КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2022 Т. 14 № 5 С. 1199–1210
DOI: 10.20537/2076-7633-2022-14-5-1199-1210

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 519.243

Определение автора текста методом сегментации

М. Ю. Воронина[a], Ю. Н. Орлов

Федеральный исследовательский центр «Институт прикладной математики им. М. В. Келдыша Российской академии наук»,
Россия, 125047, г. Москва, Миусская пл., д. 4

E-mail: [a] voronina.miu@phystech.edu

Получено 27.06.2022, после доработки — 09.08.2022.
Принято к публикации 12.08.2022.

В работе описывается метод распознавания авторов литературных текстов по близости фрагментов, на которые разделен отдельный текст, к эталону автора. Эталоном является эмпирическое распределение частот буквосочетаний, построенное по обучающей выборке, куда вошли экспертно отобранные достоверно известные произведения данного автора. Совокупность эталонов разных авторов образует библиотеку, внутри которой и решается задача об идентификации автора неизвестного текста. Близость между текстами понимается в смысле нормы в L1 для вектора частот буквосочетаний, который строится для каждого фрагмента и для текста в целом. Автором неизвестного текста назначается тот, эталон которого чаще всего выбирается в качестве ближайшего для набора фрагментов, на которые разделен текст. Длина фрагмента оптимизируется исходя из принципа максимального различия расстояний от фрагментов до эталонов в задаче распознавания «свой–чужой». Тестирование метода проведено на корпусе отечественных и зарубежных (в переводе) авторов. Были собраны 1783 текста 100 авторов суммарным объемом примерно 700 млн знаков. Чтобы исключить тенденциозность отбора авторов, рассматривались авторы, фамилии которых начинались на одну и ту же букву (в данном случае Л). Ошибка идентификации по биграммам составила 12 %. Наряду с достаточно высокой точностью данный метод обладает еще одним важным свойством: он позволяет оценить вероятность того, что эталон автора рассматриваемого текста в библиотеке отсутствует. Эта вероятность может быть оценена по результатам статистики ближайших эталонов для малых фрагментов текста. В работе исследуются также статистические цифровые портреты писателей: это совместные эмпирические распределения вероятности того, что некоторая доля текста идентифицируется на заданном уровне доверия. Практическая важность этих статистик в том, что носители соответствующих распределений практически не пересекаются для своих и чужих эталонов, что позволяет распознать эталонное распределение буквосочетаний на высоком уровне доверия.

Ключевые слова: эмпирическое распределение частот, биграммы, идентификация автора, литературный текст, ближайший эталон

Исследование выполнено при поддержке Министерства науки и высшего образования РФ, договор № 075-15-2020-808.

COMPUTER RESEARCH AND MODELING
2022 VOL. 14 NO. 5 P. 1199–1210
DOI: 10.20537/2076-7633-2022-14-5-1199-1210

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 519.243

Identification of the author of the text by segmentation method

M. Yu. Voronina[a], Yu. N. Orlov

Keldysh Institute of Applied Mathematics Russian Academy of Sciences,
4 Miusskaya sq., Moscow, 125047, Russia

E-mail: [a] voronina.miu@phystech.edu

Received 27.06.2022, after completion — 09.08.2022.
Accepted for publication 12.08.2022.

The paper describes a method for recognizing authors of literary texts by the proximity of fragments into which a separate text is divided to the standard of the author. The standard is the empirical frequency distribution of letter combinations, built on a training sample, which included expertly selected reliably known works of this author. A set of standards of different authors forms a library, within which the problem of identifying the author of an unknown text is solved. The proximity between texts is understood in the sense of the norm in L1 for the frequency vector of letter combinations, which is constructed for each fragment and for the text as a whole. The author of an unknown text is assigned the one whose standard is most often chosen as the closest for the set of fragments into which the text is divided. The length of the fragment is optimized based on the principle of the maximum difference in distances from fragments to standards in the problem of recognition of «friend–foe». The method was tested on the corpus of domestic and foreign (translated) authors. 1783 texts of 100 authors with a total volume of about 700 million characters were collected. In order to exclude the bias in the selection of authors, authors whose surnames began with the same letter were considered. In particular, for the letter L, the identification error was 12 %. Along with a fairly high accuracy, this method has another important property: it allows you to estimate the probability that the standard of the author of the text in question is missing in the library. This probability can be estimated based on the results of the statistics of the nearest standards for small fragments of text. The paper also examines statistical digital portraits of writers: these are joint empirical distributions of the probability that a certain proportion of the text is identified at a given level of trust. The practical importance of these statistics is that the carriers of the corresponding distributions practically do not overlap for their own and other people's standards, which makes it possible to recognize the reference distribution of letter combinations at a high level of confidence.

Keywords: empirical frequency distribution, bigrams, author identification, literature text, the nearest pattern

Citation: *Computer Research and Modeling*, 2022, vol. 14, no. 5, pp. 1199–1210 (Russian).

This study was supported by the Ministry of Science and Higher Education of the Russian Federation, Contract No. 075-15-2020-808.

1. Введение

Задача автоматической классификации текста по атрибутам является весьма актуальной в контексте развития информационных технологий и анализа больших данных. Кроме того, она находится на стыке наук — математической статистики и лингвистики, что может способствовать развитию обеих отраслей знания. Однако следует отметить, что взаимопроникновение методов, применяемых в указанных науках, идет весьма медленно. Отчасти это связано с тем, что многочисленные методы машинного обучения (см., например, [Stamatatos, Fakotakis, Kokkinakis, 2000; Argamon, Juola, 2011; Sudheep et al., 2013; Cappellato et al., 2014]), используемые для той или иной обработки и анализа текстовых документов, представляют собой «черный ящик» не только для лингвистов, но и, собственно, для математиков. Не ясно, насколько эти методы, настроенные на решение задачи классификации в рамках конкретно отобранного корпуса текстов, робастны применительно к другим выборкам и другим задачам. Кроме того, в силу специфики настройки сложно предложить процедуру ее коррекции.

