УДК 81-139+81-112
DOI: 10.15827/0236-235X.132.720-725

Дата подачи статьи: 29.09.20
2020. Т. 33. № 4. С. 720–725

Использование статистических индексов для различения научных и научно-популярных текстов на примере трудов А.Е. Ферсмана

Л.Г. Горбич [1], *научный сотрудник, glg@cbibl.uran.ru*
А.А. Живодеров [1], *к.ф.-м.н., старший научный сотрудник, csl@cbibl.uran.ru*

[1] *Центральная научная библиотека УрО РАН, г. Екатеринбург, 620137, Россия*

С развитием информационной техники и информационных систем актуализировалась проблема разработки методик машинной атрибуции текстов. Эти методики могут быть использованы для автоматического поиска текстов необходимого жанра и стиля и установления авторства с помощью компьютерных технологий.

В основу разработки рассматриваемой в статье методики была положена гипотеза о том, что существуют структурные особенности текста, которые позволяют без учета смыслового содержания отнести его к определенному жанру или автору на основе вычисления чисто количественных значений некоторых параметров и индексов. Авторы наряду с другими исследователями в течение ряда лет занимались разработкой таких индексов и формированием из них оптимального набора и добились в этом определенных успехов. В частности, был сформирован набор индексов, позволяющий правильно классифицировать тексты по жанру с вероятностью до 86 %.

Для решения задачи автоматической классификации научных и научно-популярных текстов авторы применили и усовершенствовали набор статистических индексов, разработанный ими ранее для атрибуции других стилей. В качестве материала исследования были взяты труды академика А.Е. Ферсмана. Одной из особенностей этого автора является стилевая двойственность – наличие большого числа принадлежащих ему как научных, так и научно-популярных текстов, что создало уникальную возможность для попытки решения задачи автоматической классификации стилей текстов, принадлежащих одному автору. В ходе работы было показано, что выборочные средние статистических индексов для текстов двух стилей достоверно различаются. Применяя методы дискриминантного анализа, логистической регрессии и ROC-кривых, авторы продемонстрировали возможность автоматической классификации текстов двух стилей и с помощью оптимизации используемого набора индексов добились существенного повышения качества классификации. Предложен также новый статистический индекс, позволяющий минимизировать вычислительные затраты и успешно (до 100 % точности) решать задачу классификации научных и научно-популярных текстов даже при использовании его в качестве единственного фактора. Результаты исследования были проверены на текстах других авторов.

Ключевые слова*: стиль текста, автоматическая классификация текстов, статистический индекс, дискриминантный анализ, логистическая регрессия, ROC-кривая.*

Стилевая двойственность – наличие большого числа как научных, так и научно-популярных текстов, вышедших из-под пера одного автора, – создает уникальную ситуацию для исследования возможностей автоматической классификации этих стилей, что составляет цель настоящей работы.

Исследование является оригинальным, поскольку задача различения текстов научного и научно-популярного стилей в явном виде не ставилась.

В работе использовались методы статистических индексов, которые были разработаны авторами в результате предыдущих исследований в области поиска методик машинной атрибуции текстов [1–3], а также стандартные методы дискриминантного анализа, логистической регрессии и метод ROC-кривых.

Краткий обзор литературы

Первое представление о том, что такое стиль текста, мы получаем еще в школе. Однако определения, понятные для людей, трудно формализуемы для автоматического, машинного применения. Вместе с тем автоматическая атрибуция текстов является весьма важной задачей, например, в сферах релевантного поиска в сети Интернет или создания тематических каталогов [4, 5].

Использование частотностей букв и их сочетаний для лингвистических целей можно

встретить уже в работах А.А. Маркова [6]. Впоследствии такой подход применялся и другими авторами [7–9].

Относительно недавний пример успешного использования частотностей отдельных букв и их сочетаний для атрибуции текстов можно найти в монографии Ю.Н. Орлова и К.П. Осминина [10]. Они используют все буквы алфавита, но ограничиваются биграммами, авторы данной статьи рассматривают только гласные, но используют в исследованиях и индекс, основанный на сочетаниях трех букв (триграммах). Кроме того, в упомянутой книге развивается подход попарного сравнения текстов – исследуемого с эталонным или с набором эталонных текстов. В настоящем исследовании используется в индексах отличие исследуемого текста от модели, где частотность определенной комбинации букв находится как произведение частотностей отдельных букв, то есть отсутствует корреляция между последовательными буквами. Это позволяет уйти от попарного сравнения текстов и получить некий «абсолютный» индекс.

