КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2020 Т. 12 № 1 С. 243–254
DOI: 10.20537/2076-7633-2020-12-1-243-254

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 519.25

Статистический анализ биграмм специализированных текстов

Н. А. Митин[a], Ю. Н. Орлов

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

E-mail: [a] mitin@keldysh.ru

Получено 21.08.2019, после доработки — 24.11.2019.
Принято к публикации 26.11.2019.

Метод спектрального анализа стохастической матрицы применяется для построения индикатора, позволяющего определять тематику научных текстов без использования ключевых слов. Эта матрица представляет собой матрицу условных вероятностей биграмм, построенную по статистике используемых в тексте символов алфавита без учета пробелов, цифр и знаков препинания. Научные тексты классифицируются по взаимному расположению инвариантных подпространств матрицы условных вероятностей пар буквосочетаний. Индикатор разделения — величина косинуса угла между правым и левым собственными векторами, отвечающими максимальному и минимальному собственным значениям. Вычислительный алгоритм использует специальное представление параметра дихотомии, в качестве которого выступает интеграл от нормы квадрата резольвенты стохастической матрицы биграмм по окружности заданного радиуса в комплексной плоскости. Стремление интеграла в бесконечность свидетельствует о приближении контура интегрирования к собственному значению матрицы. В работе приведены типовые распределения индикатора идентификации специальностей. Для статистического анализа были проанализированы диссертации по основным 19 специальностям ВАК без учета классификации внутри специальности, по 20 текстов на специальность. Выяснилось, что эмпирические распределения косинуса угла для физико-математических и гуманитарных специальностей не имеют общего носителя, поэтому могут быть формально разделены по значению этого индикатора без ошибки. Хотя корпус текстов был не особенно большой, тем не менее при произвольном отборе диссертаций ошибка идентификации на уровне 2 % представляется очень хорошим результатом по сравнению с методами, основанными на семантическом анализе. Также выяснилось, что можно составить паттерн текста по каждой из специальностей в виде эталонной матрицы биграмм, по близости к которой в норме суммируемых функций можно безошибочно идентифицировать тематику написанного научного произведения, не используя ключевые слова. Предложенный метод можно использовать и в качестве сравнительного индикатора большей или меньшей строгости научного текста или как индикатор соответствия текста определенному научному уровню.

Ключевые слова: стохастическая матрица, спектральный портрет, статистический индикатор, научный текст

© 2020 Николай Алексеевич Митин, Юрий Николаевич Орлов
Статья доступна по лицензии Creative Commons Attribution-NoDerivs 3.0 Unported License.
Чтобы получить текст лицензии, посетите веб-сайт http://creativecommons.org/licenses/by-nd/3.0/
или отправьте письмо в Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

COMPUTER RESEARCH AND MODELING
2020 VOL. 12 NO. 1 P. 243–254
DOI: 10.20537/2076-7633-2020-12-1-243-254

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 519.25

Statistical analysis of bigrams of specialized texts

N. A. Mitin[a], Yu. N. Orlov

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

E-mail: [a] mitin@keldysh.ru

Received 21.08.2019, after completion — 24.11.2019.
Accepted for publication 26.11.2019.

The method of the stochastic matrix spectrum analysis is used to build an indicator that allows to determine the subject of scientific texts without keywords usage. This matrix is a matrix of conditional probabilities of bigrams, built on the statistics of the alphabet characters in the text without spaces, numbers and punctuation marks. Scientific texts are classified according to the mutual arrangement of invariant subspaces of the matrix of conditional probabilities of pairs of letter combinations. The separation indicator is the value of the cosine of the angle between the right and left eigenvectors corresponding to the maximum and minimum eigenvalues. The computational algorithm uses a special representation of the dichotomy parameter, which is the integral of the square norm of the resolvent of the stochastic matrix of bigrams along the circumference of a given radius in the complex plane. The tendency of the integral to infinity testifies to the approximation of the integration circuit to the eigenvalue of the matrix. The paper presents the typical distribution of the indicator of identification of specialties. For statistical analysis were analyzed dissertations on the main 19 specialties without taking into account the classification within the specialty, 20 texts for the specialty. It was found that the empirical distributions of the cosine of the angle for the mathematical and Humanities specialties do not have a common domain, so they can be formally divided by the value of this indicator without errors. Although the body of texts was not particularly large, nevertheless, in the case of arbitrary selection of dissertations, the identification error at the level of 2 % seems to be a very good result compared to the methods based on semantic analysis. It was also found that it is possible to make a text pattern for each of the specialties in the form of a reference matrix of bigrams, in the vicinity of which in the norm of summable functions it is possible to accurately identify the theme of the written scientific work, without using keywords. The proposed method can be used as a comparative indicator of greater or lesser severity of the scientific text or as an indicator of compliance of the text to a certain scientific level.

Keywords: stochastic matrix, spectral portrait, statistical indicator, scientific text

Citation: *Computer Research and Modeling*, 2020, vol. 12, no. 1, pp. 243–254 (Russian).

© 2020 Nikolay A. Mitin, Yuriy N. Orlov
This work is licensed under the Creative Commons Attribution-NoDerivs 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/3.0/
or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

1. Введение

В работе приведены результаты статистического эксперимента по анализу текстовой информации, относящейся к широкому кругу научных отраслей знаний. Анализ описываемого далее эксперимента показал, что тексты научных специальностей могут быть достаточно точно классифицированы не только по экспертно определяемым ключевым словам, но и посредством анализа математических объектов, характеризующих частоту и устойчивость распределения буквосочетаний. В частности, в данной работе анализируется структура пространства, образованного собственными векторами матрицы двухбуквенных сочетаний (биграмм). Исследовалась возможность тематической кластеризации текстов только на основе анализа вероятностных распределений буквосочетаний, т. е. чисто машинным методом без участия эксперта.

