ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 106 за 2017 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Борисов Л.А., Ивченко А.Ю.,
Митин Н.А., Орлов Ю.Н.

Тематическая классификация текстов с помощью спектральных портретов

Рекомендуемая форма библиографической ссылки: Тематическая классификация текстов с помощью спектральных портретов / Л.А.Борисов [и др.] // Препринты ИПМ им. М.В.Келдыша. 2017. № 106. 22 с. doi:10.20948/prepr-2017-106
URL: http://library.keldysh.ru/preprint.asp?id=2017-106

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Л.А. Борисов, А.Ю. Ивченко,
Н.А. Митин, Ю.Н. Орлов

Тематическая классификация текстов с помощью спектральных портретов

Москва — 2017

Борисов Л.А., Ивченко А.Ю., Митин Н.А., Орлов Ю.Н.

Тематическая классификация текстов с помощью спектральных портретов

В работе рассмотрены примеры применения метода спектрального анализа несимметричных матриц для построения классификационных индикаторов при структурировании текстовой информации большого объема. Обнаружилась возможность классификации текстов по тематике на основе анализа структуры инвариантных подпространств стохастической матрицы условных вероятностей парных буквосочетаний. Выяснилось, что тексты весьма достоверно могут быть классифицированы как литературные, научные по психологии, философии, а также по естественным наукам. Индикатором разделения служит величина близости к нулю косинуса угла между левым и правым собственными векторами, отвечающими соответственно минимальному и максимальному действительным собственным значениям стохастической матрицы условных вероятностей пар буквосочетаний.

Ключевые слова: стохастическая матрица, спектральный портрет, классификация текстов

Borisov L.A., Ivchenko A.Yu., Mitin N.N., Orlov Yu.N.

Classification of text information with the use of bigram analysis

In this paper we consider examples of applying the method of spectral analysis of nonsymmetric matrices to construct indicators of classification in the structuring of textual information. The split indicator is the value of the cosine of the angle between the left and right eigenvectors, corresponding respectively to the minimum and maximum real eigenvalues of the stochastic matrix of conditional bigrams.

Keywords: stochastic matrix, spectral portrait, text classification

Работа выполнена при поддержке гранта РФФИ, проект № 15-01-07944

Содержание

Введение

В настоящей работе методы статистического анализа, развитые в ИПМ им. М.В. Келдыша РАН, применяются к задаче классификации текстовой информации достаточно большого объема (более 10 тыс. знаков). Представляет интерес нахождение статистических индикаторов, позволяющих с определенным уровнем достоверности классифицировать информацию по тем или иным атрибутам. В частности, будут исследованы тексты, представляющие собой профессиональные научные работы в областях естественных и гуманитарных наук.

В монографии [1] и работе [2] приведены примеры индикаторов для классификации текстов художественной литературы по авторам и жанрам. Индикаторами в этих работах выступали расстояния между распределениями анализируемых текстов по буквам и некоторым эталонным распределением, специфическим для определенного автора или жанра. Распределение расстояний между распределениями в нормах C и L1 позволили в этих работах определить критический уровень наилучшего разделения при кластеризации текстов по авторам и жанрам.

В работе [3] был проведен анализ точности вычислительных процедур, которые сопровождают процесс кластеризации функций распределения. Выяснилось, что непосредственное проецирование вектора большой размерности в пространство малой размерности является некорректной операцией, если базис не ортогональный, как оно и имеет место быть для авторских эталонных распределений вероятностей буквосочетаний. Ошибки проецирования связаны с неточностью определения координат проецируемого вектора, а также зависят от числа обусловленности матрицы проектора. Вектором же в рассматриваемых задачах являлась упорядоченная совокупность частот буквосочетаний в тексте. Будучи эмпирически оцененной, вероятность определенного буквосочетания известна не как элемент генеральной совокупности, характерной для данного автора, а всего лишь как выборочная частота. Поэтому по разным выборкам эта частота будет, вообще говоря, различной. Точность, с которой необходимо определять всю совокупность частот буквосочетаний для достижения требуемой точности в проецировании эмпирического вектора на пространство эталонных распределений, находится с помощью специальной вычислительной процедуры, называемой построением спектрального портрета матрицы. Эта процедура и является основным инструментом анализа, применяемого в настоящей работе для нахождения некоторых атрибутов текста. Параллельно с построением спектрального портрета с определенной вычисляемой точностью находятся собственные значения и соответствующие им собственные векторы матрицы условных вероятностей буквосочетаний.

В книге [1] было показано, что существуют два собственных вектора указанной матрицы, которые являются устойчивыми характеристиками автора

текста, причем каждый из них может выступать авторским эталоном. Выяснилось также, что угол между этими векторами имеет универсальную устойчивость и, независимо от авторов, близок к прямому, отличаясь от него на 4-5 градусов в большую сторону.

В настоящей работе мы проанализировали не литературный, а профессиональный текст. Для физико-математических научных работ оказалось, что этот угол существенно меньший и находится в пределах 1-1,5 градусов. Хотя различие между литературными и техническими текстами в этом смысле сравнительно невелико, следует подчеркнуть, что эмпирические плотности распределения косинуса угла между двумя выделенными собственными векторами для двух указанных классов текстов не имеют общего носителя. Это позволяет разделить их с нулевой (по данному корпусу текстов, разумеется) ошибкой. Объем же текстов был достаточно большим – по 100 текстов каждого типа разных авторов.

Также было выяснено, что тексты, относящиеся к гуманитарным наукам, занимают некоторое промежуточное положение по указанному индикатору между литературными текстами и естественными науками. Это позволило сделать предположение о том, что именно измеряет индикатор косинуса угла: по-видимому, он показывает степень однородности текста, которая связана со статистической устойчивостью частот буквосочетаний.

Подробному описанию методики проведения соответствующего статистического эксперимента и посвящена настоящая работа.

1. Матрица условных вероятностей биграмм

Опишем сначала основной объект нашего исследования. Рассмотрим некоторый текст на русском языке. Исключим из него все пробелы, знаки препинания, цифры и прочие символы, оставив только буквы алфавита, причем строчные и прописные буквы не различаем. Получившуюся совокупность символов будем рассматривать 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. Так, например, первое значение временного ряда – это первая буква в тексте, второе значение – вторая буква, и т.д. Тем самым каждому тексту сопоставляется временной ряд $\{x(t)\},\ x(t) \in \{а,..,я\}$. Случайный процесс, реализацией которого является ряд $\{x(t)\}$, обозначим $\xi(t)$.

Итак, $\mathbf{f}(t)$ есть вектор вероятностей, i-я компонента которого равна вероятности того, что в момент времени t реализуется буква i:

$$\mathbf{f}(t) = \begin{pmatrix} P(\xi(t) = а) \\ .. \\ P(\xi(t) = я) \end{pmatrix} \in R^{33}. \qquad (1)$$

Для текста в целом вектор (1) представляет набор эмпирических частот употребления указанных символов. Также мы можем построить распределение частот букв и для фрагмента текста определенной длины. Двигаясь скользящим окном по тексту, в каждый следующий момент времени мы получаем некоторую другую, вообще говоря, функцию распределения фрагмента текста по буквам. Для каждого натурального сдвига по тексту на l шагов построим матрицу условных вероятностей, элементы которой $P_{ij}(l,t)$ определяются формулой

$$P_{ij}(l,t) = P\big(\xi(t+l) = j \mid \xi(t) = i\big), \quad i, j \in \{a,..,я\}. \tag{2}$$

Из формулы (2) следует, что

$$\sum_j P_{ij}(l,t) = 1 \ \ \forall i, l. \tag{3}$$

Сопоставим каждой такой матрице условных вероятностей того, что через l символов после буквы i, реализовавшейся в тексте в момент t, появится буква j, оператор трансляций $\hat{P}_t(l)$ в пространстве исследуемого текста. Эти операторы переводят вектор $\mathbf{f}(t)$ в вектор $\mathbf{f}(t+l)$, так что

$$\mathbf{f}(t+l) = \hat{P}_t(l)\mathbf{f}(t). \tag{4}$$

Таким образом, $P_{ij}(l,t)$ суть элементы матриц линейных операторов, действующих на пространстве вероятностей в R^{33}. Их можно трактовать как операторы эволюции вероятностного распределения букв в рассматриваемом тексте. Будем говорить об однобуквенном распределении текста как об 1-ПФР (однобуквенная плотность функции распределения), а о двухбуквенном распределении – как о 2-ПФР. Если $l = 1$ и распределение $\mathbf{f}(t)$ стационарно, то матрица $P_{ij}(1,t) \equiv P_{ij}$ выражается через 2-ПФР $F(i, j)$ и 1-ПФР $f(i)$, которые отвечают произведению в целом:

$$P_{ij} = \frac{F(i, j)}{f(i)}. \tag{5}$$

В силу того, что

$$f(i) = \sum_j F(i, j), \tag{6}$$

причем результат суммирования инвариантен относительно перестановки аргументов 2-ПФР, хотя сама она не является симметричной, получаем, что матрица $P_{ij}(1)$ имеет одно из собственных значений, равное 1, и этому значению отвечает собственный вектор, являющийся 1-ПФР текста.

Таким образом, вероятностному распределению пар букв в литературном произведении можно сопоставить линейный оператор, что позволит исследовать тексты на близость между собой с помощью операторных норм.

Мы будем рассматривать далее собственные векторы матрицы (5), отвечающие действительным собственным значениям.

2. Построение спектрального портрета матрицы

При анализе спектра стохастической матрицы одним из основных вопросов является оценка того, в каких пределах лежит возмущение спектра матрицы при малом возмущении ее элементов. Приведем здесь определения ключевых понятий, следуя монографии С.К. Годунова [4].

По определению, 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.

Сингулярными числами σ_i матрицы P называются собственные значения эрмитовой матрицы P^+P, где знак «плюс» вверху означает эрмитово сопряжение. Все сингулярные числа неотрицательны. Занумеруем их в порядке возрастания, так что для матрицы порядка N минимальное сингулярное число есть σ_1, а максимальное σ_N.