В обзорной работе [Резанова, Романов, Мещеряков, 2013] сравнивается эффективность чисто статистических методов анализа, основанных на подсчете формальных показателей, таких как число букв, слов, знаков препинания и т. п., с экспертными методами анализа авторского стиля, оборотов речи, использования литературных приемов. Авторы приходят к выводу, что, хотя для литературоведов более ценен экспертный метод, он не обладает достаточной точностью на большом корпусе текстов и, что более существенно, не может быть адекватно реализован в виде формальной компьютерной программы. В то же время статистика букв или буквосочетаний, хотя и не имеет непосредственного литературного смысла, может быть вполне однозначно сопоставлена каждому тексту с указанием погрешности в рамках формальных критериев. Тем самым в контексте задачи машинного распознавания атрибутов текстов статистический метод более эффективен, т. е. имеет меньшую ошибку, чем экспертный.

В задачах идентификации существенной проблемой после собственно модели распознавания является оценка вероятности принятия ошибочного решения: либо ложного принятия «чужого» за «своего», либо ложного отклонения «своего», принятого за «чужого». Формально распознаваемый объект в математическом плане представляет собой некоторый числовой вектор, размерность которого равна размерности пространства параметров, измеряемых при наблюдении объекта. Например, если изучается частота встречаемости букв русского алфавита в каких-либо текстах, то такой вектор имеет размерность 33, вектор пар буквосочетаний имеет размерность $1089 = 33^2$ и т. д.

Идея статистического распознавания состоит в том, что каждый объект реализуется не точно, а с некоторой ошибкой, порождаемой конечной точностью измерений и конечной длиной изучаемого ряда данных. Тогда результатом отдельного наблюдения является некоторое выборочное распределение из генеральной совокупности параметров, предположительно отвечающих идеальной модели. При таком подходе координаты наблюдаемого объекта лежат в некотором многомерном параллелепипеде, длина стороны которого по каждой переменной соответствует разбросу данных около эталонного значения. Очевидным достаточным условием правильной идентификации объекта является непересекаемость носителей распределений параметров для разных объектов. Однако в реальности такая ситуация встречается очень редко. Поэтому принадлежность результата наблюдения тому или иному множеству определяется, вообще говоря, с некоторой ненулевой ошибкой. Вероятность ошибки зависит от правила сравнения и метрики, в которой объекты сравниваются с эталонами.

На практике эталоны представляют собой более или менее обоснованные теоретические модели или создаются по результатам обработки большого массива эмпирических данных. Технологию искусственного интеллекта, реализующего задачу распознавания, будем называть логи-

чески прозрачной, если сформулирован однозначный метод идентификации, каждый шаг которого может быть проверен, а вероятность ошибки оценивается по явно сформулированным критериям вне системы машинного обучения с перечислением априорных допущений о свойствах объектов изучаемой категории. Это определение уточняет в части оценки ошибки принятое в работе [Roscher et al., 2020] соглашение о том, что объясняемый (интерпретируемый, прозрачный) искусственный интеллект — это такой метод машинного обучения, который может быть назван «белым ящиком» и который позволяет проверить, какие именно операции и на каких этапах были проведены распознающим алгоритмом.

Статистическая идентификация состоит в нахождении наиболее вероятной генеральной совокупности, которой могло бы принадлежать данное выборочное распределение. Ошибка такого решения дается критерием Колмогорова. Если выборка достаточно большая, как в случае с анализом больших литературных произведений, то уровень доверия близок к единице. Однако выясняется, что генеральные совокупности, характеризующие авторов (т. е. авторские эталоны), весьма мало отличаются один от другого, так что формальная близость к чужому эталону также оказывается почти такой же, как и для своего эталона. Просто к своему эталону выборка оказалась чуть ближе. В связи с этим возникает необходимость разработки второго индикатора, помимо собственно расстояния между текстом и эталоном, который давал бы оценку вероятности того, что найденный ближайший сосед среди авторских эталонов не является правильным ответом. Трудность построения такого индикатора состоит в том, что он должен, с одной стороны, не зависеть от первого индикатора, т. е. от близости распределений буквосочетаний текста и эталона. Но, с другой стороны, в случае независимости индикаторов вероятность правильного распознавания двумя индикаторами меньше, чем каждым из них в отдельности. Если же индикаторы зависимы, то в такой композиции нет необходимости, ибо не приносится нового знания.

В настоящей работе представлена модель коррекции ошибки идентификации на примере задачи распознавания автора текста методом кросс-валидации. Авторским эталоном является эмпирическое распределение частот буквосочетаний, построенное по всем достоверно известным произведениям автора. Близость между текстами понимается как близость между эмпирическими распределениями частот парных буквосочетаний — биграмм в смысле нормы в L1. Автором неизвестного текста назначается тот, к эталону которого тестируемый текст находится ближе всего, т. е. используется метод ближайшего соседа. Для идентификации используется библиотека авторов, каждый из которых имеет достаточно большое количество произведений, определяющих соответствующие эталоны биграмм. Изучаемой проблемой является оценка вероятности того, что среди эталонов библиотеки нет эталона автора тестируемого текста. Для ее решения предлагается исследовать зависимость вероятности ошибочной идентификации от длины текста. Численные эксперименты показали, что правильный автор устойчиво распознается на фрагментах текста, существенно (на порядок) меньших исходного. При исключении из рассмотрения правильного эталона идентификация будет заведомо неверной, а «автору» будет отвечать следующий по близости эталон, который при фрагментации текста может показывать свойства, отличные от правильного эталона автора. Устойчивость идентификации автора фрагментов текста предлагается в качестве нового критерия корректности метода.