Цель эксперимента состояла в том, чтобы исследовать различия между так называемыми спектральными портретами матриц условных вероятностей биграмм для текстов основных специальностей, формально установленных ВАК. Этих специальностей 19: математика и физика, химия, биология, машиноведение и математическое моделирование, земледелие, история, экономика, философия, языкознание, юриспруденция, педагогика, медицина, искусствоведение, психология, социология, политология, культурология, геология, теология. В принципе, можно было бы провести более детальный эксперимент, погружаясь в подразделы указанных специальностей, но на данном этапе исследований демонстрируется работоспособность метода на верхнем уровне классификации.

Следует отметить, что в области машинной классификации текстов существует большое количество подходов, основанных в той или иной степени на семантическом анализе. В частности, вопросы автоматической обработки научной информации и выделения определенной информации из текстов с целью выявления общих тем исследований и полученных результатов применительно к медицине обсуждались в работах [Bekhuis, Demner-Fushman, 2012; Kim et al., 2011]. Методы кластеризации текстов по общности постановок задач с помощью опорных векторов и байесовской нейронной сети сравнивались в [Park, Blake, 2012] на достаточно большом корпусе текстов. Автоматическая классификация научных тестов по постановкам задач и по их решениям на основе анализа тэгов, предложенная довольно давно в работе [Chandrasekaran, 1983], имеет различные варианты развития на современном этапе (см., например, [Charles, 2011]). Однако авторы всех работ сходятся в том, что процент ошибок любых применяемых семантических методов варьируется от 20 до 30 %.

Метод, применяемый в настоящей статье, не относится к семантическим. В нем анализируются вероятности буквосочетаний в текстах определенной тематики, после чего, например, для матрицы частот биграмм строится ее спектр и определяются собственные векторы для двух специальных собственных значений. Угол между ними является индикатором принадлежности текста определенной тематической группе.

Хотя наиболее часто используемым методом для оценки спектра стохастической матрицы является построение кругов Гершгорина (см. [Yuan Lu, 2010; Kirkland, 2009]), в нашей работе используется так называемый ε-спектр, вычисляемый по процедуре, описанной в [Годунов, 1997; Голуб, Ван Лоун, 1999].

В результате проведенного анализа применительно к российским диссертациям выяснились два обстоятельства. Во-первых, тексты по каждой специальности могут быть кластеризованы по близости к некоторому эталонному распределению вероятностей биграмм, характерному для каждой специальности, причем кластеризация эта имеет весьма точное соответствие с формальной атрибутикой текстов (учебников и диссертаций). Ошибка идентификации научных текстов из корпуса, в состав которого вошли 400 текстов разных специальностей, составила всего 2 %. При этом тексты естественно-научных специальностей идентифицировались безошибочно. Во-вторых, был найден индикатор, имеющий численное выражение, распределение которого характерно для текстов определенной специальности, что позволило провести унифицированное ранжирование научных тематик.

Описываемый в работе вычислительный эксперимент проводился на программном комплексе TRIL [Орлов, Осминин, 2017], разработанном в ИПМ им. М. В. Келдыша РАН. Отметим здесь, что этот комплекс был разработан для определения авторства текста, в основе анализа лежит идея построения паттерна функции распределения n-грамм, характерных для того или иного автора. Тестирование метода, проведенное в [Орлов, Осминин, 2012; Борисов и др., 2017], показало, что метод паттернов распределений буквосочетаний не особенно эффективен для определения жанра произведения, в связи с чем для тематической классификации следовало бы разработать другой подход. В настоящей работе и предложен такой индикатор.

2. Основные понятия и постановка задачи

Рассматривается достаточно большой (более 10 тыс. знаков) текст на русском языке, из которого исключены пробелы, знаки препинания, цифры, формулы и прочие символы, кроме букв русского алфавита, причем строчные и прописные буквы не различаются. Получившаяся совокупность символов рассматривается как линейное пространство, элементы которого суть 33-мерные векторы $\mathbf{f}(t)$, представляющие собой вероятности того, что в момент t в данном тексте реализовалась одна из 33 букв алфавита. Временем в таком пространстве служит порядковый номер конкретной буквы. Введем также матрицу $\mathbf{F}(t)$ вероятностей парных буквосочетаний. Рассмотрим стохастическую матрицу условных вероятностей того, что в тексте после символа i следует символ j, элементы которой связаны с однобуквенными и двухбуквенными частотами формулой

$$P_{ij} = \frac{F_{ij}}{f_i}, \quad f_i = \sum_{j=1}^{33} F_{ij}. \tag{1}$$

Согласно результатам, представленным в [Орлов, Осминин, 2012; Борисов и др., 2017], собственные векторы матрицы P из (1), отвечающие действительным собственным значениям, являются устойчивыми индикаторами автора литературного текста. Насколько метод собственных векторов применим для анализа текстов, имеющих специализированный научный характер?

При анализе спектра стохастической матрицы одним из основных вопросов является оценка того, в каких пределах лежит возмущение спектра матрицы при малом возмущении ее элементов. По определению [Годунов, 1997], комплексное число λ принадлежит ε-спектру $\Lambda_\varepsilon(P)$ матрицы P, если существует такая возмущающая ее матрица Δ, что $\|\Delta\| \le \varepsilon \|P\|$ и $\det(\lambda I - P - \Delta) = 0$, где I — единичная матрица. В исследуемом нами случае анализа текстов возмущением является статистическая неопределенность частот буквосочетаний, вычисленная по конечному объему данных — конкретному тексту.