Числом обусловленности невырожденной матрицы P называется величина

$$\mu(P) = \frac{\sigma_N(P)}{\sigma_1(P)} = \|P\| \cdot \left\|P^{-1}\right\|. \qquad (7)$$

Резольвентой матрицы P называется матрица

$$R(\lambda) = (\lambda I - P)^{-1}. \qquad (8)$$

В терминах резольвенты ε-спектр определяется следующим образом: число λ принадлежит ε-спектру $\Lambda_\varepsilon(P)$ матрицы P, если

$$\|R(\lambda)\| \ge \frac{1}{\varepsilon\|P\|}. \qquad (9)$$

При исследовании расположения точек спектра удобно рассматривать замкнутые гладкие кривые γ_ε, представляющие изолинии ε-спектра. Контур γ_ε разбивает весь ε-спектр $\Lambda_\varepsilon(P)$ на две части – лежащие внутри и вне его. Тем самым γ_ε осуществляет дихотомию ε-спектра матрицы. Качество дихотомии $\kappa_\gamma(P)$ оценивается нормой квадрата резольвенты (9) на данной кривой:

$$\kappa_\gamma(P) = \frac{\|P\|^2}{l_\gamma} \oint_\gamma \|R(\lambda)\|^2 d\lambda. \qquad (10)$$

Здесь l_γ есть длина контура γ. Величина $\kappa_\gamma(P)$ выбрана как индикатор точности разделения спектра потому, что если на некоторой кривой γ нет точек спектра $\lambda(P)$, то норма резольвенты на такой кривой конечна, $\|R(\lambda)\|_\gamma < \infty$, как и интеграл от нее по этой кривой.

Если внутри области, ограниченной кривой γ_ε, оказалось несколько собственных значений, то с указанной точностью ε их следует считать совпадающими. Тогда подпространство с базисом из собственного и присоединенных векторов для такого кратного собственного значения будет

приближенным ε-инвариантным подпространством для оператора P. Проектор Π на это инвариантное подпространство определяется формулой:

$$\Pi_\gamma = \frac{1}{2\pi i}\oint_\gamma R(\lambda)d\lambda. \qquad (11)$$

В частности, если контур γ охватывает все собственные значения, то всякая аналитическая функция от матрицы, будучи определенной на ее спектре, имеет интегральное представление

$$f(P) = \frac{1}{2\pi i}\oint_\gamma f(\lambda)R(\lambda)d\lambda. \qquad (12)$$

Рассмотрим радиальную дихотомию, т.е. дихотомию, задаваемую кривой $\lambda = re^{i\varphi}$ при фиксированном значении r. Тогда параметр дихотомии $\kappa_r(P)$ является нормой эрмитовой матрицы $H_r(P)$, имеющей следующее интегральное представление:

$$H_r(P) = \frac{1}{2\pi}\int_0^{2\pi}\left(P^+ - re^{-i\varphi}I\right)^{-1}\left(P - re^{i\varphi}I\right)^{-1}d\varphi, \quad \kappa_r(P) = \|P\|^2\cdot\|H_r(P)\|. \qquad (13)$$

Интеграл в (13) сходится только в том случае, если на окружности $\lambda = re^{i\varphi}$ нет собственных значений матрицы P. Это представление используется при численном нахождении ε-спектра матрицы.

На рис. 1 в качестве типичного примера приведен спектральный портрет оператора P, отвечающего книге [1] одного из авторов данной работы. На этом рисунке точками отмечены собственные значения, рассчитанные по стандартным процедурам библиотеки LAPACK [5, 6], а линии уровня соответствуют областям, внутри которых находятся собственные значения при определенной ошибке в элементах матрицы. Например, линии желтого контура ограничивают область, внутри которой оказывается собственное число, если относительная ошибка в элементах матрицы равна $\varepsilon = 10^{-2}$. Границы цветовых областей являются линиями уровня для L_2-нормы резольвенты $\left\|(\lambda I - P)^{-1}\right\|$.

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 $\lambda_2 \approx -0{,}6$, и некоторый «коллективный» ноль радиуса примерно 0,3. Более детальное представление спектра требует точности входных данных на уровне 10^{-4}, для чего, согласно [1], требуется длина текста порядка 10 млн знаков. Текстов такой длины мы в данной работе не рассматриваем, поэтому соответствующая детализация точек спектра не является достоверной.

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.

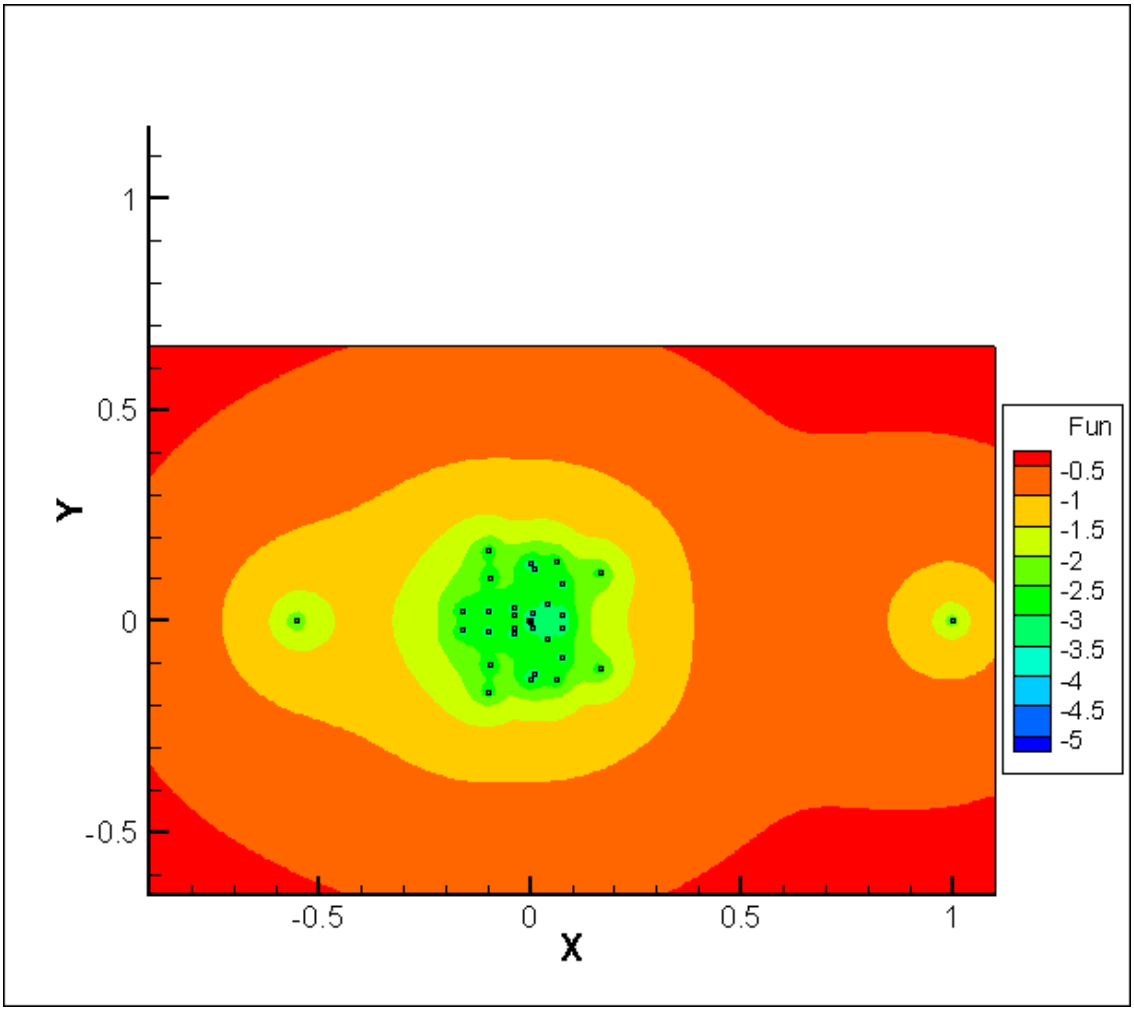

Рис. 1. Спектральный портрет текста монографии [1]

Наибольшему по модулю отрицательному собственному значению отвечают левый **u** правый **s** собственные векторы, образующие углы с вектором 1-ПФР **f**, косинусы которых изменяются по большинству текстов соответственно от -0,07 до -0,09 и от 0,29 до 0,31. Эти векторы **u** и **s**, как и вектор **f** 1-ПФР, обладают авторской устойчивостью. Более того, левый собственный вектор **u** имеет разрешающую способность по идентификации автора, более высокую, чем 1-ПФР: в норме L_1 ошибка определения автора текста по эталону, привязанному к вектору **u**, составила 0,12, тогда как для эталона **f** ошибка составила 0,15. Эти результаты были получены в [1] на корпусе из 3000 текстов, так что отличие в точности следует считать существенным. Отметим, что левый и правый собственные векторы для собственного значения $\lambda_1 = 1$ совпадают.

Таким образом, левый собственный вектор, принадлежащий наименьшему действительному собственному значению матрицы трансляций, является в некотором смысле «антиподом» вектора 1-ПФР и, хотя он не имеет такого же ясного статистического смысла, обладает как высокой авторской устойчивостью, так и высокой разрешающей способностью. Поскольку векторы **f** и **u** приближенно ортогональны, их можно считать главными направлениями матрицы условных вероятностей биграмм:

$$(u, Pf) \approx 0. \qquad (14)$$

Типовые координаты этих трех векторов приведены на рис. 2. Все векторы на этом рисунке нормированы на единицу в смысле суммы квадратов.