Тексты для анализа

В качестве материала для исследования были выделены по 22 фрагмента научных и научно-популярных текстов академика Александра Евгеньевича Ферсмана, выдающегося ученого и популяризатора науки, автора более 715 научных статей, научно-популярных книг и учебников. Размер каждого фрагмента составлял приблизительно 200 тысяч печатных знаков. Какая-либо предварительная селекция или коррекция исходных текстовых отрывков не проводилась.

Использованные показатели

В качестве факторов классификации были выбраны расчетные показатели-индексы, успешно применявшиеся авторами в других работах по классификации текстов [1–3]. Следует отметить, что эти индексы являются формально-статистическими, то есть не связанными напрямую с характерным тезаурусом или смысловыми характеристиками текстов, а значит, легко применимыми для автоматической обработки.

Индексы частотности (ИЧ) отдельных гласных букв – отношения количества вхождений какой-то буквы к общему числу гласных букв в фрагменте. Использовались частотности букв «е», «о», «э».

Индексы биграмм (ИБГ) и *триграмм* (ИТГ) гласных букв: из текста исключаются все буквы, кроме гласных, заглавные и строчные варианты написания учитываются как одинаковые буквы, буква «ё» учитывается как «е», рассчитывается количество двоек (или троек) последовательно идущих букв для каждого возможного сочетания, рассчитывается индекс по формуле критерия согласия Пирсона χ^2: $\chi^2 = \sum_{i=1}^{k} \frac{\left(p_i^{theor} - p_i^{emp}\right)^2}{p_i^{theor}}$, где p_i^{theor} – вероятность обнаружения биграмм и триграмм, вычисляемая по формуле как произведение вероятностей обнаружения отдельных букв; p_i^{emp} – вероятность, полученная из обработки реального текста.

Индексы энтропии, вычисленные исходя из представлений о хаосе и порядке в работах [11–13]. В синергетической теории информации В.Б. Вяткина полагается, что при отражении дискретных систем через совокупность своих частей происходит разделение отражаемой информации (I_0) на отраженную и неотраженную части, равные, соответственно, аддитивной негэнтропии (I_Σ) и энтропии отражения (S). Формулы этих разновидностей синергетической информации: $I_0 = \log_2 M$, $I_\Sigma = \sum_{i=1}^{N} \frac{m_i}{M} log_2 m_i$, $S = -\sum_{i=1}^{N} \frac{m_i}{M} log_2 \frac{m_i}{M}$, где M – общее количество элементов в составе системы; N – число частей системы; m_i – количество элементов в i-й части.

В качестве индексов использовалась так называемая R-функция, представляющая собой отношение аддитивной негэнтропии к энтропии отражения: $Rf = \frac{I_\Sigma}{S} = \frac{\text{порядок}}{\text{хаос}}$, а также функция развития: $Df = I_\Sigma S$.

Индексы сжимаемости текстов: отношение размера файла в байтах после сжатия стандартными алгоритмами (Deflate, Bzip2) к исходному размеру файла.

Все индексы рассчитывались с применением информационной системы Corpus, разработанной в ЦНБ УрО РАН. Для статистических расчетов использовалась программа RStudio.

Различие выборочных средних

Для оценки нормальности распределения значений индексов в выборках был применен критерий Шапиро–Уилка. Затем в зависимости от результатов теста для проверки значимости

различий выборочных средних применялись либо t-критерий Стьюдента, либо непараметрический U-критерий Манна–Уитни. В результате получено, что различия средних являются значимыми с p-значением менее 0,05 для всех индексов, за исключением частотности буквы «о». Это позволило с уверенностью перейти к поиску оптимального набора критериев для целевой классификации текстов.

Дискриминантный анализ

Для классификации научных и научно-популярных трудов А.Е. Ферсмана методом дискриминантного анализа был применен набор индексов, успешно использовавшихся в более ранних работах авторов с добавлением энтропийных параметров Rf и Df. Однако значимыми для классификации оказались только пять из них. Приведем их в порядке степени влияния: ИТГ, Df, Rf, ИБГ, Deflate. Все другие индексы слабо влияли на качество классификации.

Представим коэффициенты делящей дискриминантной функции для пяти факторов:

Constant	–9642,3,
ИБГ	–601,5,
ИТГ	161,8,
Df	504,8,
Rf	134,6,
Deflate	–114,7.

Правильность классификации для такой модели составляет более 97 %, что является хорошим результатом, превышающим аналогичные показатели, полученные авторами в предыдущих работах.

Если ограничить количество факторов до двух, то получим модель со следующими коэффициентами:

Constant	–3173,3,
ИТГ	56,7,
Df	167,3.

При использовании такой модели точность классификации составит 86 %.