Вычислительный алгоритм [Голуб, Ван Лоун, 1999], с помощью которого строится спектральный портрет матрицы, основан на специальном представлении параметра дихотомии:

$$H_r(P) = \frac{1}{2\pi} \int_0^{2\pi} \left(P^+ - re^{-i\phi} I\right)^{-1} \left(P - re^{i\phi} I\right)^{-1} d\phi. \tag{2}$$

Интеграл в (2) сходится только в том случае, если на окружности $\lambda = re^{i\phi}$ нет собственных значений матрицы P. Спектральный портрет образован изолиниями параметра дихотомии.

На рис. 1 в качестве типичного примера приведен спектральный портрет оператора P, отвечающего книге [Орлов, Осминин, 2012] одного из авторов данной работы. На этом рисунке линии уровня соответствуют областям, внутри которых находятся собственные значения при определенной ошибке в элементах матрицы. Величина ошибки вследствие возмущения или статистической неопределенности элементов матрицы указана в легенде справа от рисунка.

Например, линии желтого контура ограничивают область, внутри которой оказывается собственное число, если относительная ошибка в элементах матрицы равна $\varepsilon = 10^{-2}$.

Из рис. 1 видно, что при точности до 10^{-3} имеются фактически три достоверно различающихся собственных значения: очевидный корень $\lambda_1 = 1$, отрицательный корень, равный приблизительно $\lambda_2 \approx -0.55$, и некоторый «коллективный» ноль радиусом примерно 0.3. Более детальное представление спектра требует точности входных данных на уровне 10^{-4}, для чего, согласно [Орлов, Осминин, 2012], требуется длина текста порядка 10 млн знаков. Текстов такой длины мы в данной работе не рассматриваем, поэтому соответствующая детализация точек спектра не является достоверной и показана лишь для полноты картины.

Оказалось, что собственное значение λ_2 достаточно устойчиво по текстам и авторам, тогда как другие собственные значения, аккумулированные в окрестности нуля, не обладают таким свойством.

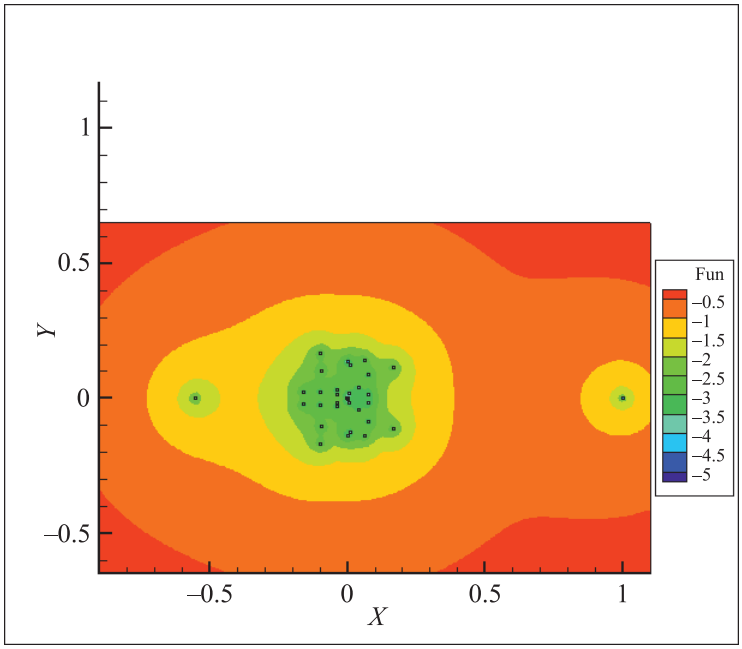

Рис. 1. Спектральный портрет текста монографии [Орлов, Осминин, 2012]; *X*, *Y* — координаты собственных значений матрицы условных биграмм

Левый собственный вектор $\mathbf{u}$, отвечающий наибольшему по модулю отрицательному собственному значению, образует угол с вектором частот $\mathbf{f}$, абсолютная величина косинуса которого для большинства текстов лежит в диапазоне от 0.07 до 0.09. Этот вектор $\mathbf{u}$, как и вектор $\mathbf{f}$, обладает авторской устойчивостью (см. [Орлов, Осминин, 2012; Борисов и др., 2017]).

Поскольку векторы $\mathbf{f}$ и $\mathbf{u}$ приближенно ортогональны, их можно трактовать как главные направления матрицы условных вероятностей биграмм:

$$(u, Pf) \approx 0. \tag{3}$$

В частности, применительно к структуре матрицы биграмм, отвечающей книге [Орлов, Осминин, 2012], оказалось, что $|\cos(\mathbf{u},\mathbf{f})| = 0.03$. Это значение в три раза меньше характерной величины модуля косинуса для литературных текстов. Чтобы выяснить, случайно ли такое сравнительно большое различие, мы вычислили указанный индикатор для нескольких профессиональных текстов — диссертаций физико-математического содержания. Оказалось, что для всех этих текстов модуль косинуса заключен в пределах от 0.025 до 0.040.