Частоты биграмм рассматривались и ранее. В работах [Хмелёв, 2000; Кукушкина, Поликарпов, Хмелёв, 2001] текст рассматривался как траектория марковского процесса присоединения букв с учетом определенных грамматических правил, для чего требовалось оценить условную вероятность появления символов в тексте. В монографии [Орлов, Осминин, 2012] исследовались различные метрики для определения разных атрибутов текста, также основанные на эмпирических частотах буквосочетаний. Однако в упомянутых работах анализ проводился на относительно небольшом корпусе текстов (порядка нескольких десятков авторов).

Отметим, что существует большое количество семантических методов идентификации (см., например, обзор [Батура, 2017]), основанных на анализе частоты употребления слов. Каждый из таких методов имеет самостоятельную область применения, но в целом это «инженерные» методики, достоверность которых по достаточно большому корпусу текстов не превосходит 0,7. Мы здесь не имеем цель сравнивать разные методы, а демонстрируем эффективность метода биграмм и связанного с ним метода фрагментации текста с последующим «голосованием» по частоте упоминания автора фрагмента.

2. Постановка задачи, основные обозначения

Концепция распознавания автора текста по эмпирическим частотам буквосочетаний основана на предположении, что каждому профессиональному писателю соответствует его персональный эталон, трактуемый как генеральная совокупность. Тогда отдельные тексты представляют собой выборку из этой совокупности и отклоняются от нее в силу конечности текста и некоторых различий в тематике между текстами. Пусть $F_a(j)$ есть частота использования символа j в эталоне автора a, где j обозначает биграмму, т. е. буквосочетание из двух идущих подряд букв. Символы других алфавитов, пробелы, знаки препинания и цифры игнорируются. Прописные и строчные буквы не различаются. Пусть также $D_a^i(j)$ есть эмпирическая частота использования символа j в i-м тексте автора a, и пусть N_a^i есть число символов в данном i-м тексте. Пусть n_a есть число произведений автора a. Тогда эмпирической оценкой эталона $F_a(j)$ автора a является взвешенное распределение по совокупности всех текстов, достоверно принадлежащих данному автору:

$$F_a(j) = \frac{1}{N_a}\sum_{i=1}^{n_a} N_a^i D_a^i(j), \quad N_a = \sum_{i=1}^{n_a} N_a^i. \tag{1}$$

Расстояние между i-м текстом автора a и эталоном $F'_a(j)$ этого же автора определяется в норме L1 по формуле

$$x_{aa}^i = \sum_{j=1}^{J} \left|D_a^i(j) - F'_a(j)\right|, \tag{2}$$

где $J = 33^2$, а штрих у эталона $F'_a(j)$ означает, что данный текст исключен из эталона (1):

$$F'_a(j) = \frac{1}{N_a - N_a^i}\sum_{\substack{k=1\\k\neq i}}^{n_a} N_a^k D_a^k(j) = \frac{1}{N_a - N_a^i}\sum_{k=1}^{n_a} N_a^k D_a^k(j) - \frac{N_a^i D_a^i(j)}{N_a - N_a^i} = \frac{N_a F_a(j) - N_a^i D_a^i(j)}{N_a - N_a^i}.$$

Тогда расстояние от i-го текста до эталона своего автора имеет вид

$$x_{aa}^i = \sum_{j=1}^{J}\left|D_a^i(j) - \frac{N_a F_a(j) - N_a^i D_a^i(j)}{N_a - N_a^i}\right| = \frac{1}{1 - \frac{N_a^i}{N_a}}\sum_{j=1}^{J}\left|D_a^i(j) - F_a(j)\right|. \tag{3}$$

Расстояние между i-м текстом автора a и эталоном $F_b(j)$ некоторого другого автора $b \neq a$ есть

$$y_{ab}^i = \sum_{j=1}^{J}\left|D_a^i(j) - F_b(j)\right|. \tag{4}$$

Объединенная формула расстояния между текстом и эталоном имеет вид

$$z_{ab}^i = \frac{1}{1 - \frac{\delta_{ab} N_b^i}{N_b}}\sum_{j=1}^{J}\left|D_a^i(j) - F_b(j)\right|, \tag{5}$$

где δ_{ab} есть символ Кронекера.

В результате авторство b неизвестного текста с $D(j)$ определяется из условия

$$z(a) = \sum_{j=1}^{J} \left| F_a(j) - D(j) \right| = \min \;\Rightarrow\; b = \arg\min z(a). \tag{6}$$

Однако в случае отсутствия эталона автора исследуемого текста ближайший эталон все равно будет найден по правилу (6), но, разумеется, получившийся ответ будет неверным. В настоящей работе предлагается использовать второй индикатор, основанный на гипотезе устойчивости «своего» эталона и неустойчивости «чужого» при разделении текста на фрагменты.

С формальной стороны задача сводится к оценке вероятности правильной идентификации выборок из разных генеральных совокупностей в зависимости от близости между этими совокупностями и длины выборки. Насколько известно авторам настоящей работы, масштабного статистического эксперимента такого рода для каких-то стандартных распределений не проводилось. Если бы распределения буквосочетаний были стационарны, оценкой ошибки был бы наибольший уровень значимости, на котором распознается выборочное распределение в соответствии с критерием Колмогорова. При этом сохранялась бы упорядоченность эталонов «чужих» авторов по расстоянию до текстов данного автора в соответствии с тем, какие эталоны близки к эталону данного автора. Однако, как показывают исследования в области статистики текстов на естественных языках, распределения частот буквосочетаний нестационарны. Именно этот аспект позволяет определить ошибку распознавания автора текста при отсутствии такового в библиотеке эталонов.

