экономике вставляется фрагмент, вырезанный из художественного текста.

Также тексты корпуса были исследованы на то, какая доля заимствований содержится в каждом тексте (отношение длины заимствованных фрагментов к длине текста в символах) и сколько различных фрагментов заимствований присутствует в тексте. На рис. 1 приведены гистограммы результатов.

Как видно, тексты в среднем содержат от 1 до 7 фрагментов заимствований. В большинстве текстов доля заимствований не превышает 4%–5%, что усложняет задачу поиска этих заимствований.

Критерии качества. В экспериментах используются критерии качества, применявшиеся в PAN-2011 [10]. Обозначим за пару (s, d) последовательность символов, помеченную экспертом как заимствование в документе d. $S = \bigcup s_i$ — совокупность всех заимствованных сегментов. За пару (r, d) обозначим последовательность, помеченную алгоритмом как заимствованную. Аналогично $R = \bigcup r_i$ — совокупность всех сегментов, которые алгоритм классифицировал как заимствованные. Рассмотрим меры качества Precision и Recall:

$$\text{Prec}(S, R) = \frac{1}{|R|} \sum_{r_j \in R} \frac{\left| \bigcup\limits_{s_i \in S} (s_i \cap r_j) \right|}{|r_j|};$$

$$\text{Rec}(S, R) = \frac{1}{|S|} \sum_{s_i \in S} \frac{\left| \bigcup\limits_{r_j \in R} (s_i \cap r_j) \right|}{|s_i|}.$$

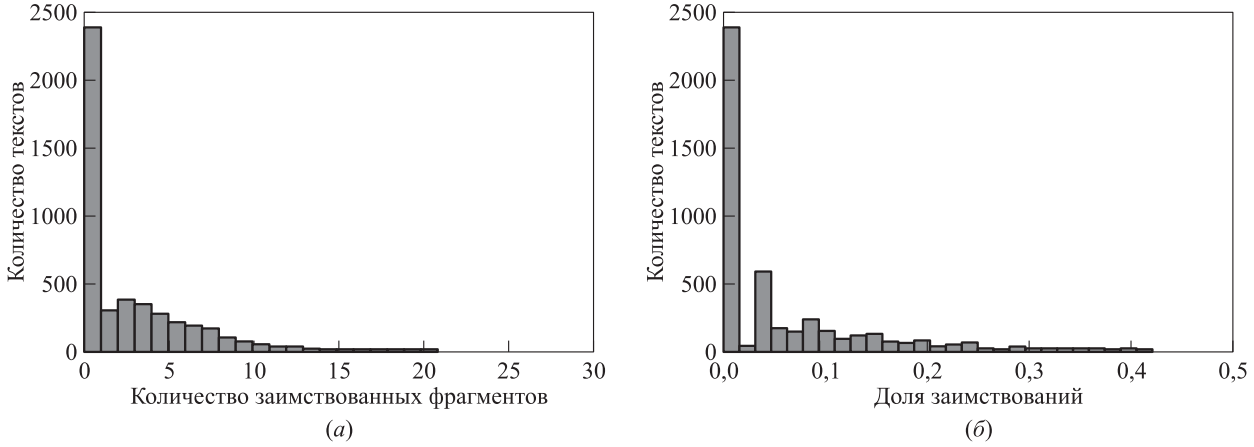

Рис. 1 Распределение текстов по количеству заимствованных фрагментов (*а*) и по доле заимствований (*б*)

Данные величины отражают точность (доля правильного распознавания заимствований по отношению ко всем выделенным сегментам) и полноту (доля правильного распознавания заимствований по отношению ко всем заимствованиям в тексте) работы алгоритма.

Вычисляется F1-мера как среднее гармоническое между Precision и Recall:

$$\mathrm{F1}(S,R) = \frac{\mathrm{Prec}(S,R)\cdot\mathrm{Rec}(S,R)}{\mathrm{Prec}(S,R)+\mathrm{Rec}\,(S,R)}\,.$$

Вычисляется величина гранулярности

$$\mathrm{gran}\,(S,R) = \frac{1}{|S_R|}\sum_{s_i\in S_R}|R_{s_i}|\,,$$

где S_R — множество заимствованных сегментов, обнаруженных алгоритмом; R_s — сегменты, отмеченные алгоритмом, которые детектируют данный сегмент заимствований s:

$$S_R = \{s|s\in S\wedge\exists r\in R : r \text{ detects } s\}\,;$$
$$R_S = \{r|r\in R\wedge r \text{ detects } s\}\,;$$
$$r \text{ detects } s : \text{if } r\cap s\neq\emptyset\,.$$

Таким образом, гранулярность показывает то, насколько мелко алгоритм разбивает заимствованные сегменты текста. Если заимствованные сегменты разделяются алгоритмом на много мелких, то гранулярность будет иметь высокие значения.