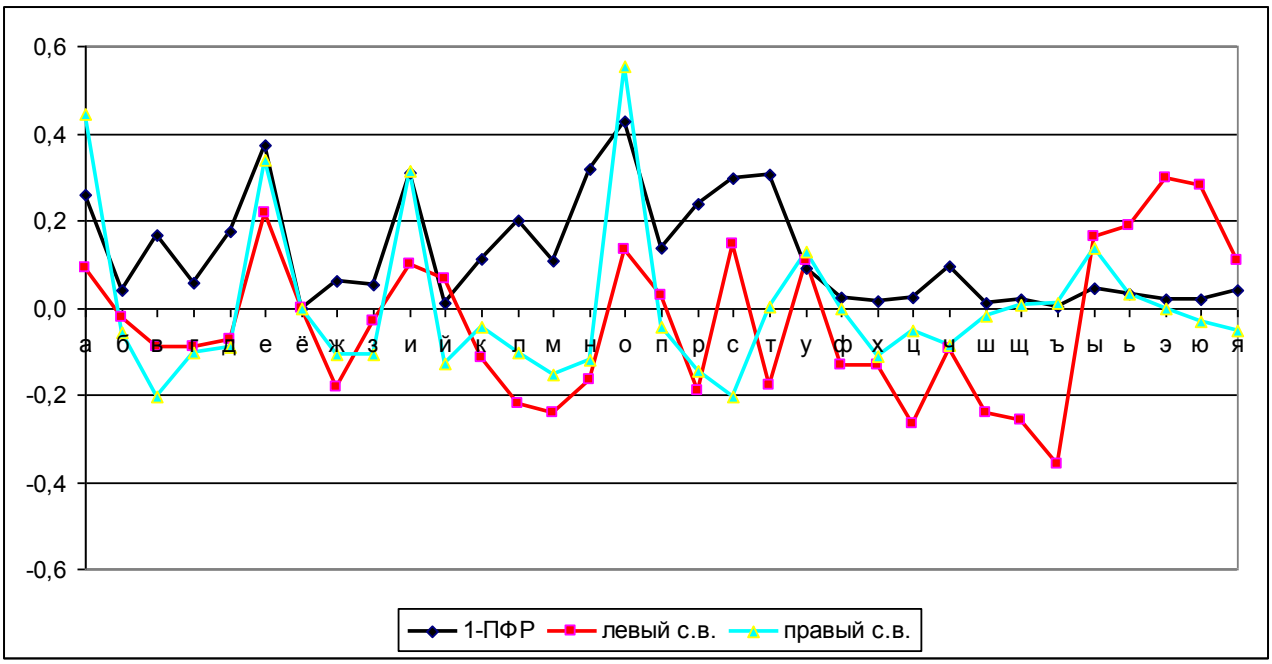

Рис. 2. Вектор 1-ПФР $\mathbf{f}$ и левый и правый векторы $\mathbf{u}, \mathbf{s}$ для книги [1]

Применительно к книге [1] оказалось, что $\cos(\mathbf{u}, \mathbf{f}) = -0{,}03$. Это значение в три раза меньше характерной величины 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 косинус заключен в пределах от -0,045 до -0,025.

Найденное различие между литературными и техническими текстами предположительно обусловлено следующим обстоятельством. Поскольку рассматриваемые собственные векторы приближенно могут считаться главными направлениями матрицы трансляций, то в идеале они ортогональны. Но точность, с которой они вычисляются, 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. В результате частая повторяемость определенных буквосочетаний обеспечивает высокую точность элементов матрицы даже на относительно небольших текстах, так что векторы оказываются определенными гораздо лучше, чем для художественной литературы.

Примеры спектральных портретов стохастических матриц (5) для авторских эталонов в смысле 1-ПФР текстов приведены в работах [1, 3, 7]. Ниже в разделах 3—5 мы приведем спектральные портреты и пары векторов

$\mathbf{u},\mathbf{f}$ для отдельных текстов известных писателей, а также для научных книг по философии, психологии и математической физике, чтобы продемонстрировать эффект, о котором идет речь в нашей работе.

3. Примеры спектральных портретов литературных текстов

Как показано в [1], статистически достоверной индивидуальностью обладают спектры операторов трансляций, отвечающих эталонному авторскому распределению 2-ПФР, поскольку в составлении эталона участвуют произведения совокупным объемом более 5 млн знаков. Однако с точностью на уровне $\varepsilon = 10^{-2}$ три области расположения собственных значений определяются вполне устойчиво и для отдельных текстов. В этой связи интересно сравнить между собой спектральные портреты отдельных литературных произведений различных писателей. На рис. 3—9 приведены спектральные портреты произведений Булгакова, Гончарова, Достоевского, Пушкина, Толстого, Тургенева, Чехова. Подчеркнем, что сами эти портреты используются только как индикаторы точности вычисляемых параметров, хотя распределение цветов достаточно специфично для каждого автора. Вычисляемые же собственные векторы могут идентифицировать более тонкие авторские характеристики.

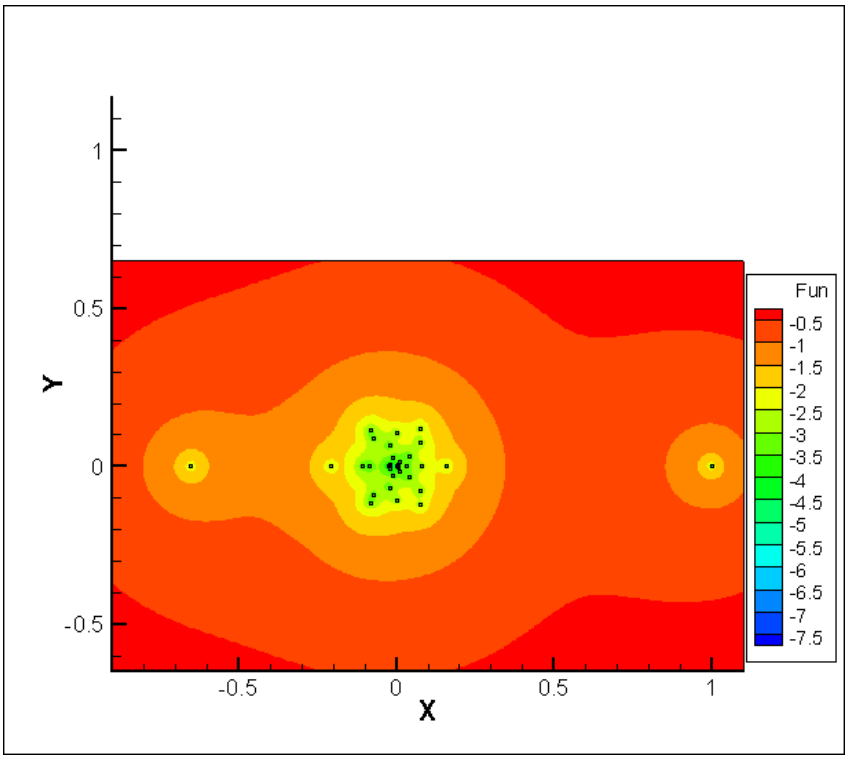

Рис. 3. Спектральный портрет Булгакова (Мастер и Маргарита), $\cos(\mathbf{u},\mathbf{f}) = -0{,}07$

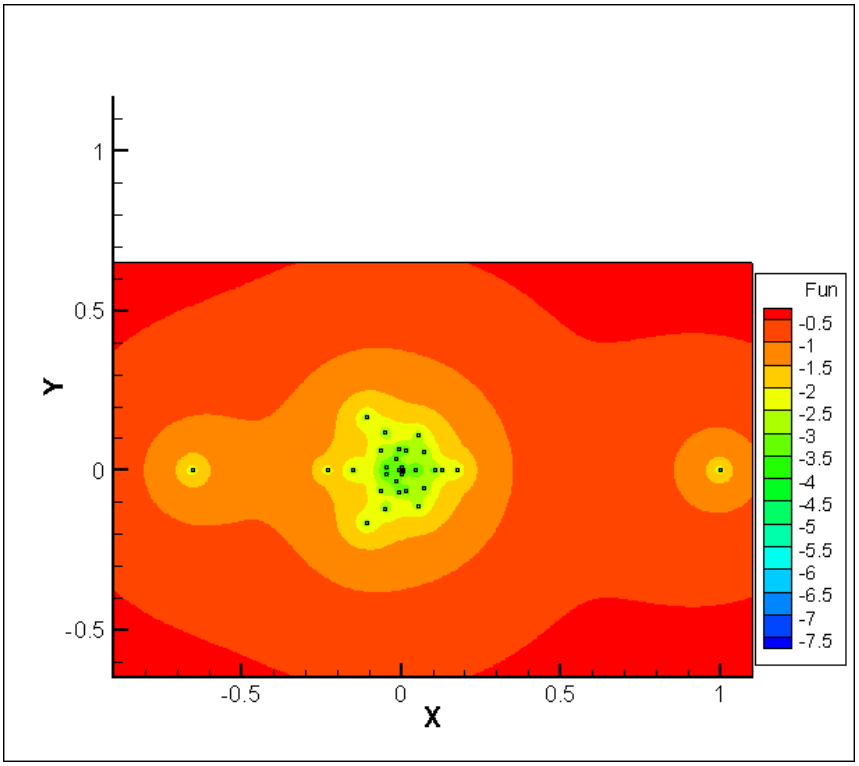

Рис. 4. Спектральный портрет Гончарова (Обломов), $\cos(\mathbf{u},\mathbf{f}) = -0{,}07$

