УДК 519.25 ; 004.8

А.С. Романов, Р.В. Мещеряков, З.И. Резанова

Методика проверки однородности текста и выявления плагиата на основе метода опорных векторов и фильтра быстрой корреляции

Проведен анализ существующих средств и подходов поиска плагиата в тексте, обоснована необходимость дополнительной проверки текста на однородность. Приводятся описание и результаты работы методики проверки текста на однородность и выявления плагиата на основе кроссвалидации, одноклассового классификатора машины опорных векторов и фильтра быстрой корреляции для определения наиболее информативных признаков текста.

Ключевые слова: плагиат, однородность текста, быстрая корреляция, кроссвалидация, одноклассовая классификация, машина опорных векторов.

Типичная система для выявления плагиата [1] представляет собой программу, сравнивающую два текста на наличие общих подстрок и предполагающую использование базы данных возможных источников заимствований. В зависимости от расположения базы данных программы для выявления плагиата можно разделить на три группы [2]:

1. «Онлайновые» системы. Позволяют производить поиск оригинальных источников в сети Интернет благодаря интеграции с поисковыми системами.

2. «Оффлайновые» системы. Позволяют проводить поиск дубликатов в пределах локальной коллекции.

3. Универсальные. Позволяют формировать собственные коллекции текстов, проводить поиск в этих коллекциях, а также использовать сеть Интернет для поиска источников заимствований.

В случае если источник заимствования не найден, любая из трех описанных систем однозначно расценивает текст как оригинальный. Однако причиной этому может служить недостаточный объем базы данных, банальное отсутствие текстового слоя в документе-«доноре» и др. Поэтому всё чаще можно слышать о фактах публикации текстов, частично заимствованных из других источников, полного плагиата [3] и даже полностью искусственно сгенерированных текстов [4]. Выявление и пресечение подобных случаев является актуальной междисциплинарной практической задачей, затрагивающей области лингвистики, криминалистики, информационной безопасности, интеллектуального анализа данных и др.

Одним из способов повышения качества работы сервисов поиска плагиата и аналогичных систем в других областях нам видится добавление проверки текста на однородность: в случае если какой-либо из фрагментов явно отличается от общего авторского стиля текста, то велика вероятность того, что этот фрагмент заимствован из другого источника.

Для проведения таких проверок можно адаптировать уже известные методы идентификации автора текста [5]:

– использовать методы статистического анализа и теории информации: методы сжатия информации, проверку статистических гипотез о равенстве средних на основе критерия Стьюдента, критерий Колмогорова–Смирнова, меру Кульбака и хи-квадарт (на использовании последней меры основан алгоритм определения плагиата в работе [6]);

– использовать методы машинного обучения. При этом нужно интерпретировать задачу поиска неоднородностей как задачу одноклассовой классификации [7], а возможное заимствование определять путем обучения и тестирования на фрагментах текста методом кроссвалидации [8]. Другой вариант – заранее обучить несколько разных классификаторов, способных определять пол автора, возраст, образование, собственно стиль конкретного автора и т.д. Подавая на вход обученных моделей вектор признаков для отдельных фрагментов, можно, например, выявить в тексте с явно мужским стилем отдельные фрагменты, написанные женщиной; в тексте, соответствующем возрастной группе 18–25 лет, – группы предложений, характерные для пожилых людей, и т.д., как это делалось в работе [9];

– использовать специализированные методы, такие как, например, метод накопительных сумм (QSUM) [10–12]. Для проведения анализа выбирается пара характеристик, являющихся функциями предложения. Затем производится подсчет этих характеристик для каждого предложения и вычисляется среднее значение для всего текста. После считаются отклонения от средних значений для каждого предложения и строится накопительная сумма отклонений: начиная с нуля и затем последовательно прибавляя отклонения остальных предложений. Для каждой характеристики строится масштабированный график, на котором отображаются значения сумм для каждого этапа вычисления накопительной суммы. Графики однородного стиля должны практически совпадать. Неоднородный текст покажет их несовпадение. Главным преимуществом метода QSUM является то, что он дает отклонения хронологически, отображая их накопленную сумму. Основной недостаток метода – интерпретация полученных графиков. Для объективного вынесения решения об однородности текста можно использовать регрессионный анализ и методы машинного обучения, как это сделано в работе [2], – точность такого модифицированного метода доходит до 75% даже без тонкой настройки параметров, однако сильно зависит от позиции «вставки» и общего размера текста.