По описанным величинам вычисляется итоговая мера качества pladget:

$$\mathrm{pladget}(S,R) = \frac{F1(S,R)}{\log_2(1+\mathrm{gran}(S,R))}\,.$$

Формальная постановка задачи. Для обнаружения заимствований исходный текст d разбивается на сегменты t_i:

$$d = \cup t_i\,.$$

Для каждого сегмента вычисляется вектор признаков $\mathbf{t}_i$ и строится статистика $\sigma(\mathbf{t}_i)$. Затем происходит детектирование выбросов среди значений статистики на основании ее отклонения от среднего значения

$$\sigma_{\mathrm{avr}}(d) = \frac{1}{N}\sum_{i=1}^{N}\sigma(\mathbf{t}_i)\,,$$

где N — число сегментов в тексте. Если отклонение превышает заданный порог δ_{susp}, то сегмент считается заимствованным:

$$|\sigma(\mathbf{t}_i)) - \sigma_{\mathrm{avr}}(d)| > \delta_{\mathrm{susp}}\,.$$

Корпус документов D разбивается на обучающую и тестовую выборки:

$$D = D_{\mathrm{test}}\cup D_{\mathrm{learn}}\,.$$

При обучении параметры алгоритма $\mathbf{w}$ настраиваются таким образом, чтобы улучшить меры качества работы алгоритма. При фиксированном способе разбиения текста мера Granularity не изменяется, так как она зависит от мелкости разбиения. Тогда для увеличения итоговой меры качества Pladget достаточно улучшить F1-меру:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbf{W}}\mathrm{F1}(S,R)\,,$$

т. е. требуется вектор параметров $\mathbf{w} = (l_{\mathrm{segm}}, n, \delta_{\mathrm{susp}})$, максимизирующий F1-меру. Здесь l_{segm} — минимальная длина сегмента; n — ширина окна сглаживания; δ_{susp} — порог выброса. Более подробно параметры описаны в вычислительном эксперименте (см. разд. 5).

3 Базовый эксперимент

Целью базового эксперимента ставилась проверка гипотезы о том, что заимствованные сегменты текста имеют отличные от среднего вектора значения признаков.

В качестве такого признака была выбрана частота встречаемости слов. Каждому слову ставится в соответствие число

$$\mathrm{fr\_class}_w = \log_2\frac{n_{\max}}{n_w}\,, \qquad (1)$$

где $n_{\max}$ — число вхождений наиболее часто употребимого слова в тексте; n_w — частота вхождений слова w в этом предложении.

В качестве основного признака использовались квантили распределения данной величины внутри окна фиксированной ширины.

Обозначим за $m^j = \overline{x^j}$ среднее значение j-го признака для рассматриваемого документа, за r^j — среднеквадратичное отклонение. Тогда нормализованный признак j для сегмента i рассчитывается по формуле:

$$t_i^j = \frac{x^j - m^j}{r^j}\,.$$

За сегменты t_i были выбраны предложения текста. Для каждого предложения t_i строился вектор признаков $\mathbf{t}_i$ и затем подсчитывалось отклонение от усредненного по всему тексту вектора $\mathbf{t}_{\mathrm{avr}}$ в L1-метрике:

$$\sigma(\mathbf{t}_i) = ||\mathbf{t}_i - \mathbf{t}_{\mathrm{avr}}|| = \sum_{j=1}^{l}|t_i^j - t_{\mathrm{avr}}^j|\,. \qquad (2)$$

Эксперимент проводился на одном из текстов конкурсной коллекции PAN-2011.