Найденное различие между литературными и техническими текстами предположительно обусловлено следующим обстоятельством. Заметим, что точность, с которой вычисляются собственные векторы, зависит от точности вероятностей буквосочетаний, образующих матрицу P_{ij}. Для научных текстов по физике и математике характерен не очень большой объем используемых слов по сравнению с литературными произведениями, причем специфические словосочетания (например, дифференцируемость функции, уравнение теплопроводности и т. п.) имеют тенденцию повторяться. В результате частая повторяемость определенных буквосочетаний обеспечивает высокую точность элементов матрицы даже на относительно небольших текстах, так что рассматриваемые собственные векторы оказываются определенными гораздо лучше, чем для художественной литературы. На рис. 2 приведен характерный пример спектрального портрета для классического литературного произведения.

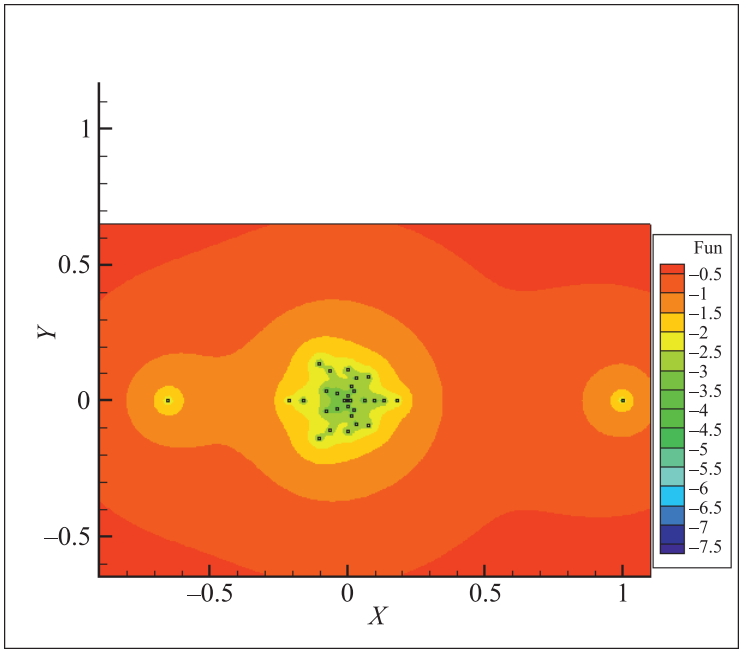

Рис. 2. Спектральный портрет романа И. С. Тургенева «Отцы и дети»; X, Y — координаты собственных значений матрицы условных биграмм; $\left|\cos\left(\mathbf{u},\mathbf{f}\right)\right| = 0.09$

Как показано в [Орлов, Осминин, 2012], авторские спектральные портреты сохраняют индивидуальность, характеризующую писателя: их форма приблизительно совпадает для разных произведений одного и того же автора.

А как обстоит в этом смысле дело с научными текстами?

3. Индикатор кластеризации научных текстов

Оказалось, что распределения косинуса угла для текстов по техническим наукам не имеют общего носителя с распределением косинуса для литературных произведений (см. рис. 3, красный и черный графики). Это наблюдение привело авторов настоящей работы к идее проанализировать тексты основных специальностей ВАК. Были отобраны в среднем по 20 текстов разных авторов из каждой научной области, построены соответствующие матрицы условных вероятностей биграмм и вычислены абсолютные величины косинусов углов между векторами главных направлений. Последующие выводы относятся, естественно, только к анализируемому корпусу текстов, собранному, впрочем, без предпочтений к определенным авторам и тематикам исследований.

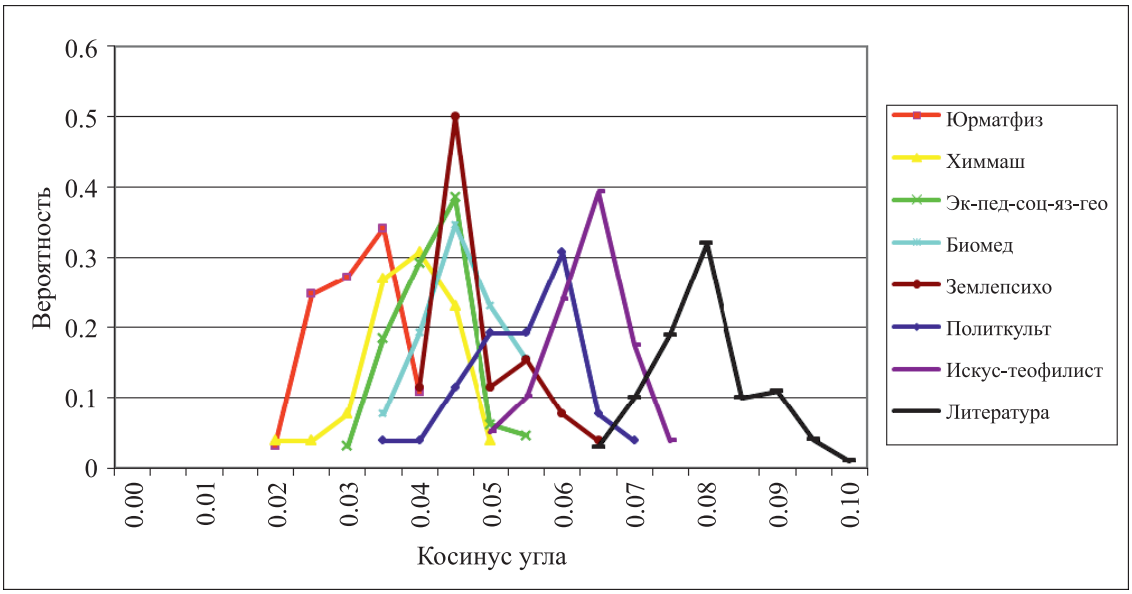

Рис. 3. Распределения модуля косинуса угла между собственными векторами главных направлений для текстов различных специальностей

Выяснилось, что существует достаточно четкое разделение специализированных текстов по данному параметру, который может считаться индикатором определенной тематической классификации. Так, физико-математическим наукам отвечает косинус порядка 0.025–0.035, химическим — 0.035–0.045, политическим — 0.045–0.065, философским — 0.065–0.075. Распределения частично перекрываются, но важно, что существуют специальности с нулевым пересечением носителя распределения данного параметра. В целом тексты гуманитарного характера весьма четко отделяются от технических текстов.

