Н.А. Митин, Ю.Н. Орлов

**Статистический индикатор тематики
научных текстов**

Рекомендуемая форма библиографической ссылки

Митин Н.А., Орлов Ю.Н. Статистический индикатор тематики научных текстов // Проектирование будущего. Проблемы цифровой реальности: труды 1-й Международной конференции (8-9 февраля 2018 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2018. — С. 74-80. — URL: <http://keldysh.ru/future/2018/11.pdf> doi:[10.20948/future-2018-11](https://doi.org/10.20948/future-2018-11)

Статистический индикатор тематики научных текстов

Н.А. Митин, Ю.Н. Орлов

Институт прикладной математики им. М.В.Келдыша РАН

Аннотация. Метод спектрального анализа стохастической матрицы применяется для построения индикатора, позволяющего определять тематику научных текстов без использования ключевых слов. Научные тексты классифицируются по взаимному расположению инвариантных подпространств матрицы условных вероятностей пар буквосочетаний. Индикатор разделения – косинус угла между собственными правым и левым собственными векторами, отвечающими максимальному и минимальному собственным значениям.

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

Рассматривается достаточно большой (более 10 тыс. знаков) текст на русском языке, из которого исключены пробелы, знаки препинания, цифры 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 :

$$P_{ij} = \frac{F_{ij}}{f_i}. \quad (1)$$

Согласно результатам, представленным в [1, 2], собственные векторы матрицы (1), отвечающие действительным собственным значениям, являются устойчивыми индикаторами автора текста.

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, комплексное число λ принадлежит ε -спектру $\Lambda_\varepsilon(P)$ матрицы P , если существует такая возмущающая ее матрица Δ , что $\|\Delta\| \leq \varepsilon\|P\|$ и

2. Новые возможности и перспективы

$\det(\lambda I - P - \Delta) = 0$, 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, с помощью которого строится спектральный портрет матрицы, основан на специальном представлении параметра дихотомии:

$$H_r(P) = \frac{1}{2\pi} \int_0^{2\pi} (P^+ - re^{-i\varphi} I)^{-1} (P - re^{i\varphi} I)^{-1} d\varphi. \quad (2)$$

Интеграл в (2) сходится только в том случае, если на окружности $\lambda = re^{i\varphi}$ 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 [1] одного из авторов данной работы.

Рис. 1. Спектральный портрет текста монографии [1]

На этом рисунке 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 $\varepsilon = 10^{-2}$.

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.

Наибольшему по модулю отрицательному собственному значению отвечают левый **u** правый **s** собственные векторы, образующие углы с вектором частот **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**, обладают авторской устойчивостью.

Поскольку векторы **f** и **u** приближенно ортогональны, их можно трактовать как главные направления матрицы условных вероятностей биграмм:

$$(u, Pf) \approx 0. \quad (3)$$

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. 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. В результате частая повторяемость определенных буквосочетаний обеспечивает высокую точность элементов матрицы даже на относительно небольших текстах, так что эти векторы оказываются определенными гораздо лучше, чем для художественной литературы. На рис. 2 приведен характерный пример спектрального портрета для классического литературного произведения.

2. Новые возможности и перспективы

Рис. 2. Спектральный портрет романа И.С. Тургенева «Отцы и дети»,
 $\cos(\mathbf{u}, \mathbf{f}) = -0,09$

Авторские спектральные портреты сохраняют индивидуальность, характеризующую писателя: их форма приблизительно совпадает для разных произведений одного и того же автора.

Оказалось, что распределения косинуса угла для текстов по техническим наукам не имеют общего носителя с распределением косинуса для литературных произведений (см. рис. 3). Это наблюдение привело авторов настоящей работы к идее проанализировать тексты основных специальностей ВАК (их 19). Были отобраны по 20 текстов разных авторов из каждой научной области, построены соответствующие матрицы условных вероятностей биграмм и вычислены косинусы углов между векторами главных направлений. 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. В целом тексты гуманитарного характера весьма четко отделяются от технических текстов.

Рис. 3. Распределения косинуса угла для текстов различных специальностей

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

Рис. 4. Упорядочение специальностей ВАК по среднему косинусу

Предлагается следующий способ ранжирования специальностей ВАК. Считаем, по определению, математику наукой. Тогда максимальный

2. Новые возможности и перспективы

косинус математического текста (0,04) определяет правую границу научности. Левая же граница ненаучности 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. Получаем, что юриспруденция на 100% является наукой, что, в общем, логично.

Есть 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.

Есть четыре специальности, целиком лежащие между наукой и литературой: это земледелие, психология, политология, культурология.

Кроме того, есть четыре специальности, которые следует трактовать не как науки, ибо доля науки в их распределениях нулевая, а как плохая литература – по доле распределения косинуса угла в литературной

области, определяемой произведениями русских классиков (нижняя граница косинуса 0,065). Самая «плохая» литература – это история, «почти литература» – искусствоведение (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 или физическая статья, 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, ироничных детективов и боевиков имеют косинус угла порядка 0,05, поэтому такие тексты не следует считать литературой, а относить, скорее, к публицистике.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 16-01-00342.

Литература

1. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС / Книжный дом «ЛИБРОКОМ», 2012. – 312 с.
2. 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.