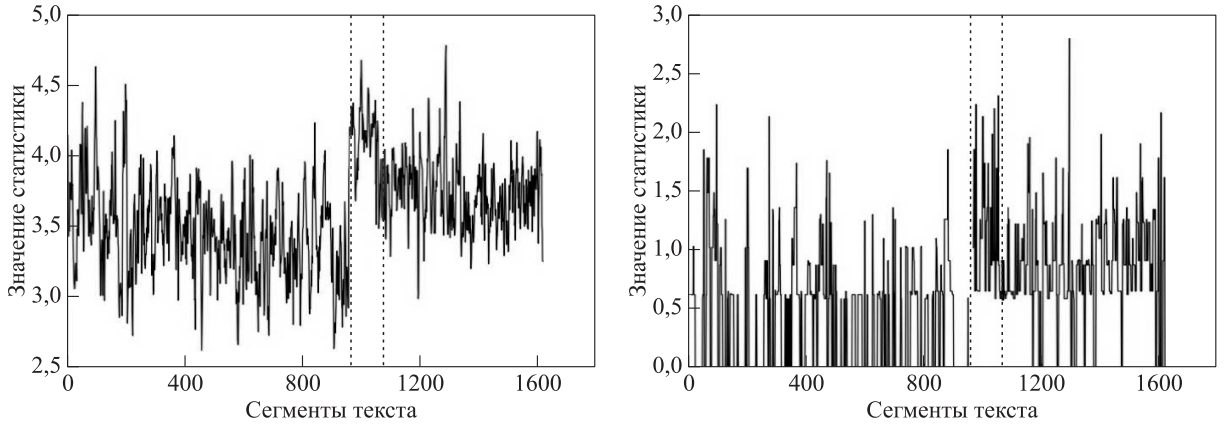

Рис. 2 Отклонение признакового вектора от среднего

На рис. 2 показано отклонение признакового вектора каждого предложения от усредненного вектора. Пунктирными линиями выделены предложения, помеченные экспертом как заимствованные.

Видно, что заимствованные фрагменты имеют характерные выбросы из области средних значений отклонения. Однако некоторые предложения, не являющиеся заимствованными, также сильно отличаются от усредненного признакового вектора. На основании этого можно сделать вывод, что использование только данного признака недостаточно для решения поставленной задачи.

4 Описание алгоритма

Модель. Предлагаемый алгоритм работает с частотными признаками, предоставляющими описание текста. В качестве такого признака выбран признак частоты встречаемости слов, описанный в формуле (1).

Исходный текст подвергается предобработке: удаляются служебные символы, все буквы переводятся в нижний регистр. Также из текста удаляются стоп-слова.

Сегментирование текста. Текст разбивается на предложения. Затем формируется разбиение текста на сегменты t_i: если длина очередного предложения меньше минимальной длины сегмента l_{segm}, к этому предложению добавляется следующее за ним — процесс повторяется, пока длина сегмента t_i не превысит заданную минимальную длину. Минимальная длина сегмента l_{segm} является настраиваемым параметром алгоритма.

Построение статистики и детектирование аномалий. Для каждого сегмента t_i текста строится вектор признаков. Затем строится статистика $\sigma(\mathbf{t}_i)$ на основе отклонения вектора признаков от усредненного по всему тексту вектора (2).

Полученная статистика сглаживается методом скользящего среднего: новые значения статистики $\sigma'(\mathbf{t}_i)$ вычисляются по формуле:

$$\sigma'(\mathbf{t}_i) = \frac{1}{2n+1} \sum_{k=i-n}^{i+n} \sigma(\mathbf{t}_i)\,,$$

где n — ширина сглаживания, которая также является настраиваемым параметром. Значения в крайних точках вычисляются по формулам (N — число сегментов):

$$\sigma'(\mathbf{t}_i) = \frac{1}{i+n+1} \sum_{k=0}^{i+n} \sigma(\mathbf{t}_i)\,;$$

$$\sigma'(\mathbf{t}_i) = \frac{1}{i+n+1} \sum_{k=i-n}^{N} \sigma(\mathbf{t}_i)\,.$$

Полученные значения статистики $\sigma'(\mathbf{t}_i)$ исследуются на выбросы. Если в ряде статистики присутствует аномалия, превышающая заданный порог δ_{susp}, то сегмент t_i, отвечающий этому выбросу, помечается как заимствованный.

Минимальная длина сегмента, ширина окна сглаживания и порог выброса настраиваются на обучающей выборке путем максимизации F1-меры.