Можно упорядочить специальности ВАК по среднему косинусу. Соответствующая диаграмма приведена на рис. 4.

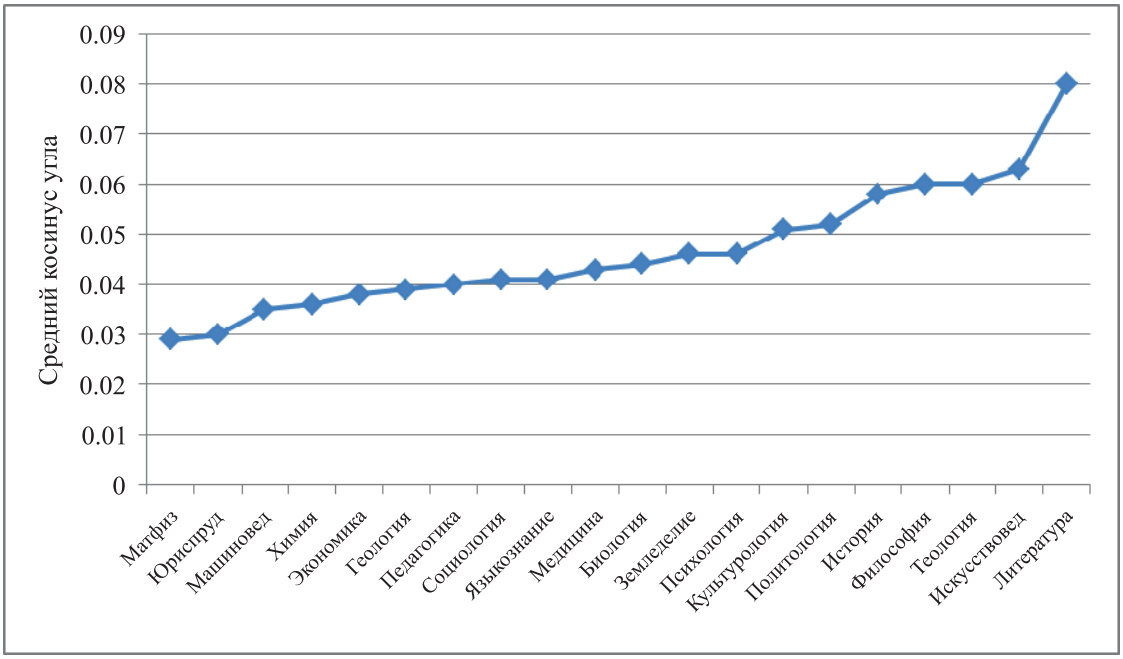

Рис. 4. Упорядочение специальностей ВАК (+ художественная литература) по среднему косинусу угла между главными направлениями матрицы условных вероятностей биграмм

На основе проведенного эксперимента предлагается следующий способ ранжирования специальностей ВАК. Считаем, по определению, математику наукой. Тогда максимальный косинус математического текста (0.040) определяет правую границу научности в смысле строгости изложения, характерной для научно-технических текстов. Левая же граница условной ненаучности (нестрогости) определяется минимальным значением косинуса угла литературных текстов по произведениям русских классиков (0.065). Специальности ВАК естественно ранжировать по доле распределения косинуса, лежащего в области научности, т. е. это значение функции распределения косинуса угла данной специальности в точке 0.04.

Оказалось, что к точным наукам кроме математики и физики следует отнести также и юриспруденцию, которая по стилю изложения (отсылки к законам по аналогии использования определений, лемм или теорем) близка к математике, а вот специальность 05 (математическое моделирование и машиноведение) частично выходит за выделенную границу.

Были отобраны девять специальностей, отдельные тексты которых написаны в научном стиле (рис. 5), т. е. за этими специальностями можно закрепить научный статус, если избавиться от присутствующих в них текстов, написанных недостаточно строго.