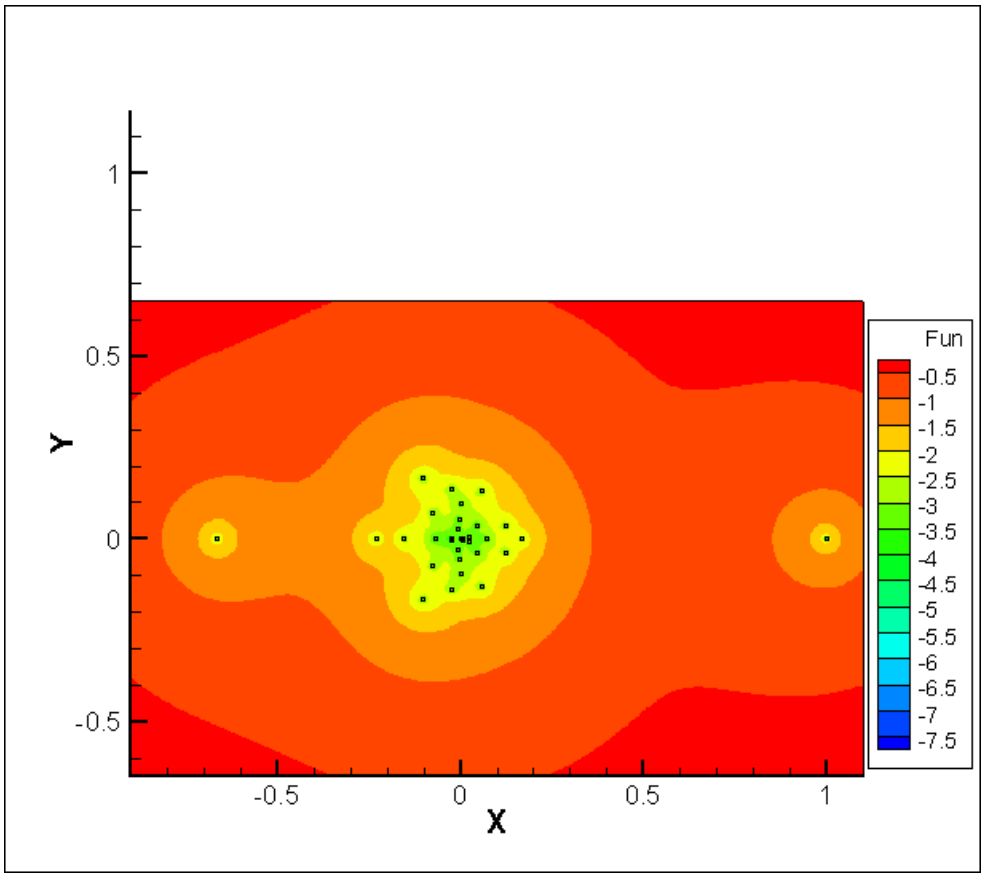

Рис. 5. Спектральный портрет Достоевского (Записки из мертвого дома), $\cos(\mathbf{u},\mathbf{f}) = -0{,}07$

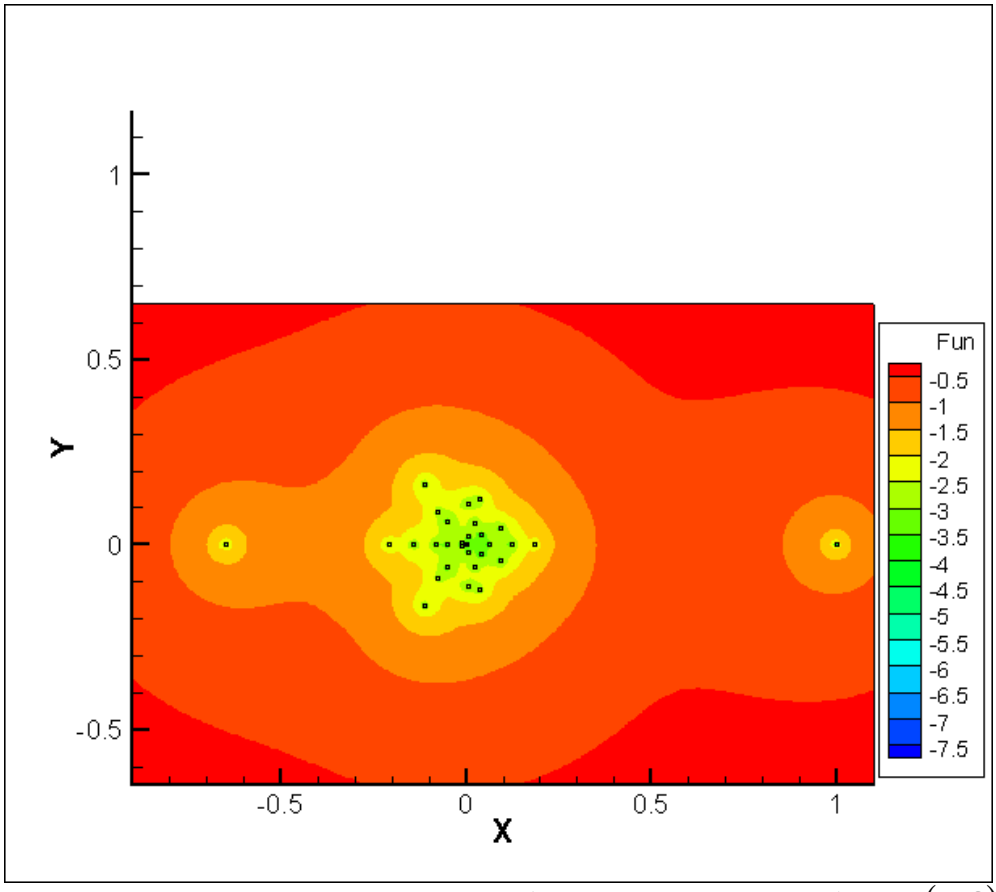

Рис. 6. Спектральный портрет Пушкина (Капитанская дочка), $\cos(\mathbf{u},\mathbf{f}) = -0{,}08$

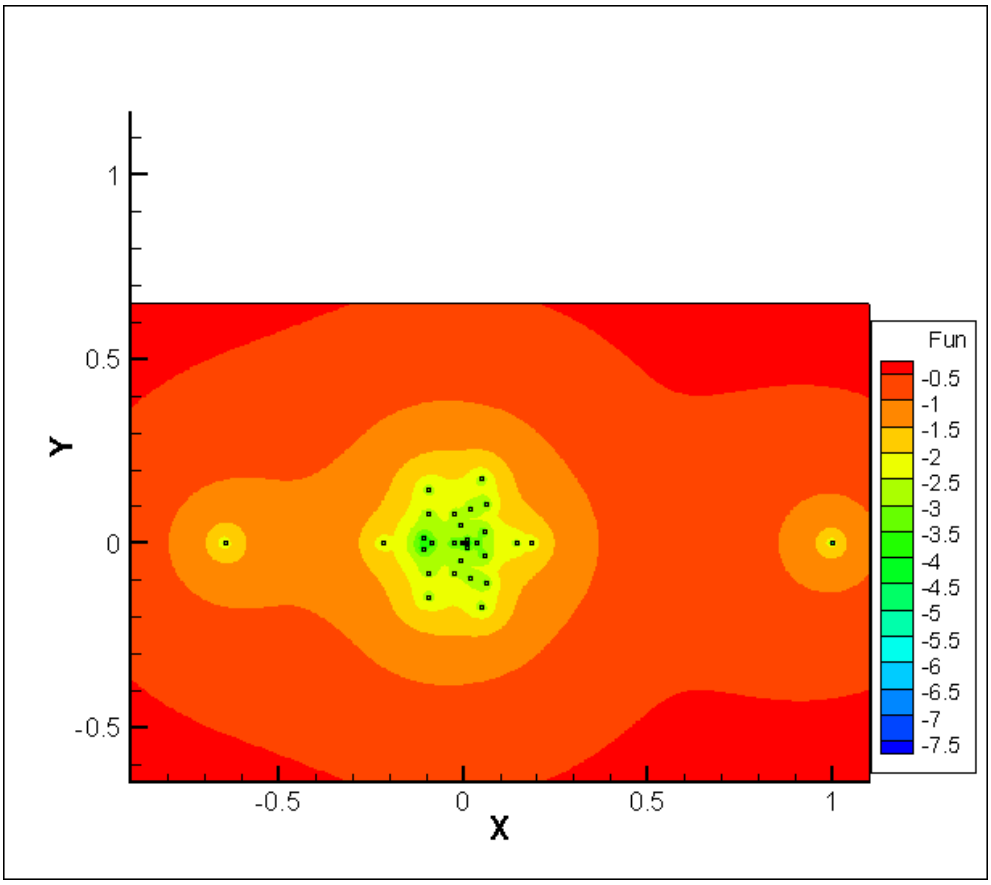

Рис. 7. Спектральный портрет Толстого (Воскресение)
$\cos(\mathbf{u},\mathbf{f}) = -0{,}09$

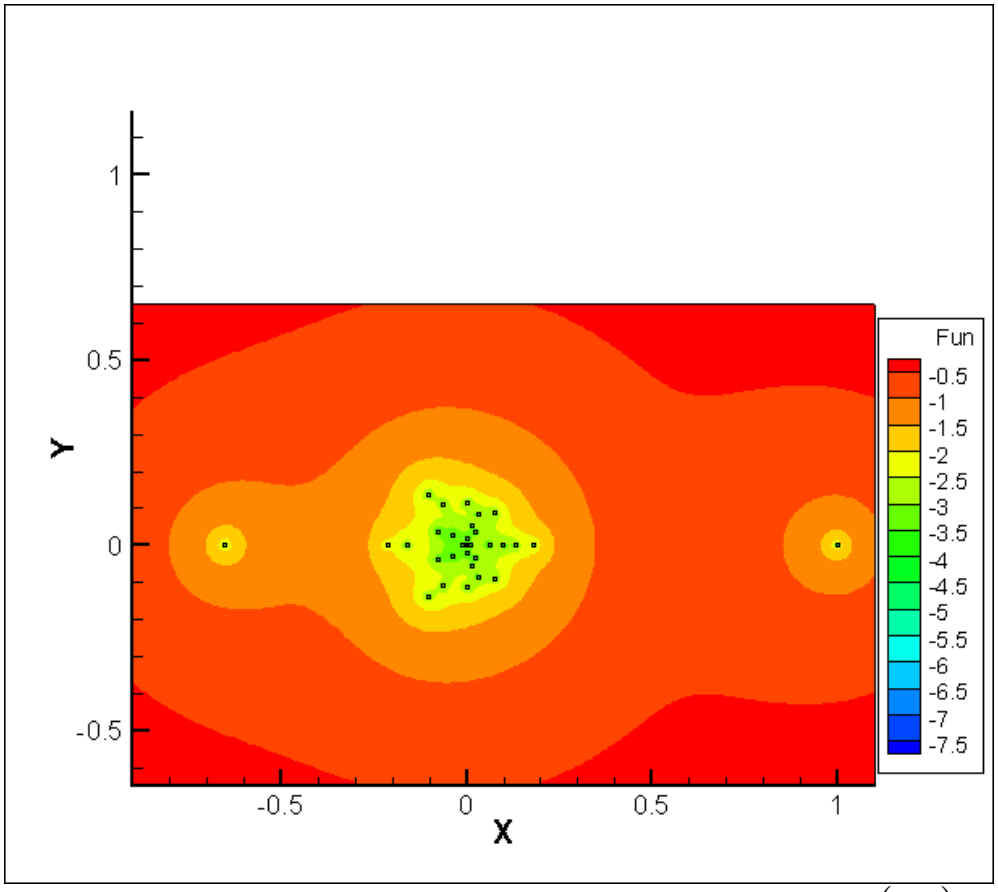

Рис. 8. Спектральный портрет Тургенева (Отцы и дети), $\cos(\mathbf{u},\mathbf{f}) = -0{,}09$