5 Вычислительный эксперимент

Алгоритм настраивался на частях 1–5 корпуса PAN-2011 путем максимизации F1-меры. Тестирование проводилось на частях 6–10 корпуса. Оптимальные параметры после настройки: $\hat{l}_{\text{segm}} = 450$; $\hat{n} = 8$; $\hat{\delta}_{\text{susp}} = 0{,}37$.

На рис. 3 и 4 приведены примеры работы алгоритма. Серые участки обозначают заимствованные фрагменты, пунктирными линиями обозначен порог выброса значений стилевой функции.

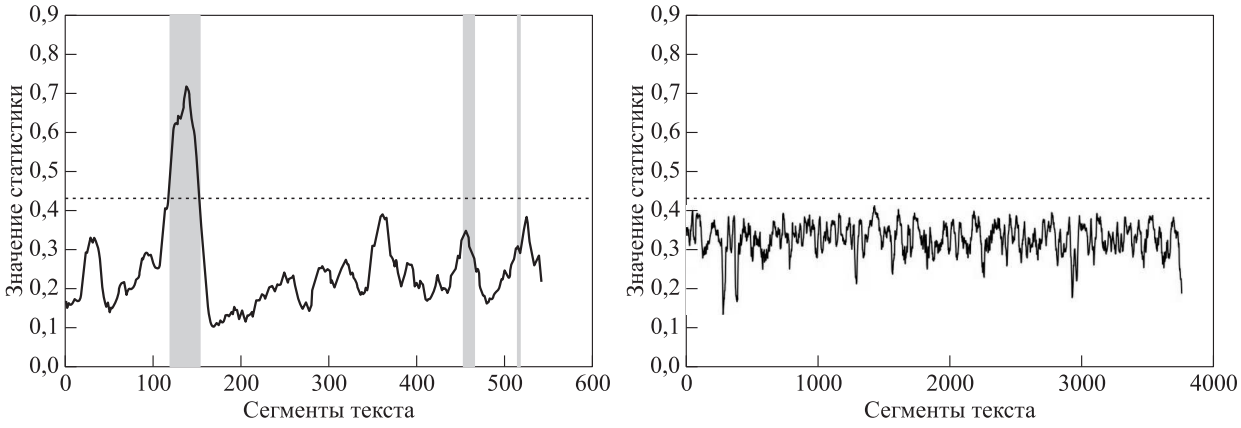

Рис. 3 Результаты на обучающей выборке

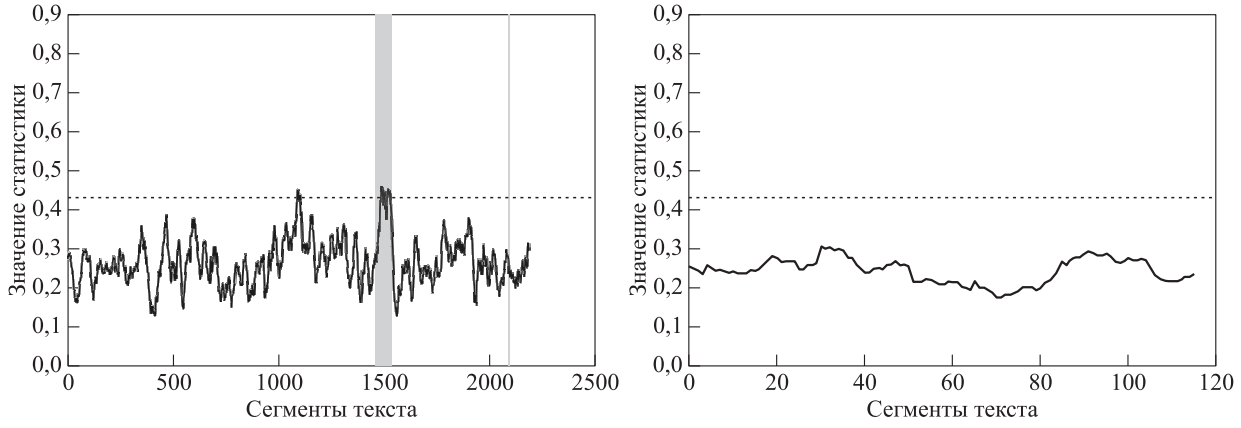

Рис. 4 Результаты на тестовой выборке

Сравнение качества алгоритмов на корпусе PAN-2011

Алгоритм	Precision	Recall	F1	Granularity	Pladget
Предлагаемый авторами	0,27	0,28	0,28	1,04	0,28
Oberreuter	0,34	0,31	0,33	1,00	0,33
Kestemont	0,11	0,43	0,17	1,03	0,17

Результаты работы и сравнение с двумя алгоритмами приведены в таблице.