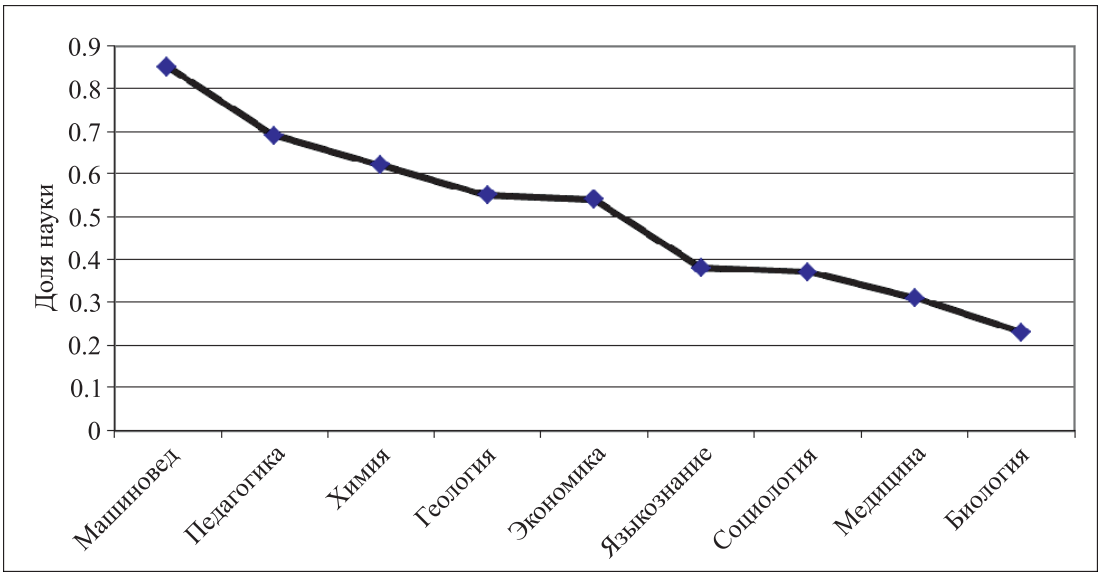

Рис. 5. Упорядочение специальностей ВАК по доле строгости в рассуждениях

Из диаграммы рис. 5 видно, что снижение доли научности изложения не связано напрямую с изменением тематики с технической на гуманитарную. По-видимому, науки, в которых еще достаточно велика описательная составляющая, такие как медицина и биология, требуют известного развития до достижения определенной строгости в манере описания. Отсутствие такой строгости позволяет отдельным авторам писать тексты, например, не «по медицине», а «о медицине», что несколько девальвирует научную составляющую всего корпуса текстов по данной тематике. Разумеется, можно считать, что каждая специальность уникальна и не должна быть похожа, например, на математику. Однако в таком случае следовало бы ожидать, что специальности отделены одна от другой по распределению индикатора, тогда как на рис. 3 видны значительные перекрытия распределений, дрейфующих вправо по мере уменьшения строгости в рассуждениях. Наукой же принято считать не вообще писательскую деятельность людей, а такую деятельность, которая использует «научный метод», т. е. алгоритмизируемый метод рассуждений, не зависящий от конкретного рассуждающего субъекта. В этой связи представляет интерес ранжирование того, что официально считается наукой, по степени использования научного метода в совокупности работ по каждой специальности.

Интересно, что существуют четыре специальности, распределения указанного параметра для которых целиком лежат между наукой и литературой, не имея с этими двумя общего носи-

теля: это земледелие, психология, политология, культурология. Этот факт может быть интерпретирован как аргумент в пользу того, что существуют такие области знаний, которые не являются в принципе ни наукой, ни литературой, а отвечают определенным объективно существующим проявлениям различных субъективных аспектов бытия социума.

Кроме того, имеются четыре специальности, которые следует трактовать не как науки, ибо доля науки в их распределениях нулевая, а как «недостаточно художественная» литература — по доле распределения косинуса угла в литературной области, определяемой произведениями русских классиков (нижняя граница косинуса 0.065). В этом смысле оказалось, что история и философия — это, в сущности, не научные специальности, а плохо (с точки зрения писательского мастерства) написанные литературные произведения, а искусствоведение — это «почти литература» (рис. 6).

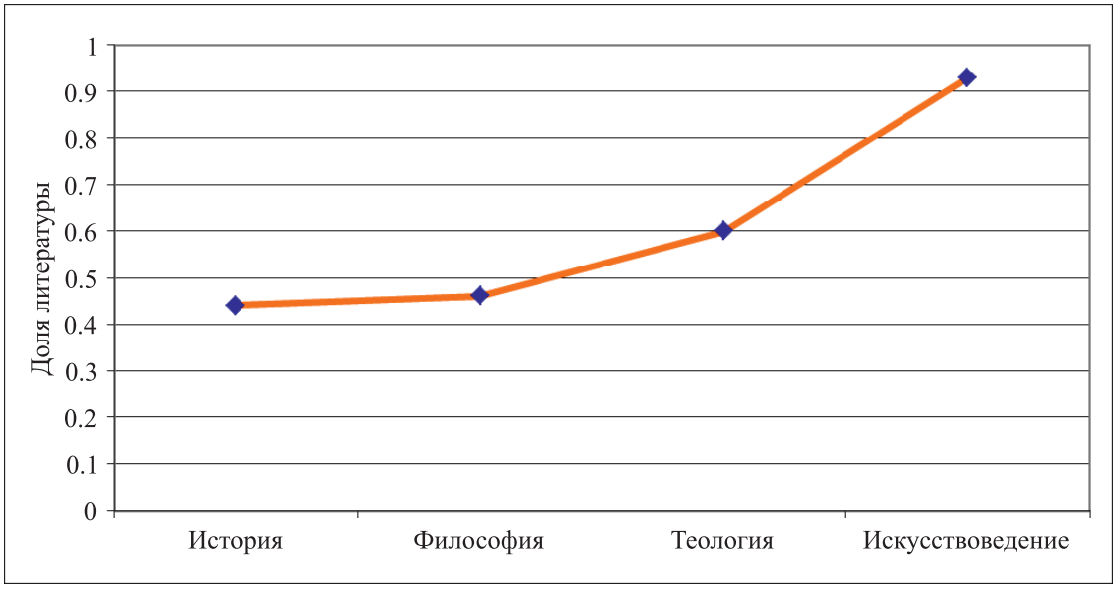

Рис. 6. Антиупорядочение специальностей ВАК по доле литературности в произведениях