Пусть тестируется достаточно длинный текст, который может быть разрезан на некоторое количество фрагментов. Уменьшая длину фрагмента текста, мы увеличиваем тем самым статистическую неопределенность оценивания выборочных частот. При этом было замечено, что эталоны «чужих» авторов, следующие за правильным эталоном в порядке увеличения расстояния до полного текста, не имеют устойчивой упорядоченности: с уменьшением длины фрагмента на втором месте оказываются эталоны разных авторов. Следовательно, если убрать из библиотеки эталон «своего» автора, то ближайшие к тексту эталоны не будут иметь свойства устойчивой идентификации текста при уменьшении длины фрагмента, тогда как правильный эталон такой устойчивостью обладает. Тогда, связав с каждым автором распределение вероятностей ошибочной идентификации фрагмента текста в зависимости от его длины, получим еще один набор распределений, с которыми будут сравниваться эмпирически получаемые вероятности ошибочной идентификации в предположении, что первичная идентификация проведена правильно.

Применительно к исследуемой задаче минимальная длина фрагмента выбирается равной 1 тыс. символов. Хотя эта малость избыточная, поскольку такие выборки не являются репрезентативными, она принята для большей полноты анализа. Таких идущих подряд неперекрывающихся фрагментов исходного текста из N знаков имеется $n = \left[\frac{N}{1000}\right] + 1$. Первые $n-1$ фрагментов имеют длину 1000 знаков, а последний фрагмент — длину $l = N - 1000 \cdot \left[\frac{N}{1000}\right]$. Если длина последнего фрагмента меньше половины минимальной (т. е. меньше 500 знаков), то этот фрагмент объединяется с предпоследним. Если же она больше 500 знаков, то фрагмент анализируется так же, как и остальные.

Для экономии вычислений библиотека прочитывается только один раз. Для каждого текста составляются распределения биграмм $f_k^{(1)}(j)$, $k = 1, 2, \ldots, n-1$, для фрагментов длины 1000 знаков. Из этих распределений затем составляются распределения $f_k^{(2)}(j)$ фрагментов длин 2000 $\left($таких фрагментов будет $\left[\frac{n-1}{2}\right]\right)$, $f_k^{(3)}(j)$ для длины 3000 и т. д. до того момента, пока количество фрагментов не станет равным двум (разрезание текста пополам).

Для фрагментов длины $s \cdot 1000$ распределения биграмм имеют вид

$$f_k^{(s)}(j) = \frac{1}{s}\left(f_{s(k-1)+1}^{(1)}(j) + f_{s(k-1)+2}^{(1)}(j) + \ldots + f_{sk}^{(1)}(j)\right), \quad k = 1, 2, \ldots, \left[\frac{n-1}{s}\right]. \tag{7}$$

Распределение $D(j)$ полного текста включает все фрагменты:

$$D(j) = \frac{1000}{N}\sum_{k=1}^{n-1} f_k^{(1)}(j) + \left(1 - (n-1)\frac{1000}{N}\right) f_n^{(1)}(j). \tag{8}$$

Затем по формуле (1) из распределений (8) составляются авторские эталоны.

3. Анализируемый корпус и результаты тестирования

Рассматривается корпус из литературных текстов на русском языке ста отечественных и зарубежных (в переводе) авторов, имеющих не менее шести произведений длиной не менее 30 тыс. знаков. Из этих авторов 40 оказались российскими, а 60 — иностранными. Общее количество рассмотренных текстов составило 1782 произведения суммарным объемом примерно 700 млн знаков. Поскольку полный перечень текстов с их атрибутами (автор, название, выходные данные издания, число символов, переводчик — если есть) занимает по объему более 100 страниц в формате данной статьи, то мы ограничимся только перечислением авторов в формате «Фамилия, Имя — количество текстов, доля ошибок идентификации». Авторы отбирались исходя из того, что они (а) известные писатели, (б) имеют не менее шести крупных произведений.