Описанный алгоритм использует частоты распределения слов. Сегментирование текста происходит по группам предложений. Для определения заимствованных фрагментов исследуются значения статистики каждого сегмента.

На корпусе PAN-2011 алгоритм показал сравнимые результаты с победителем конкурса — алгоритмом Oberreuter. Также было проведено сравнение с алгоритмом Kestemont'a, занявшим второе место на конкурсе. Качество работы предлагаемого алгоритма значительно превышает качество работы алгоритма Kestemont'a.

6 Анализ ошибок

Результаты работы предлагаемого алгоритма зависят от длины документа. При анализе небольших по объему текстов сглаживание приводит к существенной потере информации об аномальных значениях статистики. При малой ширине сглаживания шумовые выбросы вызывают ложное срабатывание алгоритма.

7 Заключение

Предлагаемый алгоритм использует распределение частот слов внутри текста для нахождения

заимствованных сегментов. Сегментирование текста осуществляется по группам предложений. Для каждого сегмента строится статистика. Затем ряд статистики для всего текста сглаживается методом скользящего среднего. Полученные значения исследуются на отклонение от среднего значения для выявления заимствованных сегментов.

Алгоритм был настроен и протестирован на корпусе PAN-2011. Алгоритм Oberreuter [6], модификацией которого является предлагаемый алгоритм, показал на этом же корпусе результаты в 0,32 по F1-мере. Таким образом, описанный алгоритм показал сравнимые результаты при работе с тем же корпусом документов.

Дальнейшие исследования могут быть направлены на более точную настройку параметров алгоритма, подбор параметров в зависимости от длины рассматриваемого текста, поиск новых признаков, которые будут точнее выявлять заимствованные фрагменты, а также поиск новых способов нахождения заимствований.

Авторы выражают свою благодарность доктору физико-математических наук В. В. Стрижову, а также кандидату физико-математических наук Ю. В. Чеховичу за ценные советы при планировании исследования и рекомендации по оформлению статьи.

Литература

1. *Никитов А. В., Орчаков О. А., Чехович Ю. В.* Плагиат в работах студентов и аспирантов: проблема и методы противодействия // Университетское управление: практика и анализ, 2012. № 5. С. 61–68.
2. *Zechner M., Muhr M., Kern R., Granitzer M.* External and intrinsic plagiarism detection using vector space models // CEUR Workshop Proceedings, 2009. Vol. 502. P. 47–55.
3. *Tschuggnall M., Specht G.* Countering plagiarism by exposing irregularities in authors grammars // European Intelligence and Security Informatics Conference. — IEEE, 2013. P. 15–22.
4. *Eissen S. M., Stein B.* Intrinsic plagiarism detection // Advances in information retrieval / Eds. M. Lalmas, A. MacFarlane, S. M. Rüger, *et al.* — Lecture notes in computer science ser. — Springer, 2006. Vol. 3936. P. 565–569.
5. *Stamatatos E.* Intrinsic plagiarism detection using character n-gram profiles // CEUR Workshop Proceedings, 2009. Vol. 502. P. 38–46.
6. *Oberreuter G., L'Huillier G., Ríos S. A., Velásquez J. D.* Outlier-based approaches for intrinsic and external plagiarism detection // Knowlege-based and intelligent information and engineering systems / Eds. A. König, A. Dengel, K. Hinkelmann, *et al.* — Lecture notes in computer science ser. — Springer, 2011. Vol. 6882. P. 11–20.
7. *Bensalem I. Rosso P., Chikhi S.* Intrinsic plagiarism detection using n-gram classes // Conference on Empirical Methods in Natural Language Processing Proceedings. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2014. P. 1459–1464.
8. *Vartapetiance A., Gillam L.* Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. http://epubs.surrey.ac.uk/id/eprint/766727.
9. *Kuznetsov M., Motrenko A., Kuznetsova R., Strijov V.* Methods for intrinsic plagiarism detection and author diarization. http://ceur-ws.org/Vol-1609/16090912.pdf.
10. *Potthast M., Stein B., Barron-Cedeno A., Rosso P.* An evaluation framework for plagiarism detection // 23rd Conference (International) on Computational Linguistics Proceedings. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. P. 997–1005.