Найденный тематический индикатор может использоваться и для оценки качества научной публикации. Например, если некоторый текст имеет косинус на уровне 0.03, то можно сказать, что он написан как математическая, физическая или юридическая работа, а не так, как художественная литература, и даже не так, как медицинский справочник. И наоборот, если книга, претендующая на то, чтобы быть учебным пособием по некоторой технической специальности, имеет косинус порядка 0.07, то это однозначно не учебник, а в лучшем случае научно-популярная литература. Интересно отметить, что так называемые «сериальные писатели» любовных романов, иронических детективов и боевиков имеют косинус угла порядка 0.06, поэтому такие тексты относятся скорее не к художественной литературе, а к публицистике. Также добавим, что небезызвестный «Корчеватель: алгоритм типичной унификации точек доступа и избыточности» имеет косинусный индикатор, равный 0.07, т. е. это текст, характерный для «плохой литературы», а не научного труда по технической специальности.

4. Статистика расстояний между научными текстами

Следует подчеркнуть, что отбор научных текстов в соответствии с той или иной специальностью, к которой они априори были отнесены в соответствии с индексацией УДК, был скорректирован, причем не экспертным образом, а в результате машинной кластеризации. Анализировались только те тексты, которые не только имели формальную тематическую принадлежность определенной специальности, но и попали в общий кластер, для которого было

построено эталонное распределение биграмм. Эталон составлялся следующим итерационным способом. Пусть, например, имеется N1 текстов по некоторой специальности C1 и N2 по специальности C2. Каждый из этих текстов сравнивается с эталонами f_1 и f_2, отвечающими средневзвешенным распределениям вероятностей биграмм, построенным по текстам каждой из специальностей, причем на момент сравнения данный конкретный текст исключается из состава эталона. Если все тексты при таком тестировании были отнесены к тем специальностям, которым соответствовали их первоначальные эталоны, то кластеризация была построена правильно с самого начала. Если же какой-либо текст оказывался при тестировании в другом кластере, то он исключался из рассмотрения как не отвечающий тематике. В итоге были построены эталонные распределения специальностей ВАК, распределение расстояний между которыми показано на рис. 7.

Рис. 7. Распределение расстояний между эталонами биграмм для специальностей ВАК

Медиана этого распределения составляет 0.27. Заметим, что оптимальный уровень разделения своих и чужих текстов для биграмм по вычислительным экспериментам [Орлов, Осминин, 2017] составил 0.19. Для распределения, представленного на рис. 7, соответствующее значение квантиля равно 0.13. Это показывает, что эталоны разных специальностей в большинстве своем представляют совершенно различные тексты. Сравнительно близкими являются эталоны языкознания, культурологии, социологии, искусствоведения, политологии. Тем не менее все тексты, принадлежащие разным специальностям, безошибочно распознаются по близости к эталонам своих специальностей. Отметим, что эталоны естественно-научных специальностей отстоят довольно далеко как от гуманитарных эталонов, так и один от другого. Они образуют правую часть распределения.

Распределения расстояний между текстами определенной тематики и соответствующим эталоном не зависят от специальности. Эти распределения расстояний до «своих» и «чужих» эталонов показаны на рис. 8.

Благодаря значительно различающимся модам распределений расстояний «свой–чужой» тексты различных специальностей идентифицируются практически безошибочно. Отметим также, что расстояния между отдельными текстами даже внутри одной специальности имеют распределение, близкое к виду «чужой» на рис. 8, т. е. кластеризация проводится именно через близость к эталону, а не попарно.

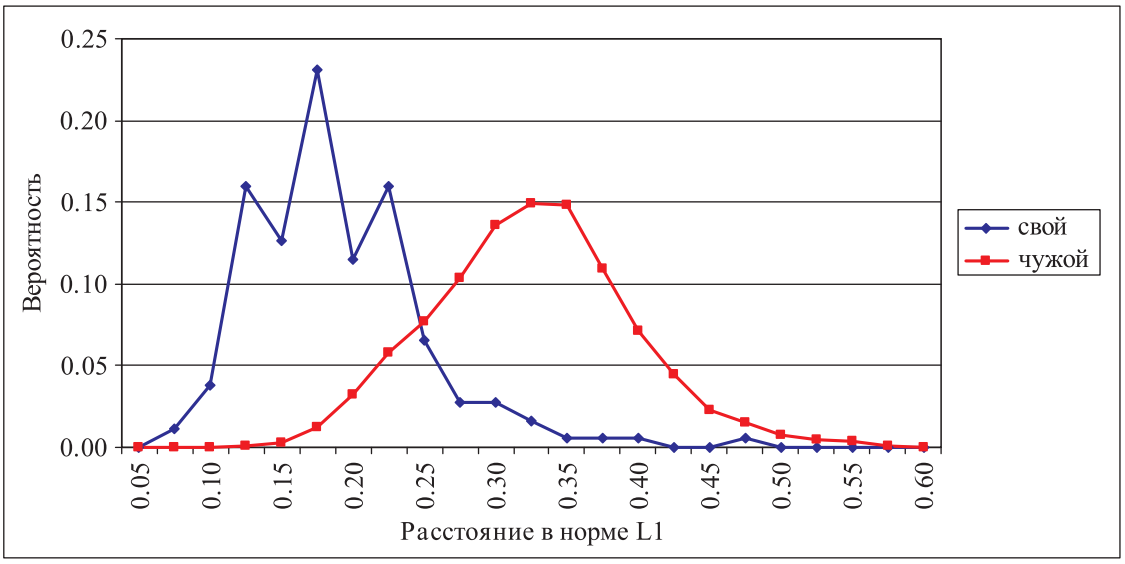

Рис. 8. Распределение расстояний от текста до эталонных биграмм для специальностей ВАК

5. Заключение

В работе построен статистический индикатор формального отнесения научного текста к определенной области знаний. Его значение в том, что, во-первых, он показывает, что тексты определенной научной тематики близки не только семантически, но и по структуре своего линейного пространства. Во-вторых, появилась возможность оценивать каждый конкретный текст на предмет соответствия специальности в виде численной характеристики, а не только по мнению эксперта.