В данной работе предлагается комбинированная методика проверки однородности текста и выявления плагиата, включающая последовательное использование:

1) метода отбора информативных признаков, основанного на быстрой корреляции (FCBF);

2) метода машинного обучения машина опорных векторов (SVM).

Выбор информативных признаков текста. Как уже было сказано, ключевую роль в вопросе проверки однородности текста играет выбор признаков. Характеристика должна слабо контролироваться автором на сознательном уровне, быть устойчивой к изменению стиля внутри текстов одного и того же автора и быть способной статистически разделить двух и более авторов с заданной точностью. Сложность заключается как в выборе этих характеристик, так и в методике их сравнения.

Для отбора информативных признаков текста в работе используется метод многомерного отбора-FCBF (Fast Correlation-Based Filter) [13]. Метод начинает работать с полным множеством доступных для анализа признаков, использует меру симметричной неопределенности для определения зависимостей между признаками и позволяет найти подмножество, лучше всего описывающее данную предметную область, путем поиска и последовательного исключения малоинформативных признаков.

Мера симметричной неопределенности рассчитывается как

$$SU(X,Y)=2\left[\frac{H(X)-H(X|Y)}{H(X)+H(Y)}\right]=SU(Y,X),$$

где $H(X)$, $H(Y)$ – энтропии случайных величин, имеющих соответственно *i* и *j* состояний:

$$H(X)=-\sum_i P(x_i)\log_2(P(x_i)),$$

$H(X|Y)$ – условная энтропия:

$$H(X|Y)=-\sum_j P(y_j)\sum_i P(x_i|y_i)\log_2(P(x_i|y_j)),$$

$P(x_i)$, $P(y_i)$ – априорные вероятности для всех значений *X* и *Y*, $P(x_i|y_j)$ – апостериорная вероятность *X* при известных *Y*.

Значение *SU*, равное единице, свидетельствует о том, что, используя первый признак, можно точно предсказать значение второго, тогда как нулевое значение означает полную независимость признаков.

Пусть имеется набор данных *S*, состоящий из *C* классов и описывающийся *N* признаками. Для получения итогового подмножества признаков выполняются следующие действия:

1. Путем последовательного расчета меры для всех признаков и сравнения с заданным пороговым значением δ получают множество S' релевантных классу *C* признаков $\forall F_i \in S'$, $i=\overline{1,N}$, $SU_{i,c}>\delta$, где $SU_{i,c}$ обозначает корреляцию признака F_i и класса *C*.

2. Определение доминантных признаков таких, что для F_i ($F_i \in S$, $SU_{i,c}>\delta$) не существует $F_j \in S'$ ($j \neq i$), для которого $SU_{j,i}>SU_{i,c}$, где $SU_{j,i}$ – количественная оценка степени корреляции

признака F_i и других релевантных признаков из множества S'. Признак с самым большим значением $SU_{i,c}$ является доминантным признаком всегда.

3. Если найден F_j, для которого условие из п. 2 не выполняется, то считаем его избыточным по отношению к F_i. Обозначим S_{P_i} как множество, содержащее все возможные избыточные признаки по отношению к F_i.

4. Пусть $F_i \in S'$ и множество S_{P_i} не пустое. Разделим S_{P_i} на два класса: $S_{P_i}^{+} = \{F_j \mid F_j \in S_{P_i}, SU_{j,c} > SU_{i,c}\}$ и $S_{P_i}^{-} = \{F_j \mid F_j \in S_{P_i}, SU_{j,c} \le SU_{i,c}\}$.

5. Если $\left|S_{P_i}^{+}\right| = 0$, то F_i можно считать доминантным признаком, не продолжать поиск избыточных признаков для элементов множества $S_{P_i}^{-}$, а удалить их.

6. Если $\left|S_{P_i}^{+}\right| \ne 0$, то необходимо проверить его элементы: если среди них не найдено доминантных признаков, то следовать п. 5, иначе – удалить признак F_i, а решение относительно удаления признаков $S_{P_i}^{-}$ принимать на основе других признаков S'.

Проверка текста на однородность. Итоговый алгоритм проверки текста на однородность представлен на рис. 1.