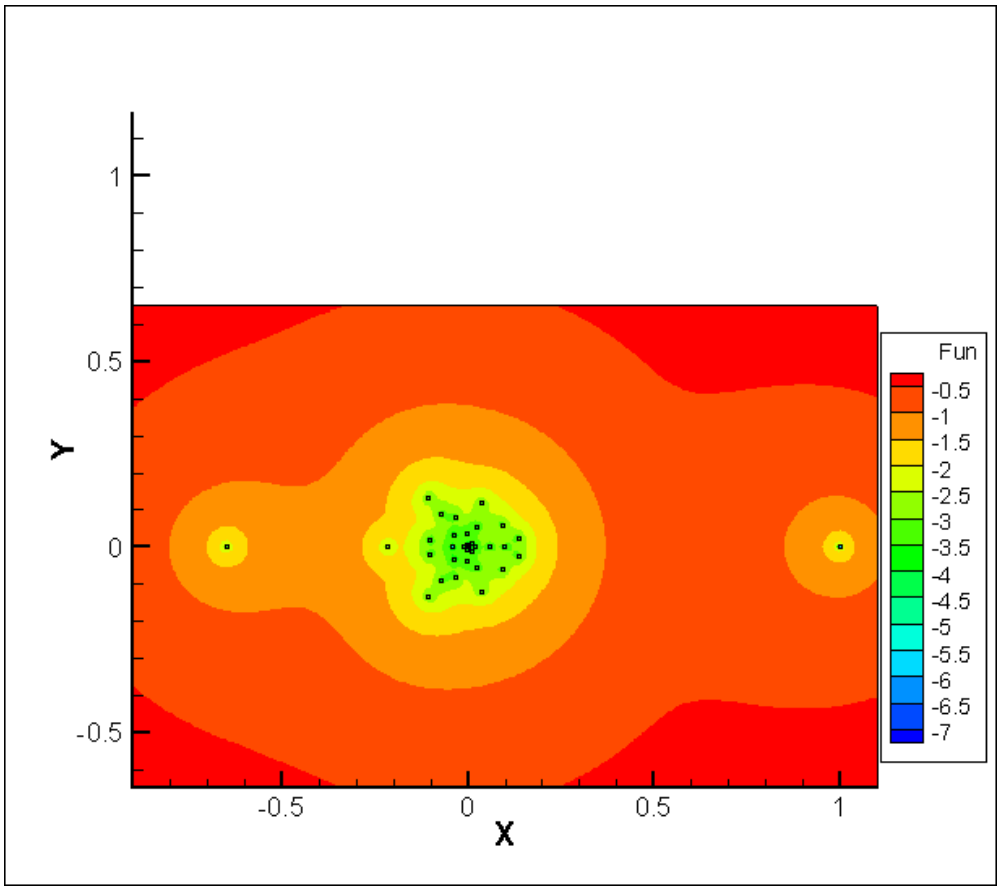

Рис. 9. Спектральный портрет Чехова (Дом с мезонином)
$\cos(\mathbf{u},\mathbf{f}) = -0{,}08$

В целом «авторские» спектральные портреты сохраняют свою индивидуальность, характеризующую писателя: их форма приблизительно совпадает для разных произведений одного и того же автора.

Что касается скалярного произведения $(\mathbf{u},\mathbf{f})$, оно для большинства писателей лежит в указанных ранее границах от -0,07 до -0,09. Насколько интересной является ситуация практически полной ортогональности векторов главных направлений, становится понятно из теста, результаты которого представлены далее в разделах 4—6.

4. Анализ физико-математических текстов

Тексты научно-технического содержания отличаются от литературных в плане разнообразия используемых слов и конструкций, естественно, в меньшую сторону. Кроме того, часть предложений оканчивается формулами, которые при анализе не учитываются, так что некоторые фразы не имеют логического завершения. Выяснилось, что последнее обстоятельство не существенно для анализа, поскольку физико-математические тексты, состоящие только из введения и описания идеи или эксперимента, имеют точно такие же статистические свойства, как и фрагменты с доказательством теорем.

Ниже на рис. 10—16 показаны спектральные портреты текстов такого рода (учебников, научных статей и диссертаций). Их основное отличие от пиктограмм, изображенных на рис. 3—9 в том, что для них позиционирование собственных значений матрицы гораздо более точное.

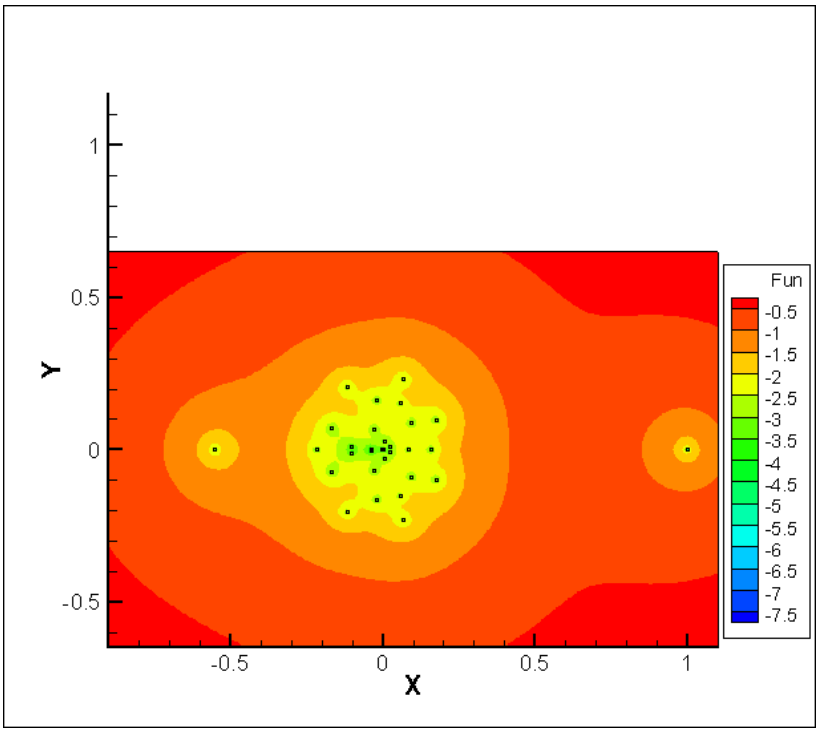

Рис. 10. Григорий Амосов (Квантовые динамические системы),
$\cos(\mathbf{u},\mathbf{f}) = -0{,}034$

Например, у Чехова на рис. 9 при точности 10^{-3} (зеленая зона) известно всего 4 собственных значения из 33, а у Амосова (рис. 10) при той же точности их 24. В результате оказывается, что косинус угла между сопряженными векторами главных направлений оказывается для физико-математических текстов в 2,5-3 раза меньше по абсолютной величине, чем для литературных.

Характерно, что для физико-математических текстов спектральный портрет не является персональным идентификатором автора, а зависит как от автора, так и от темы научной работы. Например, статистический анализ литературных текстов (работа [1], рис. 1) имеет заметно отличающийся спектр по сравнению с квантовыми кинетическими уравнениями (автор тот же, рис. 14). Поэтому классификационные свойства собственно спектральных портретов ограничены по сравнению с индикатором косинуса угла, который имеет достаточно узкое распределение в соответствии с тематикой научных исследований.

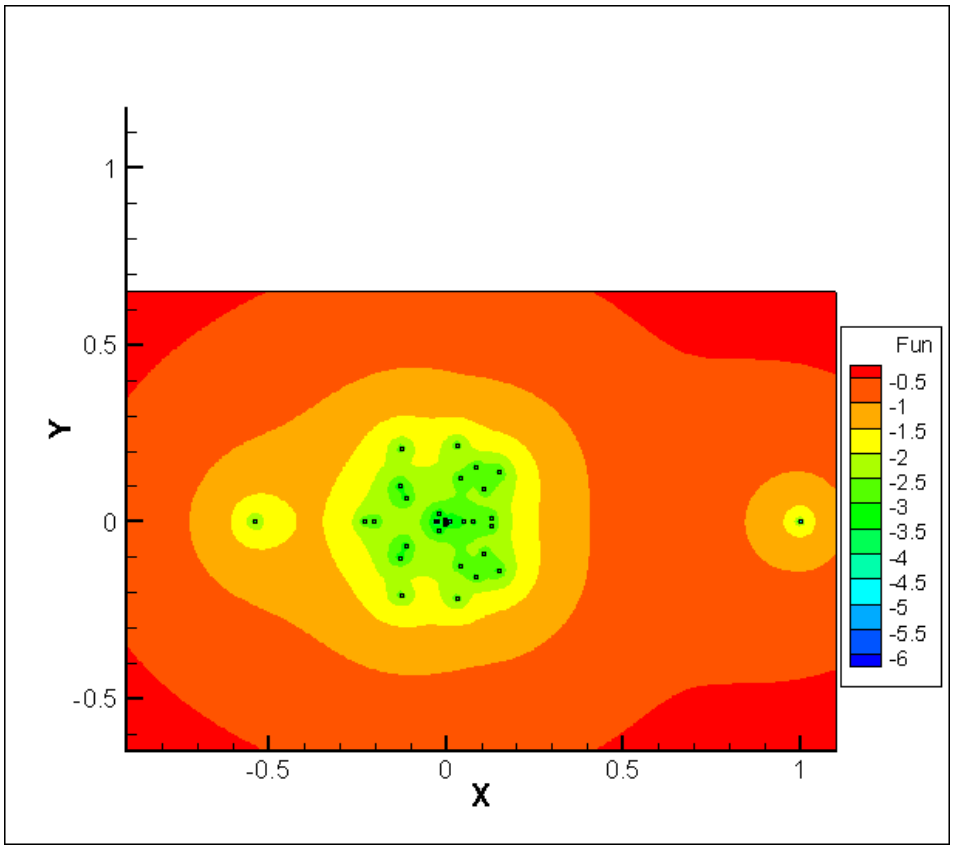

Рис. 11. Татьяна Гермогенова (Интегральное уравнение переноса), $\cos(\mathbf{u},\mathbf{f}) = -0{,}026$