Айтматов Чингиз — 10, 0,0; Акунин Борис — 10, 0,0; Блаватская Елена — 7, 0,0; Донцова Дарья — 71, 0,014; Достоевский Фёдор — 10, 0,0; Дяченко Марина — 12, 0,083; Кристи Агата — 35, 0,0; Лавкрафт Говард — 9, 0,22; Лаврова Ольга — 20, 0,15; Лагин Лазар — 6, 0,67; Ладлэм Роберт — 27, 0,15; Лазарчук Андрей — 17, 0,0; Лайонз Вайолетт — 6, 0,33; Лайтман Михаэль — 25, 0,0; Ламли Брайан — 14, 0,29; Ламур Луис — 99, 0,04; Ларионова Ольга — 10, 0,2; Ластбадер Эрик — 20, 0,1; Латынина Юлия — 20, 0,1; Ле Гуин Урсула — 19, 0,32; Ле Карре Джон — 12, 0,17; Леблан Морис — 13, 0,077; Легостаев Андрей — 11, 0,18; Леженда Валентин — 8, 0,25; Лейбер Фриц — 17, 0,47; Лейстер Мюррей — 15, 0,13; Лем Станислав — 23, 0,39; Ленин Владимир — 10, 0,1; Леонов Николай — 31, 0,032; Лесков Николай — 7, 0,14; Ли Миранда — 6, 0,5; Ли Танит — 16, 0,19; Ли Шарон — 9, 0,0; Ли Эйна — 10; 0,5; Лимонов Эдуард — 9, 0,22; Линдгрен Астрид — 7, 0,29; Линдсей Джоанна — 43, 0,047; Липатов Вил — 12, 0,42; Липскеров Дмитрий — 9, 0,0; Литвинова Анна — 25, 0,0; Лиханов Альберт — 17, 0,24; Логинов Святослав — 11, 0,45; Ломер Кит — 23, 0,22; Лондон Джек — 24, 0,25; Лорд Джеффри — 33, 0,0; Лори Андре — 10, 0,0; Лоса Марио — 9, 0,44; Лоуелл Элизабет — 20, 0,1; Лоуренс Стефани — 15, 0,13; Лукин Евгений — 13, 0,15; Лукьяненко Сергей — 36, 0,19; Лэйтон Эдит — 7, 0,29; Лэнгтон Джоанна — 19, 0,16; Люис Клайв — 13, 0,62; Лютый Алексей — 9, 0,0; Маканин Владимир — 15, 0,067; Макбейн Лори — 8, 0,38; Макбейн Эд — 47, 0,085; Макдональд Грегори — 11, 0,0; Макдональд Джон — 13, 0,077; Макдональд Росс — 22, 0,0; Макеев Алексей — 21, 0,048; Маккаммон Роберт — 13, 0,077; Маккефри Энн — 25, 0,04; Маклин Алистер — 25, 0,28; Макнот Джудит — 19, 0,11; Малинин Евгений — 12, 0,17; Малышева Анна — 19, 0,0; Мамин-Сибиряк Дмитрий — 10, 0,1; Манн Генрих — 8, 0,12; Манн Томас — 13, 0,077; Маркес Габриэль — 10, 0,4; Марриет Фредерик — 21, 0,14; Март Михаил — 10, 0,0; Мартин Джордж — 23, 0,61; Мартин Кэт — 13, 0,077; Мартынов Георгий — 10, 0,1; Мартьянов Андрей — 14, 0,071; Марш Найо — 10, 0,1; Мейсон Конни — 24, 0,083; Мельников Руслан — 10, 0,2; Мельникова Ирина — 13, 0,15; Меррит Абрахам — 11, 0,36; Мерфи Уоррен — 72, 0,0; Мзареулов Константин — 14, 0,071; Милевская Людмила — 18, 0,0; Миллер Генри — 11, 0,18; Миллер Линда — 10, 0,2; Набоков Владимир — 10, 0,2; Полякова Татьяна — 9, 0,0; Сталин Иосиф — 17, 0,18; Тургенев Иван — 10, 0,0; Устинова Татьяна — 13, 0,0; Хайнлайн Роберт —

38, 0,13; Хайсмит Патриция — 7, 0,29; Хмелевская Иоанна — 50, 0,06; Холт Виктория — 27, 0,074; Чандлер Рэймонд — 11, 0,0; Чейни Питер — 9, 0,22; Шолохов Михаил — 7, 0,0.

Для каждого произведения описанного корпуса были построены распределения биграмм фрагментов (7). Были также построены авторские эталоны биграмм и распределения расстояний между распределениями символов в текстах и эталонах согласно методике, описанной выше. Выяснилось, что вероятность ошибочной идентификации автора отдельного текста составила 0,12.

На рис. 1 приведено совместное распределение расстояний «свой–чужой» для данного корпуса. Оси расстояний от текста до эталонов авторов разбиты на 20 классовых интервалов с шагом 0,03. Приведенная на рисунке поверхность показывает долю текстов, расстояние от которых до своего эталона попало в указанную ячейку «Свой», а до чужого — в ячейку «Чужой». Поскольку почти весь носитель этого распределения расположен справа от побочной диагонали квадрата, то это означает, что почти всегда расстояние до своего эталона меньше, чем до чужого.

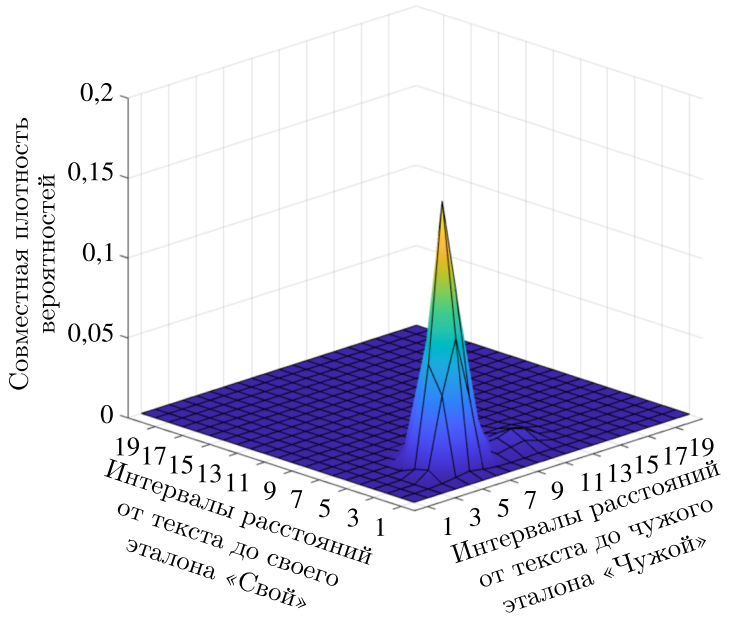

Рис. 1. Совместная плотность распределения расстояний биграмм между текстами и эталонами «свой–чужой»

Далее сделаем одно важное замечание, относящееся к адекватности применяемой гипотезы распознавания автора. Данный метод позволяет провести так называемый очищающий отбор. Он состоит в следующем. Сначала все ошибочно определяемые тексты исключались из библиотеки, причем если в результате у некоторого автора оказывалось менее шести произведений, то он целиком исключался из данного корпуса. После этого составлялись новые эталоны и идентификация проводилась снова. Разумеется, поскольку эталоны менялись, то опять появлялись тексты, авторы которых неверно идентифицировались. Такие тексты снова исключались. Важно подчеркнуть, что после шести описанных итераций в корпусе осталось 88 авторов (38 российских и 50 зарубежных) и 1452 текста, автор каждого из которых идентифицировался безошибочно. На этом очищенном корпусе и был проведен основной статистический эксперимент по идентификации фрагментов текстов в зависимости от их длины. Отметим, что во многих статистических моделях, не основанных на анализе содержательной части изучаемых систем, поэтапное исключение ошибочных результатов приводит к почти полному исключению элементов выборки: таковы, например, регрессионные модели.