Поступила в редакцию 30.01.17

METHODS FOR INTRINSIC PLAGIARISM DETECTION

K. F. Safin[1,2], M. P. Kuznetsov[3], and M. V. Kuznetsova[1,2]

[1]Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation
[2]Antiplagiat JSC, 33 Varshavskoe Shosse, Moscow 117105, Russian Federation
[3]"Forecsys" LLC, 42 Vavilov Str., Moscow 119333, Russian Federation

Abstract: There are two ways to find plagiarism in documents: "external" and "intrinsic" plagiarism detection. External plagiarism detection is the task with a known set of possible references. Intrinsic plagiarism detection aims at discovering plagiarism by analyzing only the document by itself. The paper investigates the methods of intrinsic plagiarism detection. The authors developed a plagiarism detection method based on constructing statistics from the features of the document parts and detecting outliers. The proposed algorithm was tested on the PAN-2011 collection for intrinsic plagiarism detection.

Keywords: natural language processing; intrinsic plagiarism detection; outliers detection

DOI: 10.14357/19922264170308

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (project 16-07-01155).

References

1. Nikitov, A. V., O. A. Orchakov, and Ju. V. Chehovich. 2012. Plagiat v rabotakh studentov i aspirantov: Problema i metody protivodeystviya [Plagiarism in works of undergraduate and graduate students: Problem and methods of counteraction]. *Universitetskoe upravlenie: Praktika i analiz* [University Management: Practice and Analysis] 5:61-68.
2. Zechner, M., M. Muhr, R. Kern, and M. Granitzer. 2009. External and intrinsic plagiarism detection using vector space models. *CEUR Workshop Proceedings*. 502:47–55.
3. Tschuggnall, M., and G. Specht. 2013. Countering plagiarism by exposing irregularities in authors grammars. *European Intelligence and Security Informatics Conference Proceedings*. IEEE. 15–22.
4. Eissen, S. M., and B. Stein. 2006. Intrinsic plagiarism detection. *Advances in information retrieval*. Eds. M. Lalmas, A. MacFarlane, S. M. Rüger, *et al.* Lecture notes in computer science ser. Springer. 3936:565–569.
5. Stamatatos, E. 2009. Intrinsic plagiarism detection using character n-gram profiles. *CEUR Workshop Proceedings*. 502:38–46.
6. Oberreuter, G., G. L'Huillier, S. Ríos, and J. Velásquez. 2011. Outlier-based approaches for intrinsic and external plagiarism detection. *Knowlege-based and intelligent information and engineering systems*. Eds. A. König, A. Dengel, K. Hinkelmann, *et al.* Lecture notes in computer science ser. Springer. 6882:11–20.
7. Bensalem, I., P. Rosso, and S. Chikhi. 2014. Intrinsic plagiarism detection using n-gram classes. *Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics. 1459–1464.
8. Vartapetiance, A., and L. Gillam. Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. Available at: http://epubs.surrey.ac.uk/id/eprint/766727 (accessed September 23, 2013).
9. Kuznetsov, M., A. Motrenko, R. Kuznetsova, and V. Strijov. Methods for intrinsic plagiarism detection and author diarization. Available at: http://ceur-ws.org/Vol-1609/16090912.pdf (accessed September 6, 2016).
10. Potthast, M., B. Stein, A. Barrón-Cedeño, and P. Rosso. 2010. An evaluation framework for plagiarism detection. *23rd Conference (International) on Computational Linguistics Proceedings*. Stroudsburg, PA: Association for Computational Linguistics. 997–1005.

Received January 30, 2017

Contributors

Safin Kamil F. (b. 1995) — student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation; junior researcher, Antiplagiat JSC, 33 Varshavskoe Shosse, Moscow 117105, Russian Federation; kamil.safin@phystech.edu

Kuznetsov Mikhail P. (b. 1989) — Candidate of Science (PhD) in physics and mathematics; analyst, "Forecsys" LLC, 42 Vavilov Str., Moscow 119333, Russian Federation; mikhail.kuznecov@phystech.edu

Kuznetsova Margarita V. (b. 1990) — PhD student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation; Head of Departament, Antiplagiat JSC, 33 Varshavskoe shosse, Moscow 117105, Russian Federation; kuznetsova@ap-team.ru