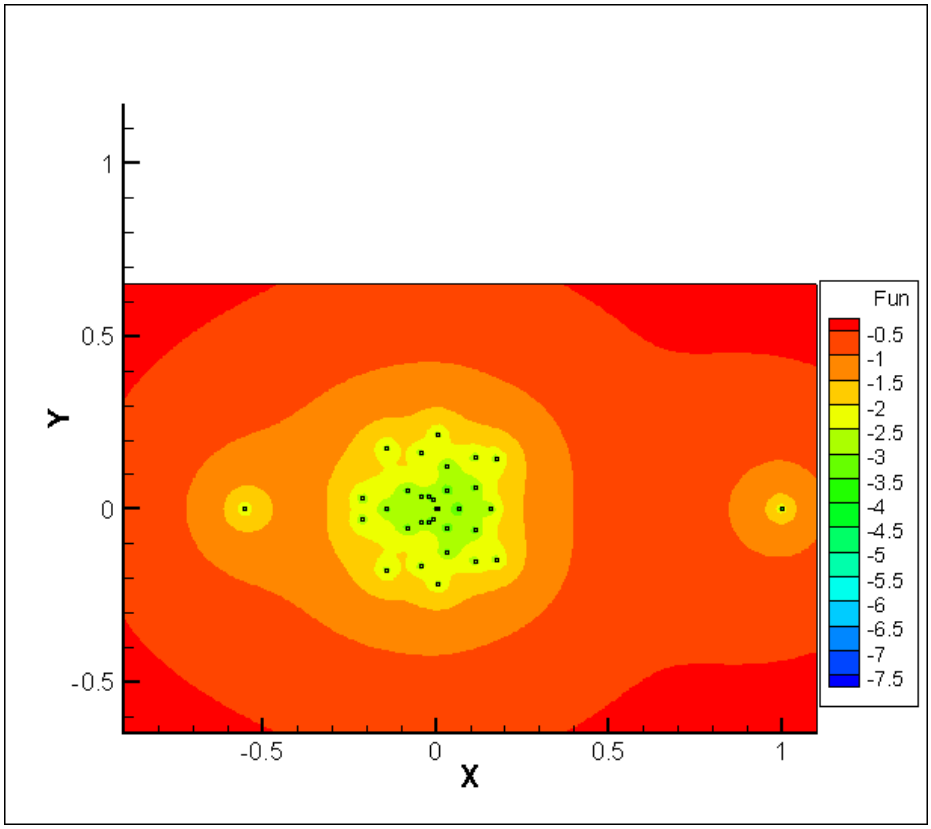

Рис. 12. Наталия Любимова (Метод БГК для квантовых ферми-газов),
$\cos(\mathbf{u},\mathbf{f}) = -0{,}040$

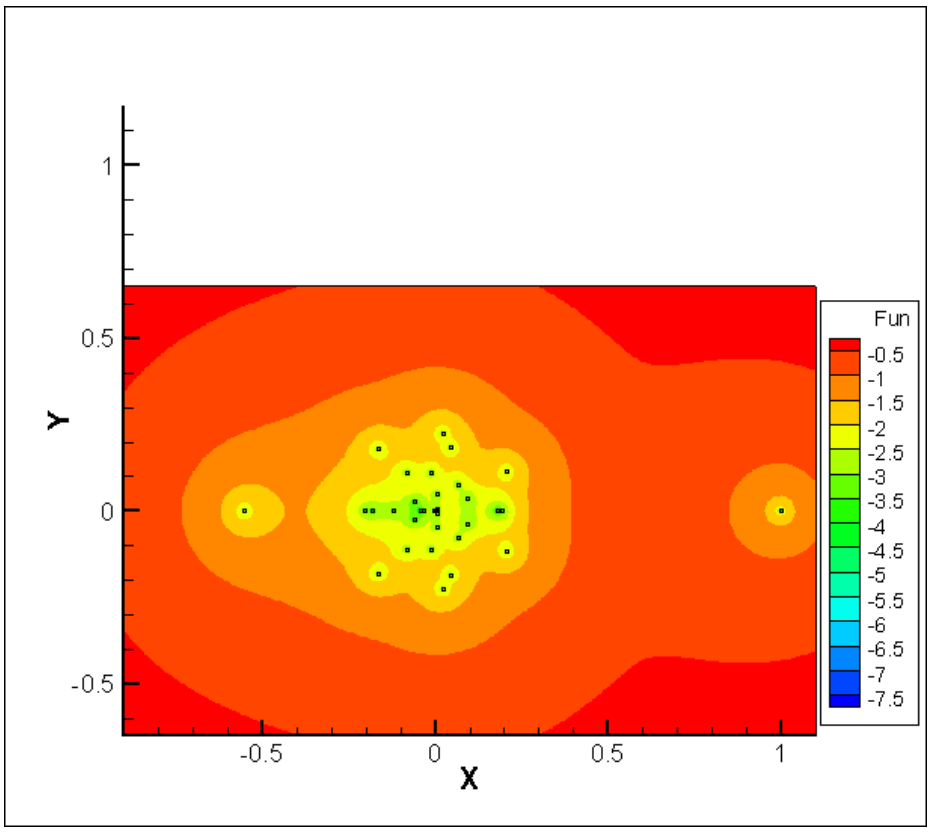

Рис. 13. Джеймс Максвелл (Труды по кинетической теории, перевод),
$\cos(\mathbf{u},\mathbf{f}) = -0{,}035$

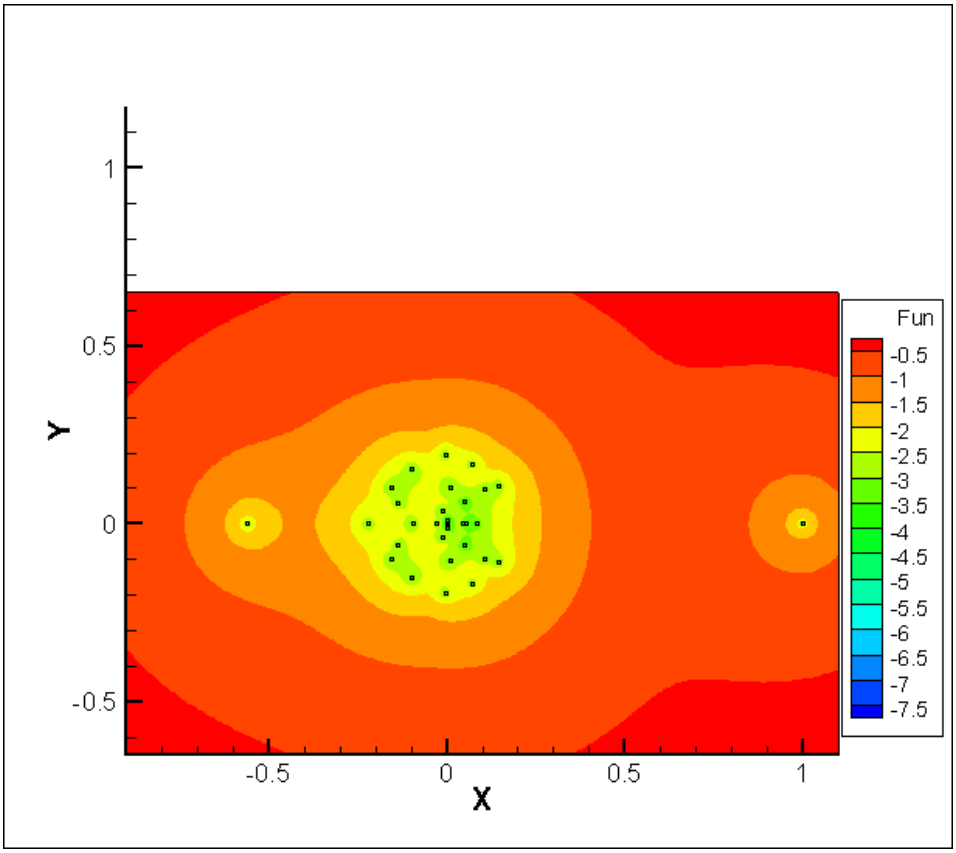

Рис. 14. Юрий Орлов (Квантовые кинетические уравнения),
$\cos(\mathbf{u},\mathbf{f}) = -0{,}035$

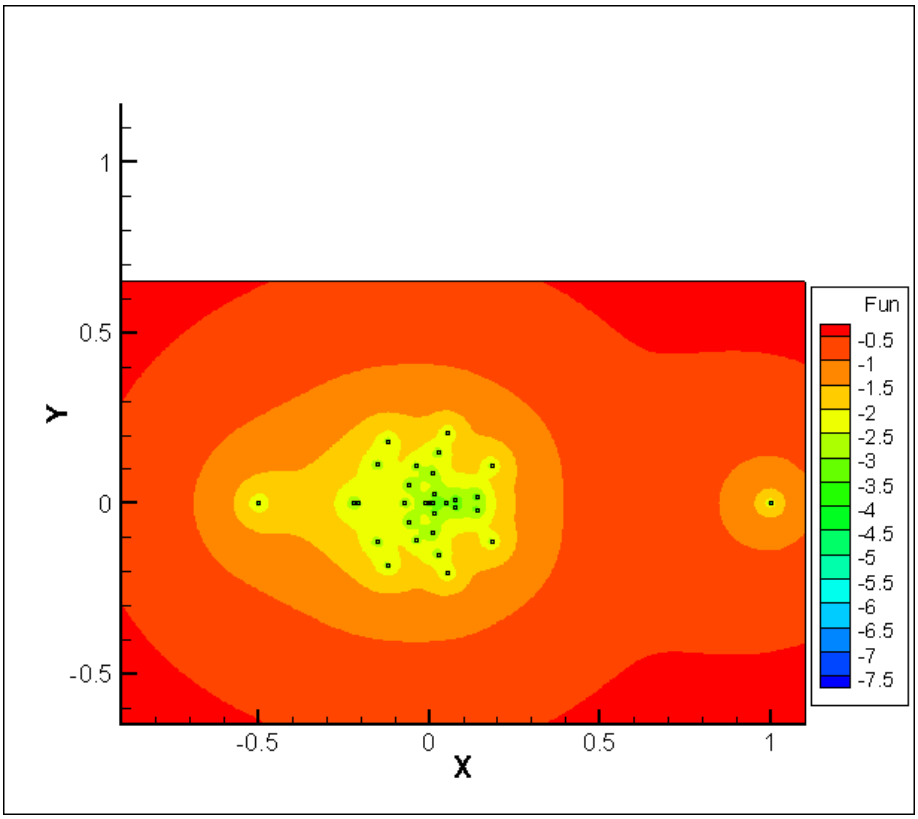

Рис. 15. Константин Осминин (Нестационарные временные ряды),
$\cos(\mathbf{u},\mathbf{f}) = -0{,}033$