Для текстов очищенного корпуса проводится фрагментация по формуле (7). После этого для i-го текста автора a определяется доля $\nu_i^a(\tau)$ фрагментов длины τ в тысячах знаков, автором которых является «свой» автор a. По совокупности n_a текстов данного автора можно построить эмпирическое дискретное распределение $\varphi_\tau^a(\nu)$ для каждой мелкости τ фрагментации текста. Например, если разбить всю область возможных значений $\nu \in [0;\ 1]$ на h промежутков с шагом $\frac{1}{h}$, то $\varphi_\tau^a\left(\frac{k}{h}\right)$, $k = 1, 2, \ldots, h$, есть эмпирическая частота доли (как случайной величины) правильно идентифицированных фрагментов длины τ, заключенной в промежутке $\nu \in \left[\frac{k-1}{h};\ \frac{k}{h}\right]$. При этом $\forall\tau$ $\sum\limits_{k=1}^{h} \varphi_\tau^a\left(\frac{k}{h}\right) = 1$.

В более общем случае строится совместное распределение $f^b(\tau, \nu)$, характеризующее автора b. Это распределение строится по данным $\nu_i^b(\tau)$ по достоверным текстам автора b. Область значений (в долях) $\tau \in [0;\ 1]$ разбивается на L классовых интервалов, после чего определяется, какая доля произведений автора b попала в ячейку $\left[\frac{n-1}{L};\ \frac{n}{L}\right] \times \left[\frac{k-1}{h};\ \frac{k}{h}\right]$ носителя распределения $f^b(\tau, \nu)$. Эта доля и есть собственно эмпирическая оценка этого распределения в виде $f^b\left(\frac{n}{L}, \frac{k}{h}\right)$. На рис. 2 приведен пример такого распределения для автора Луиса Ламура по данным 98 текстов. Видно, что ошибка достаточно мала уже на долях текста порядка 0,01, то есть на 3–4 тыс. знаков.

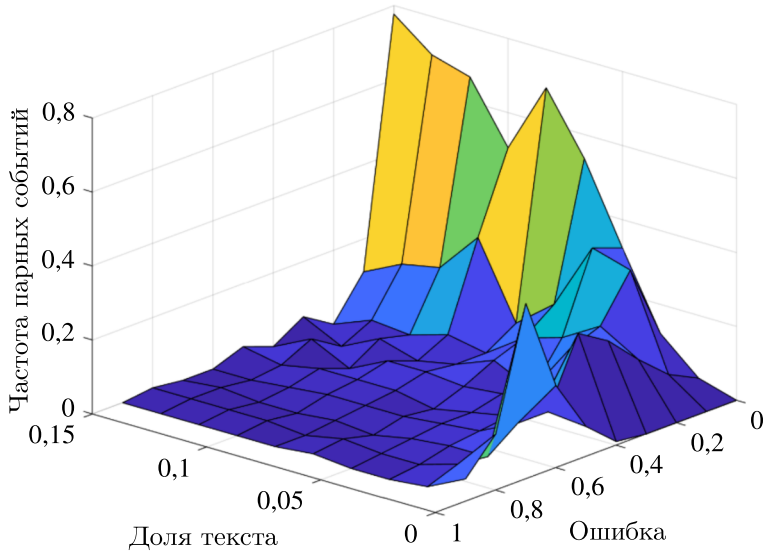

Рис. 2. Совместное распределение вероятности ошибки и доли фрагмента

По текстам из очищенного корпуса определяется такая последовательность фрагментаций $\{\tau_0, \tau_0 + 1, \ldots, \tau_0 + s\}$ в тысячах знаков, для которой при каждом τ доля правильно идентифицируемых фрагментов $\nu_i^a(\tau)$ отлична от нуля. Например, выяснилось, что если разделить полный текст на два равных фрагмента, то с вероятностью примерно 0,95 автор полного текста идентифицируется также как автор обоих фрагментов, но существуют примеры, когда он не опознается для одного или для обоих фрагментов. В то же время на каждом уровне фрагментации $\frac{1}{4}$, $\frac{1}{16}$, $\frac{1}{32}$ от полного текста правильный автор всегда присутствует как автор хотя бы одного фрагмента. Этот эмпирический факт можно использовать для корректировки результата распознавания полного текста. В качестве иллюстрации идеи на диаграмме рис. 3 приведены эмпирические частоты, с которыми чужие авторы встречаются как авторы хотя бы одного фрагмента по пересечению ответов для всех трех уровней фрагментации.