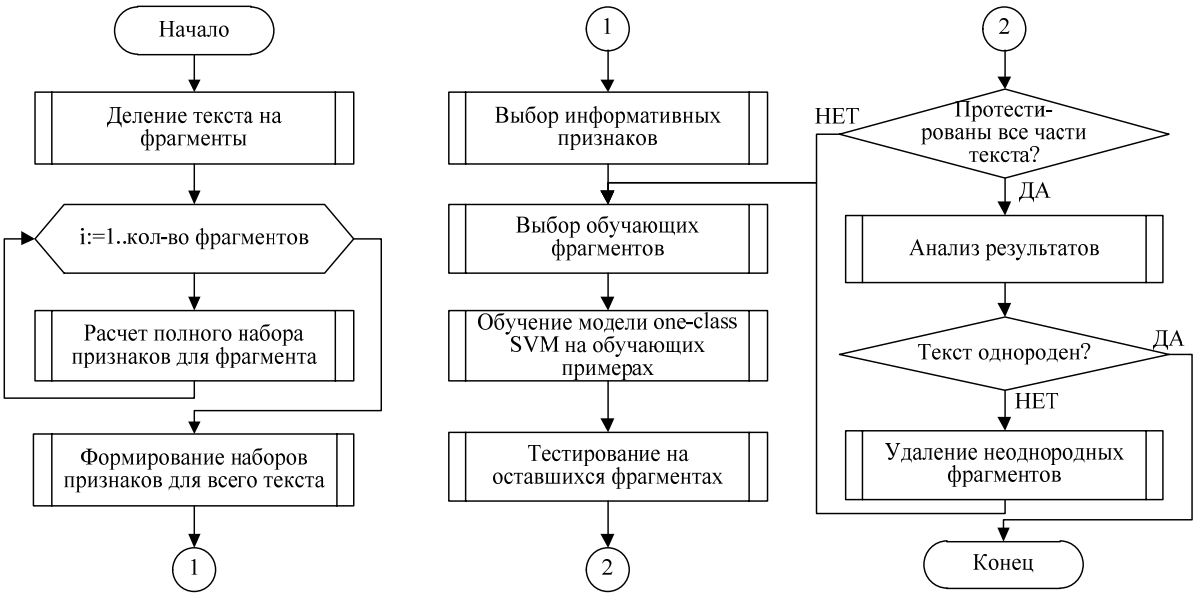

Рис. 1. Алгоритм проверки текста на однородность

Решение относительного того, отличается ли текст от общего авторского стиля, принимается классификатором на основе машины опорных векторов (SVM), показавшей отличные результаты при решении ряда смежных задач [14–16]. Однако учитывая специфику задачи, используется модификация метода – одноклассовый SVM и его реализация в libsvm [17], позволяющие проводить обучение только на основе примеров одного класса.

Для выделения неоднородных фрагментов из текста используется метод кроссвалидации – фрагменты текста делятся в определенной пропорции на обучающие и тестовые примеры, производится обучение классификатора и проверка, далее тестовая часть берется другая, после чего процедура повторяется. Найденные неоднородные фрагменты удаляются, и поиск начинается сначала. Критерием останова может служить, например, максимальная доля фрагментов, которые можно удалить из текста, или другая эвристика. Следует отметить, что метод не чувствителен к позиции «вставки» относительно начала текста.

При делении текста на фрагменты в случае, если в тексте присутствует явное членение на предложения, предпочтительно использовать их границы. Если границы предложения явно не

обозначены, используется набор слов между двумя знаками препинания, либо фрагменты фиксированной длины.

Экспериментальная часть. В исследовании использовались следующие корпуса текстов:

1) корпус прозаических текстов русских писателей XVIII–XX вв. (всего 215 текстов 50 авторов);

2) научные статьи по филологии, истории, праву, экономике и другим общественным и гуманитарным наукам, взятые из электронного архива журнала «Вестник Томского государственного университета» [18] (всего 500 текстов, написанных без соавторства).

Все тексты были предварительно размечены: в автоматическом режиме определены границы предложений и проведен морфологический анализ. Большие тексты были разбиты на более мелкие (по 100 предложений). Для имитации плагиата в каждый из полученных текстов было добавлено единым блоком от 1 до 10 предложений, взятых из текстов другого автора. Для корпуса статей, по возможности, дополнительно учитывалось научное направление. Всего было получено порядка 1000 таких примеров с заранее известными позициями и объемами вставок для каждого из корпусов.

В качестве сравниваемых характеристик использовались единицы символьного уровня текста, элементы грамматики, идиосинкразические и специальные признаки текста, в том числе:

– признаки, предложенные Мортоном: длина предложения (в словах) и комбинация слов, начинающихся с гласной буквы, и коротких слов из двух-четырех букв;

– наборы биграмм и триграмм символов, разделенные по частотному признаку;

– наборы слов и сочетаний слов, разделенные по частотному признаку;

– грамматические классы слов и сочетания грамматических классов;

– словари соответствующих научных дисциплин;

– словари мужских и женских признаков текста и др.