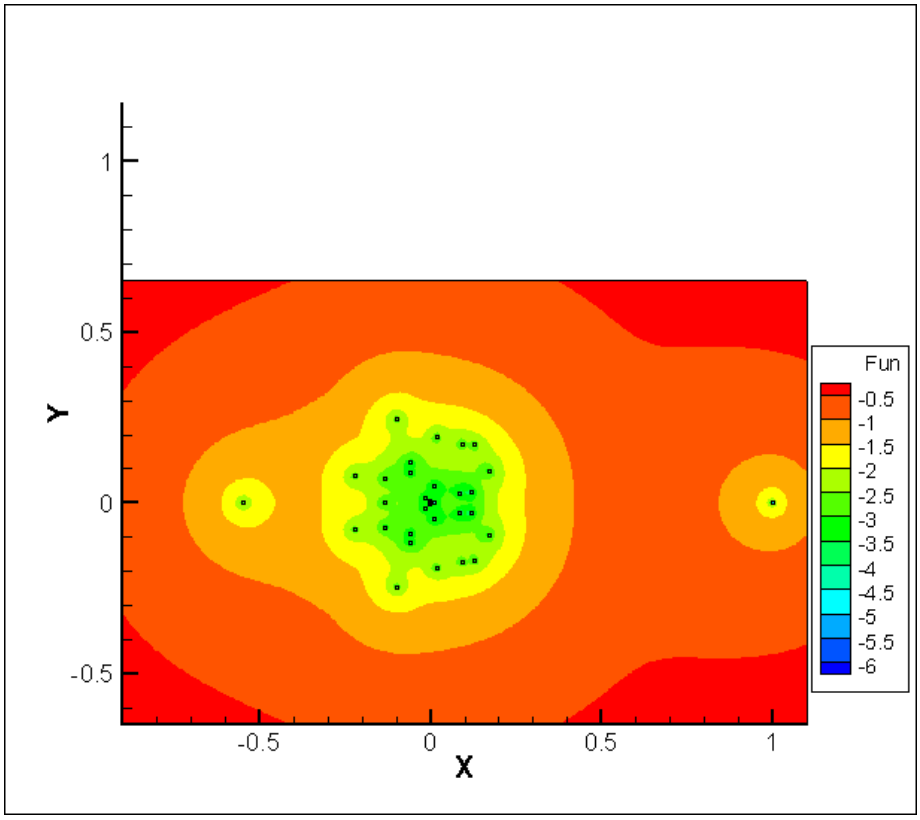

Рис. 16. Всеволод Сакбаев (Динамика квантовых систем с вырожденным гамильтонианом), $\cos(\mathbf{u},\mathbf{f}) = -0{,}025$

Аналогичный статистический эксперимент был поставлен для профессиональной литературы по химии. Были рассмотрены, в частности, следующие произведения, длина которых составляла не менее 20 страниц: «Аффинная хроматография», «Неорганические иониты», «Обогащение урана», «Органические реакции», «Химия кремнезема», «Химия плутония» и др., всего 10 произведений. Для полноценной статистики этого, конечно, не достаточно, но нашей целью было показать, что большая часть произвольно взятых текстов определенной тематики вполне может быть сгруппирована по указанному параметру. Выяснилось, что косинус угла между рассматриваемыми векторами для химических произведений в среднем несколько дальше отстоит от прямого, чем для физико-математических.

5. Эталонные распределения текстов гуманитарных направлений

Поскольку выяснилось, что с высокой статистической достоверностью технические тексты могут быть отделены от художественной литературы, естественно возник вопрос о том, могут ли тексты других научных направлений быть столь же четко разделены по своей тематике. Для этого были отобраны 30 текстов (на русском языке и переводных), которые по экспертному мнению авторов настоящей работы представляют образцы профессионального исследования в областях психологии, философии и теологии.

Психология: Блюма Зейгарник «Патопсихология», Владимир Леви «Искусство быть собой», Дэвид Майерс «Интуиция», Александр Морозов «Деловая психология», Джозеф Чилтон Пирс «Биология трансцендентного», Зигмунд Фрейд «Толкование сновидений», Роберт Хаэр «Пугающий мир психопатов», Зиглер Хьелл «Теории личности», Дэвид Шапиро «Невротические стили», Карл Густав Юнг «Психологические типы».

Философия: Фрэнсис Бэкон «Новый Органон», Георг Вильгельм Фридрих Гегель «Наука логики», Эдмунд Гуссерль «Картезианские размышления», Иммануил Кант «Критика чистого разума», Огюст Конт «Дух позитивной философии», Карл Поппер «Объективное знание», Бертран Рассел «Человеческое познание», Мартин Хайдеггер «Время и бытие», Густав Шпет «Искусство как вид знания», Дэвид Юм – Собрание сочинений.

Теология: Григорий Алфеев «Православие», Николай Бердяев «Самопознание», Дитрих Бонхеффер «Следуя Христу», Тимоти Уэр «Православная церковь», Вальтер Каспер «Бог Иисуса Христа», Серен Кьеркегор «Страх и трепет», Ганс Кюнг «Христианский вызов», Владимир Лосский «Очерк мистического богословия», Павел Флоренский «Столп и утверждение истины», Жак Маритен «Знание и мудрость», Пауль Тиллих «Систематическая теология».

Следует подчеркнуть, что перечисленные три группы авторов были сформированы автоматически посредством кластеризации текстов по их попарной близости в терминах 2-ПФР. Кластеризация и построение тематических эталонных распределений были проведены по программе TRIL [8]. Физико-математические и химические тексты тоже присутствовали в общей группе, но их отделение в самостоятельные кластеры не вызывает особого удивления. Интересно то, что философские произведения по теории познания отделились от теологических, которые тоже, в сущности, философские и тоже по теории познания, но в другом смысле.

В результате оказалось, что каждый из текстов ближе всего по расстоянию между эталонной 2-ПФР и 2-ПФР текста в норме L1 (сумма модулей разностей эмпирических частот буквосочетаний) к эталону своего кластера. При этом сам текст на момент сравнения исключался из эталона своего кластера. Характерное расстояние до эталона своего кластера для всех рассматриваемых текстов составило величину 0,12, тогда как до чужого кластера среднее расстояние примерно в два раза больше и равно 0,23. Этот факт показывает существование объективных различий в рассматриваемых текстах.

Выясним тогда, каково распределение индикатора $\cos(\mathbf{u},\mathbf{f})$ для каждой группы. Ниже на рис. 17—19 приведены спектральные портреты для некоторых текстов каждой группы.

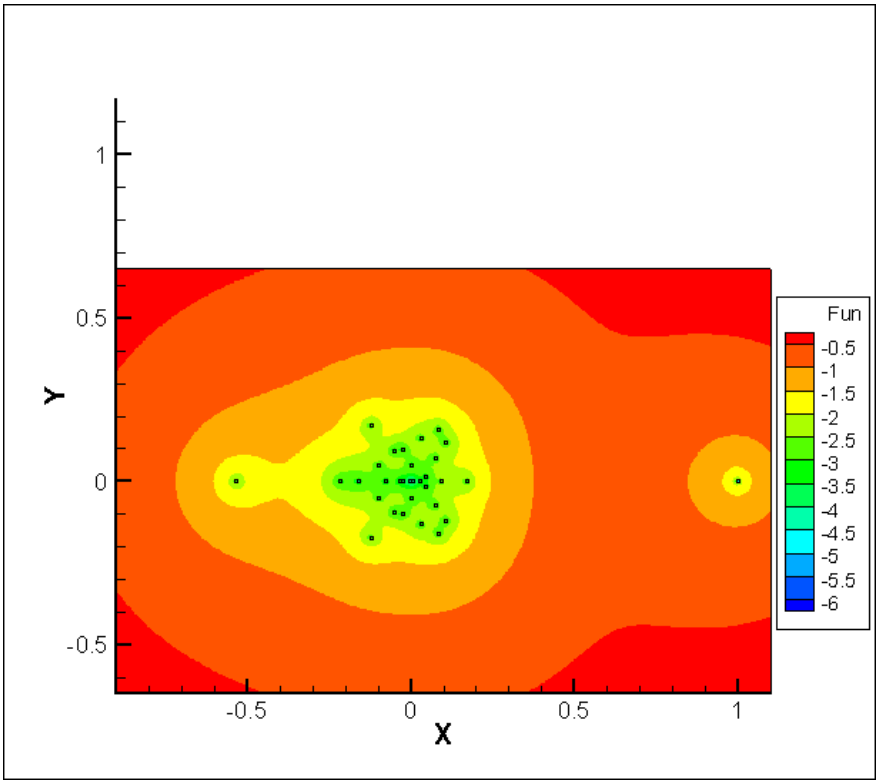

Рис. 17. Карл Густав Юнг (Психологические типы), $\cos(\mathbf{u},\mathbf{f}) = -0{,}044$