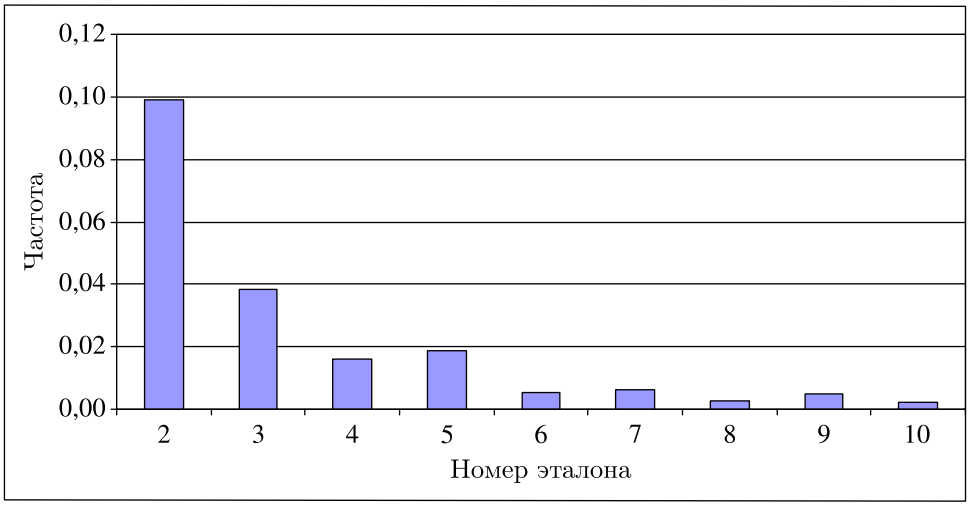

Рис. 3. Частота встречаемости чужого автора как автора фрагмента на каждом из трех уровней фрагментации в зависимости от номера эталона этого автора по отношению к полному тексту

Из рис. 3 видно, что чем ближе эталон чужого автора к тексту, тем больше вероятность ошибочной идентификации именно этого автора. Хотя этот результат кажется вполне естественным, он косвенно подтверждает адекватность отнюдь не очевидной концепции распознавания автора по близости к эталонному распределению по символам (не по словам!) как к генеральной совокупности.

Заметим теперь, что по неправильно идентифицированным полным текстам тот автор, который назначен первым, присутствует не в 90 % фрагментированных текстов, как правильный автор в очищенном корпусе, а всего в 47 %. Тем самым относительно примерно половины ошибочных случаев идентификации можно будет сделать корректирующий вывод о том, что распознавание не признается верным.

Следовательно, данный метод коррекции может быть использован также в качестве индикатора того, что автор предположительно отсутствует в данной библиотеке эталонов. Допустим, что некоторый текст был распознан как текст автора a. На очищенном корпусе этот результат является достоверным. Но если исключить мысленно из этого корпуса данного автора a, то автором рассматриваемого текста будет назначен следующий за a автор b, т. е. второй ближайший эталон к данному тексту. Разделив затем изучаемый текст на фрагменты длин $\{\tau_0, \tau_0 + 1, \ldots, \tau_0 + s\}$, находим доли $\nu^b(\tau_0)$, $\nu^b(\tau_0 + 1)$ и т. д., отвечающие распознаванию части фрагментов как текстов автора b. Если какая-нибудь одна из этих величин окажется нулевой, то мы сделаем вывод, что исходное распознавание полного текста было ошибочным.

В целом по корпусу «очищенных» текстов свой автор довольно часто оказывается и самым идентифицируемым автором фрагментов. Однако не следует полагать, что фрагментация может заменить собой по точности полный текст. На рис. 4 приведена зависимость ошибки распознавания автора по самому частому автору фрагментов в зависимости от длины фрагмента. Видно, что с уменьшением длины фрагмента ошибка возрастает, то есть находятся такие тексты, фрагменты которых чаще опознаются чужими эталонами. Даже при делении текста пополам нашелся один такой пример, когда ни одна из частей не опознается своим автором.

Таким образом, если снабдить результат распознавания критерием «принять как верный» или «отклонить как неверный» по подтверждению вышеописанного второго индикатора, то все правильно идентифицированные тексты останутся таковыми, а среди остальных будут верно выявлены тексты, авторы которых опознаны неправильно. Анализ показал, что в основном ошибочно опознается один текст из набора произведений какого-либо автора, и этот текст отстоит

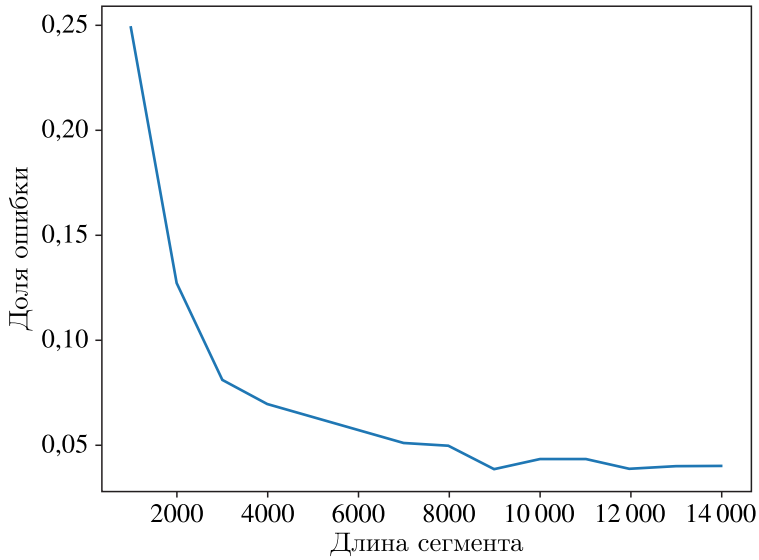

Рис. 4. Ошибка идентификации автора текста, определяемого по наиболее частому распознаваемому автору у фрагментов, для правильно идентифицированного полного текста

от соответствующего эталона достаточно далеко. Следовательно, такой текст написан заметно иначе, чем все прочие произведения данного автора, и потому естественно считать, что эталона такой новой авторской ипостаси нет в библиотеке. «Очищенный корпус» дает возможность оценки точности вывода о том, что «автора нет в библиотеке». Если исключить из рассмотрения ближайший к тексту эталон, то ошибкой проверяемого утверждения будет являться доля вторых по близости эталонов, которые при фрагментации текста остаются ближайшими, то есть устойчивыми. Доля таких ситуаций оказалась равной 0,069. Распространяя этот результат на весь исходный корпус текстов, приходим к выводу, что доля ошибочных принятий решений снизится до примерно 7 % вместо исходных 12 %. Следовательно, предложенный фильтр, основанный на фрагментации исходного текста, позволяет примерно вдвое снизить ошибку неверного распознавания автора текста.