Для оценки качества работы метода использовалась F-мера, представляющая собой гармоническое среднее между точностью P и полнотой классификации R:

$$F = 2\frac{P \cdot R}{P + R},$$

Полученные результаты приведены в таблице.

Результаты экспериментов

Корпус	Объем «вставки»									
	1	2	3	4	5	6	7	8	9	10
Произведения русских авторов XVIII–XX вв., %	64	75	89	85	85	84	79	80	82	78
Статьи по общественным и гуманитарным наукам, %	55	70	76	73	73	77	68	67	73	65

Более точные результаты для первого корпуса можно объяснить тем, что писатели обладают ярко выраженным авторским стилем, поэтому вставка «чужеродного» текста может быть обнаружена сравнительно легко. Данный вывод подтверждается предыдущими результатами исследований [5] и экспертами-лингвистами. В корпусе научных статей авторский инвариант выражен в меньшей степени. Свои особенности накладывают также лексические, морфологические и синтаксические особенности научного стиля. В целом следует отметить достаточно высокие результаты для обоих корпусов, из которых следует, что предложенный подход является перспективным и требует более тщательной проверки.

Заключение. В данной статье рассмотрена важная междисциплинарная практическая задача – выявление неоднородных фрагментов в тексте и плагиата. Обоснована необходимость использования методов проверки текста на однородность авторского стиля наряду с классическими алгоритмами текстового поиска. Предложена методика поиска неоднородных фрагментов в тексте, основанная на использовании кроссвалидации и одноклассовой классификации методом машины опорных векторов. Выбор информативных критериев предлагается делать автоматически на основе фильтра быстрой корреляции. Полученные экспериментальные результаты позволяют сделать вывод о достаточно высокой точности работы метода и перспективности предложенного подхода для решения поставленной задачи.

Полученные результаты не являются окончательными для данной работы. Мы планируем базироваться на них в своих будущих исследованиях. Как возможные направления развития темы рассматриваются следующие задачи:

1. Расширение корпуса научных статей и апробация методики на статьях, относящихся к естественным и техническим наукам.

2. Создание специального корпуса, объединяющего тексты, в которых вставка инородных предложений производится человеком осмысленно с учетом жанра, темы, контекста и прочих особенностей конкретного текста: как реальные примеры плагиата, так и специальные тексты, подготовленные экспертами. Апробация методики на этом корпусе.

3. Экспертная лингвистическая оценка полученных результатов и усовершенствование методики за счет добавления полученной дополнительной информации.

4. Полная автоматизация предложенного подхода и создание автоматизированной системы для проверки текста на однородность и определения плагиата.

Литература

1. Дягилев В.В. Архитектура сервиса определения плагиата, исключающая возможность нарушения авторских прав / В.В. Дягилев, А.А. Цхай, С.В. Бутаков // Вестник НГУ. Сер.: Информационные технологии. – 2011. – Т. 9, вып. 3. – С. 23–29.

2. Романов А.С. Модификация метода накопительных сумм для проверки однородности текста и выявления плагиата // Электронные средства и системы управления: матер. докл. IX Междунар. науч.-практ. конф. (30–31 октября 2013 г.): в 2 ч. – Ч. 2. – Томск: В-Спектр, 2013. – С. 30–38.

3. Экспертизы. Вольное сетевое сообщество «Диссернет» [Электронный ресурс]. – Режим доступа: http://www.dissernet.org/expertise, свободный (дата обращения: 19.04.2014).

4. Шумская А.О. Выбор параметров для идентификации искусственно созданных текстов // Доклады Том. гос. ун-та систем управления и радиоэлектроники. – 2013. – № 2 (28). – С. 126–128.

5. Романов А.С. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста / А.С. Романов, А.А. Шелупанов, Р.В. Мещеряков. – Томск: В-Спектр, 2011. – 188 с.

6. Седов А.В. Анализ неоднородностей в тексте на основе последовательностей частей речи [Электронный ресурс] / А.В. Седов, А.А. Рогов // Современные проблемы науки и образования. – 2013. – № 1. – Режим доступа: www.science-education.ru/107-8339, свободный.