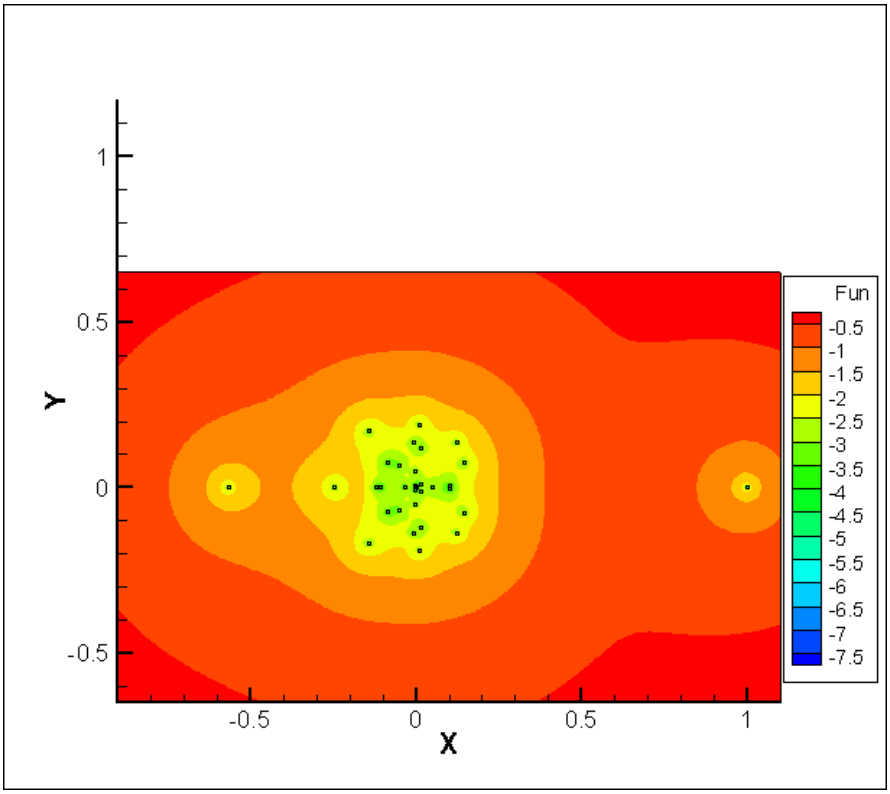

Рис. 18. Иммануил Кант (Критика чистого разума), $\cos(\mathbf{u},\mathbf{f}) = -0{,}055$

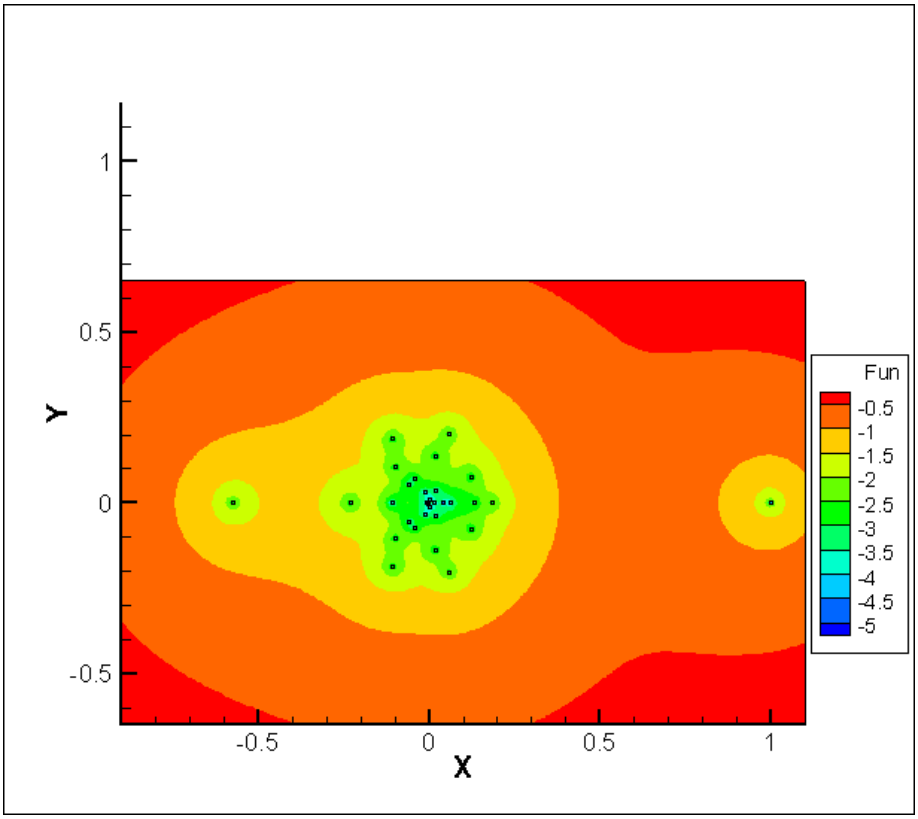

Рис. 19. Николай Бердяев (Самопознание), $\cos(\mathbf{u},\mathbf{f}) = -0{,}069$

Выяснилось, что по распределению косинуса угла между двумя собственными векторами теологи и философы совпадают, в среднем значение косинуса для них составило -0,06. Из рассмотренных направлений гуманитарных наук философия и теология по этому показателю находятся ближе всех к художественной литературе. Характерное же значение «психологического» косинуса составило -0,045, что довольно близко к точным наукам.

Заключение

В работе исследована возможность введения статистического числового индикатора, определяющего тематическую направленность профессиональных текстов. Основной результат представлен на рис. 20, где показаны распределения косинуса угла между векторами **f** и **u** для пяти групп текстов. Для сбора статистики было использовано: литературных текстов 100, научно-технических 100, текстов из гуманитарных наук 30. Выяснилось, что тексты могут быть классифицированы по устойчивости собственных векторов матрицы условных биграмм, связанной со словарным разнообразием текста.

В дальнейшем предполагается проведение анализа по более полному корпусу научных текстов с целью уточнения возможностей их автоматической классификации по отраслям знаний.

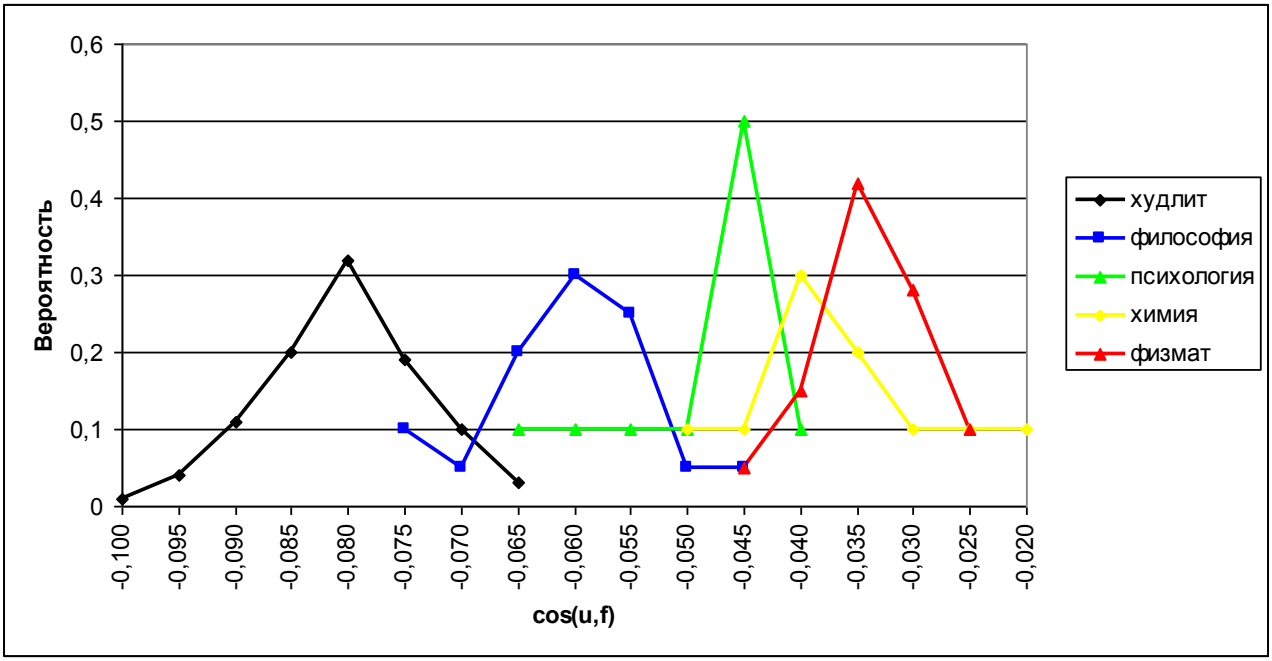

Рис. 20. Эмпирические распределения косинуса угла между векторами главных направлений матрицы условных биграмм для профессиональных текстов

Литература

1. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 312 с.

2. Орлов Ю.Н., Осминин К.П. Определение жанра и автора литературного произведения статистическими методами // Прикладная информатика, 2010. Т. 26. № 2. С. 95-108.

3. Ивченко А.Ю., Орлов Ю.Н. Практические аспекты задачи распознавания образов // Препринты ИПМ им. М.В. Келдыша. 2016. № 17. 20 с.
URL: http://library.keldysh.ru/preprint.asp?id=2016-17

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

5. Голуб Дж., Ван Лоун Ч. Матричные вычисления. Москва: Мир, 1999. – 546 с.

6. Linear Algebra PACKage. URL: http://www.netlib.org/lapack/

7. Арутюнов А.А., Борисов Л.А., Зенюк Д.А., Ивченко А.Ю., Кирина-Лилинская Е.П., Орлов Ю.Н., Осминин К.П., Федоров С.Л., Шилин С.А. Статистические закономерности европейских языков и анализ рукописи Войнича // Препринты ИПМ им. М.В. Келдыша. 2016. № 52. 36 с.
URL1: http://library.keldysh.ru/preprint.asp?id=2016-52

8. Орлов Ю.Н., Осминин К.П. Программный комплекс TRIL для идентификации языка, автора и жанра литературного текста / Свидетельство о государственной регистрации программы для ЭВМ № 2017611570 от 06.02.2017.