4. Заключение

В работе предложен новый метод коррекции результатов машинной идентификации автора текста. Этот метод использует в определенном смысле идею размножения выборок, что позволяет построить корректирующий индикатор без ухудшения точности основного индикатора. Проблема коррекции связана с тем, что результатом применения основного индикатора является не статистика, а одно число, которое далее интерпретируется в рамках метода ближайшего соседа. Фрагментация текста позволяет получить статистику потенциальных авторов. И хотя сама по себе фрагментация ухудшает точность распознавания, она тем не менее позволяет использовать в качестве индикатора новый показатель, а именно устойчивость своего автора и неустойчивость чужого. В силу приближенно равномерной случайности ошибок распознавания, которые связаны с достаточно далекими от правильного автора эталонами (шестой и далее), ошибки такого рода могут быть исключены с помощью фрагментации.

Данный метод позволяет также оценить вероятность того, что эталон правильного автора отсутствует в библиотеке. Точность этой гипотезы примерно такая же, как и точность метода идентификации правильного автора, т. е. 7 %. Заметное снижение ошибки на достаточно большом корпусе текстов показывает, что данный метод весьма эффективен. При этом мы не обсуждаем причин, по которым реально присутствующий в библиотеке автор не опознается верно.

Однако предварительный анализ показывает, что большинство таких неправильно распознаваемых текстов расположено далеко от всех авторов, а не только от своего, и лишь случайно один из них оказался ближайшим соседом. Это можно понимать и так, что автор написал текст в манере, отличной от других своих произведений, и в этом смысле он выступает как некоторый новый автор, эталон которого действительно отсутствует в библиотеке. Вопрос кластеризации текстов внутри произведений одного и того же автора представляет собой тему специального исследования, но уже сейчас можно сказать, что в ряде случаев такая кластеризация действительно приводит к улучшению распознавания. В частности, это относится к так называемым сериальным писателям, которые выпускают, например, несколько фантастических романов в жанре звездных войн, а затем пишут ряд исторических романов или фэнтези. Именно поэтому мы трактуем верное отклонение результата ошибочной идентификации как коррекцию. Дальнейшее исследование позволит формализовать этот подход в терминах расстояний до эталонных двумерных распределений авторов.

Список литературы (References)

Батура Т. В. Методы автоматической классификации текстов // Программные продукты и системы. — 2017. — Т. 30, № 1. — С. 85–99.
Batura T. V. Metody avtomaticheskoi klassifikatsii tekstov [Methods of automatic texts classification] // Programmnye produkty i sistemy. — 2017. — Vol. 30, No. 1. — P. 85–99 (in Russian).

Кукушкина О. В., Поликарпов А. А., Хмелёв Д. В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. — 2001. — Т. 37, вып. 2. — С. 96–109.
Kukushkina O. V., Polikarpov A. A., Khmelev D. V. Opredelenie avtorstva teksta s ispol'zovaniem bukvennoi i grammaticheskoi informatsii [Recognition of the text author with the use of letter and grammar information] // Problemy peredachi informatsii. — 2001. — Vol. 37, No. 2. — P. 96–109 (in Russian).

Орлов Ю. Н., Осминин К. П. Методы статистического анализа литературных текстов. — М.: Эдиториал УРСС, 2012. — 312 с.
Orlov Yu. N., Osminin K. P. Metody statisticheskogo analiza literaturnykh tekstov [Methods of statistical analysis of the literature texts]. — Moscow: Editorial URSS, 2012. — 312 p. (in Russian).

Резанова З. И., Романов А. С., Мещеряков Р. В. О выборе признаков текста, релевантных в автороведческой экспертной деятельности // Вестник Томского государственного университета. Филология. — 2013. — Т. 26, № 6. — С. 38–52.
Rezanova Z. I., Romanov A. S., Mescheriakov R. V. O vybore priznakov teksta, relevantnyh v avtorovedcheskoi i ekspertnoi dejatelnosti [On the texts attributes choice for the expertize and author identification] // Bulletin of Tomsk State University. Philology. — 2013. — Vol. 26, No. 6. — P. 38–52 (in Russian).

Хмелёв Д. В. Распознавание автора текста с использованием цепей А. А. Маркова // Вестник МГУ. Сер. 9: Филология. — 2000. — № 2. — С. 115–126.
Khmelev D. V. Raspoznavanie avtora teksta s ispol'zovaniem tsepei A. A. Markova [Recognition of the text author with the use of Markov chains] // Vestnik MGU. Ser. 9: Filologiya. — 2000. — No. 2. — P. 115–126 (in Russian).

Argamon S., Juola P. Overview of the international authorship identification competition at PAN-2011 // Conference: CLEF 2011 Labs and Workshop, Notebook Papers. — Amsterdam, 2011. — P. 1–10.

Cappellato L., Ferro N., Halvey M. et al. CLEF 2014 Labs and Workshops // Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org). — 2014.

Roscher R., Bohn B., Duarte M. F., Garcke J. Explainable mashine learning for scientific insights and discoveries // IEEE Access. — 2020. — Vol. 8. — P. 42200–42216.

Stamatatos E., Fakotakis N., Kokkinakis G. Automatic text categorization in terms of genre and author // Computational Linguistics. — 2000. — Vol. 26 (4). — P. 471–495.

Sudheep E. M., Chinchu J., Puthussery A., Sasi N. K. Text classification for authorship attribution analysis // Advanced Computing: An International Journal (ACIJ). — 2013. — Vol. 4, No. 5. — P. 1–10.