7. Stein B. Intrinsic plagiarism analysis with meta learning / B. Stein, S. Meyer zu Eissen [Электронный ресурс]. – Режим доступа: http://www.uni-weimar.de/medien/webis/research/events/pan-07/pan07-papers-final/stein07-intrinsic-plagiarism-analysis-with-meta-learning.pdf, свободный (дата обращения: 19.04.2014).

8. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. – М.: Физматлит, 2004. – Т. 13. – С. 5–36.

9. Mechti S. A framework for plagiarism detection based on author profiling / S. Mechti, M. Jaoua, H. Belghith // Notebook for PAN at CLEF 2013 [Электронный ресурс]. – Режим доступа: http://www.clef-initiative.eu/documents/71612/c7a0e432-dd82-46b1-ab9e-5d0dd98c3a8d, свободный (дата обращения: 19.04.2014).

10. Morton A.Q. Literary Detection: How to prove authorship and fraud in literature and documents. – New York : Scribner's, 1978. – 221 p.

11. Farringdon J.M. Analyzing for authorship / J.M. Farringdon with contributions by A.Q. Morton, M.G. Farringdon, M.D. Baker. – Cardiff : University of Wales Press, 1996. – 324 p.

12. Holmes D. Forensic stylometry: A review of the qsum controversy / D. Holmes, F.J. Tweedie // Revue informatique et statistique dans les sciences humanies. – 1995. – № 31. – P. 19–47.

13. Yu L. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution / L. Yu, H. Liu // Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03). – 2003. – P. 856–863.

14. Романов А.С. Идентификация автора текста с помощью аппарата опорных векторов / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: матер. ежегод. междунар. конф. «Диалог–2009» (Бекасово, 27–31 мая 2009 г.). – М.: РГГУ, 2009. – Вып. 8 (15). – С. 432–437.

15. Романов А.С. Идентификация авторства коротких текстов методами машинного обучения / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: по

матер. ежегод. междунар. конф. «Диалог» (Бекасово, 26–30 мая 2010 г.). – М.: Изд-во РГГУ, 2010. – Вып. 9 (16). – С. 407–413.

16. Романов А.С. Определение пола автора короткого электронного сообщения / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: матер. ежегод. Междунар. конф. «Диалог» (Бекасово, 25 – 29 мая 2011 г.). – М.: Изд-во РГГУ, 2011. – Вып. 10 (17). – С. 620–626.

17. Chang C.-C. LIBSVM: a library for support vector machines / C.-C. Chang, C.-J. Lin // ACM Transactions on Intelligent Systems and Technology [Электронный ресурс]. – 2011. – Режим доступа: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf, свободный (дата обращения: 19.04.2014).

18. Резанова З.И. Задачи авторской атрибуции текста в аспекте гендерной принадлежности (к проблеме междисциплинарного взаимодействия лингвистики и информатики) / З.И. Резанова, А.С. Романов, Р.В. Мещеряков // Вестник Том. гос. ун-та. – 2013. – № 370. – С. 24–28.

19. Давыдова Е.М. Модель образовательного процесса с учетом требований работодателя // Доклады Том. гос. ун-та систем управления и радиоэлектроники. – 2013. – № 4 (30). – С. 177–181.

Романов Александр Сергеевич
Канд. техн. наук, доцент каф. комплексной информационной безопасности электронно-вычислительных систем (КИБЭВС) ТУСУРа
Тел.: 8 (382-2) 41-34-26
Эл. почта: alexx.romanov@gmail.com

Мещеряков Роман Валерьевич
Д-р техн. наук, профессор каф. КИБЭВС ТУСУРа
Тел.: 8 (382-2) 41-34-26
Эл. почта: mrv@ieee.org

Резанова Зоя Ивановна
Д-р фил. наук, зав. каф. общего, славяно-русского языкознания и классической филологии
Национального исследовательского Томского государственного университета,
профессор каф. русского языка и литературы
Национального исследовательского Томского политехнического университета
Тел.: 8 (382-2) 52-67-89
Эл. почта: resso@rambler.ru

Romanov A.S., Meshcheryakov R.V., Rezanova Z.I.

Plagiarism detection and text homogeneity checking technique based on one-class support machine and fast correlation-based filter

The article provides an analysis of the existing tools and approaches for identifying text plagiarism, justifying the need for additional verification of the text homogeneity. The article presents the description and results of the technique for the purpose of determining the plagiarism in the text, based on cross-validation, one-class SVM classifier and fast correlation-based filter.

Keywords: plagiarism, text homogeneity, fast correlation based filter, cross-validation, one-class classification, support vector machine.