Отметим здесь, что исключение из научных текстов формул и иных знаков не приводит к искажению косинуса угла. Формула в тексте является некоторым нечитаемым знаком, суть которого характеризуется словами, его сопровождающими. Анализ произвольно составленного текста из несвязанных смыслом слов, не имеющих специально подобранного звучания (рифмы и т. п.), всегда дает результат, близкий к «плохой литературе». Разработанный индикатор фактически основан на устойчивости элементов матрицы условных биграмм, т. е. на устойчивости пар буквосочетаний, и эта устойчивость определяется последовательностью используемых в данной отрасли слов. Если переставлять не слова, а предложения, то косинус анализируемого угла практически не меняется.

По-видимому, аналогичные исследования можно провести и для подразделов каждой отдельной специальности.

Список литературы (References)

Борисов Л. А., Ивченко А. Ю., Митин Н. А., Орлов Ю. Н. Тематическая классификация текстов с помощью спектральных портретов // Препринты ИПМ им. М. В. Келдыша. — 2017. — № 106. — 22 с. — URL: http://library.keldysh.ru/preprint.asp?id=2017-106 (дата обращения: 20.08.2019).

Borisov L. A., Ivchenko A. Yu., Mitin N. N., Orlov Yu. N. Tematicheskaya klassifikatsiya tekstov s pomoshch'yu spektral'nykh portretov [Classification of text information with the use of bigram analysis] // Preprinty IPM im. M. V. Keldysha [Keldysh Institute Preprints]. — 2017. — No. 106. — 22 p. — URL: http://library.keldysh.ru/preprint.asp?id=2017-106 (accessed 20.08.2019) (in Russian).

Годунов С. К. Современные аспекты линейной алгебры. — Новосибирск: Научная книга, 1997. — 388 с.

Godunov S. K. Sovremennye aspekty lineinoi algebry [Modern Aspects of Linerar Algebra]. — Novosibirsk, Nauchnaya kniga, 1997. — 388 p. (in Russian).

Голуб Дж., Ван Лоун Ч. Матричные вычисления. — М.: Мир, 1999. — 546 с.
Golub Gene H., Van Loan Charles F. Matrichnye vychisleniya [Matrix computations]. — Moscow: Mir, 1999. — 546 p. (in Russian).

Орлов Ю. Н., Осминин К. П. Свидетельство о регистрации программы для ЭВМ «Программный комплекс TRIL для идентификации языка, автора и жанра литературного текста». Правообладатель: ИПМ им. М. В. Келдыша РАН. Свидетельство о государственной регистрации № 2017611570 от 06.02.2017.
Orlov Yu. N., Osminin K. P. Svidetel'stvo o registratsii programmy dlya EVM «Programmnyi kompleks TRIL dlya identifikatsii yazyka, avtora i zhanra literaturnogo teksta». Pravoobladatel': IPM im. M.V. Keldysha RAN. Svidetel'stvo o gosudarstvennoi registratsii № 2017611570 ot 06.02.2017 [Certificate of registration of the computer program No. 2017611570 from 06.02.2017 “Software TRIL for identification of the language, author and genre of a literary text”] (in Russian).

Орлов Ю. Н., Осминин К. П. Методы статистического анализа литературных текстов. — М.: Эдиториал УРСС / Книжный дом «ЛИБРОКОМ», 2012. — 312 с.
Orlov Yu. N., Osminin K. P. Metody statisticheskogo analiza literaturnykh tekstov [Methods of Statistical Analysis of literary texts]. — Moscow: Editorial URSS / Knizhnyi dom “LIBROKOM”, 2012. — 312 p. (in Russian).

Bekhuis T., Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers // Artificial Intelligent Med. — 2012. — Vol. 55 (3). — P. 197–207.

Chandrasekaran B. Towards a taxonomy of problem solving types // AI Magazine. — 1983. — Vol. 4 (1).

Charles M. Adveribals of result: Phraseology and functions in the problem-solution pattern // J. of English for Academic Purposes. — 2011. — Vol. 10 (1).

Kim S. N., Martinez D., Cavedon L., Yencken L. Automatic classification of sentences to support evidence based medicine // BMC Bioinformatics. — 2011. — Vol. 12 (2). — S5.

Kirkland S. Subdominant eigenvalues for stochastic matrices with given column sums // Electronic Journal of Linear Algebra. — 2009. — Vol. 18. — P. 784–800.

Park D. H., Blake C. Identifying Comparative Claim Sentences in Full-Text Scientific Articles // Proc. of the 50-th Annual Meeting of the Association for Computational Linguistic, Jeju, Rep. of Korea. — 2012, 12 July. — P. 1–9.

Yuan Lu. The estimation for the Eigenvalues of stochastic matrices // Journal of Mathematics Research. — 2010. — Vol. 2 (3). — P. 177–181.