Логистическая регрессия

Для определения оптимального набора классифицирующих факторов применялся также метод логистической регрессии. Подбор коэффициентов регрессии проводился для разных наборов факторов с целью получения наиболее компактного списка. При подборе учитывались значения *Pr* для каждого фактора. Учет межфакторного взаимодействия не улучшал качество получаемой модели. С целью определения пороговой вероятности для классификации использовались метод ROC-кривых [14, 15], а также визуальная оценка оптимальной границы отсечения исходя из графиков.

Получены два варианта классификационных моделей, которые могут быть приняты как оптимальные. Первая модель основана на трех факторах – ИБГ, ИТГ и Bzip2.

Коэффициенты логистической регрессии для трех факторов:

Intercept	21,63,
ИБГ	–822,66,
ИТГ	223,61,
Bzip2	–109,41.

Обнаружено, что пересечение ROC-кривых чувствительности и специфичности соответствует пороговой вероятности 0,634. При таком значении правильно классифицируются 41 из 44 (93 %) фрагментов текстов. Однако, если установить порог вероятности по визуальной оценке (он будет равным 0,3), можно добиться правильной классификации 42 фрагментов (96 %).

Вторая модель использует всего два фактора – ИБГ и ИТГ, но демонстрирует несколько худшую классификацию на обучающей выборке.

Коэффициенты логистической регрессии для двух факторов следующие:

Intercept	–4,11,
ИБГ	–381,87,
ИТГ	130,53.

Пересечение ROC-кривых соответствует вероятности 0,609, при этом правильно классифицируются 39 фрагментов текста (89 %). При установке порога отсечения равным 0,4 правильная классификация возрастает до 41 текста (93 %).

Мультиколлинеарность и снижение размерности

К недостаткам полученных моделей следует отнести то, что факторы ИБГ и ИТГ оказались сильно скоррелированными: коэффициент корреляции Пирсона составил 0,95. Использование лишь одного из них ухудшает качество моделей. Однако представляется возможным свести эти два фактора к одному, поскольку при линейной аппроксимации зависимостей для разных стилей текстов коэффициент наклона прямой существенно различается, а свободный член близок к нулю. На рисунке

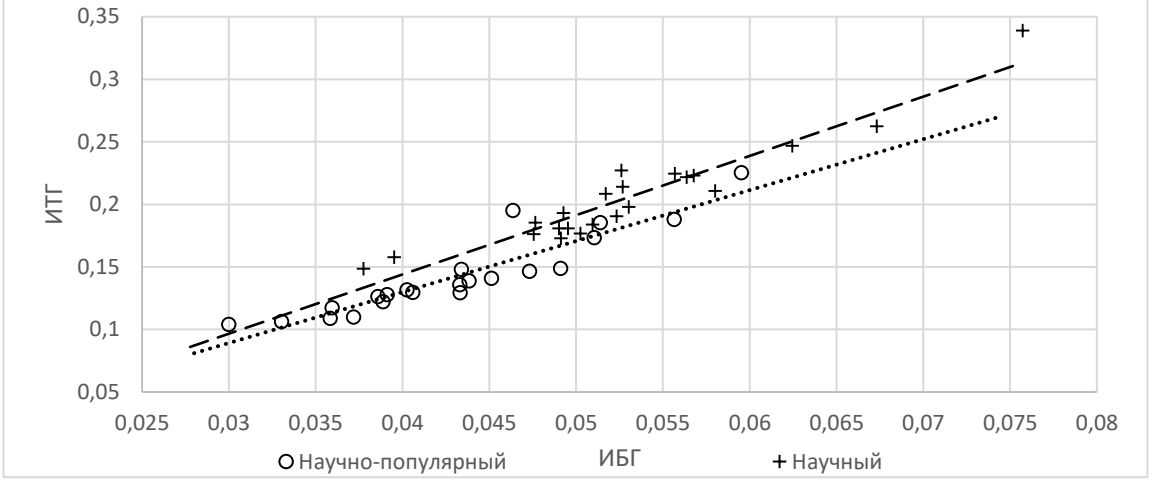

Диаграмма рассеяния для ИБГ и ИТГ

Scatter plot of Trigram index vs Bigram index

представлена диаграмма рассеяния для этих факторов с аппроксимирующими прямыми.

В качестве нового классифицирующего фактора предлагается использовать отношение ИТГ/ИБГ. Можно использовать и обратное соотношение, но, поскольку ИТГ систематически выше, чем ИБГ, значения для первого варианта оказываются больше единицы и визуально оценивать такой показатель проще, чем дробное число.

При использовании предлагаемого комплексного классифицирующего фактора задача классификации сводится к нахождению его порогового значения, то есть всего одного числа. Для рассматриваемой выборки порог по ROC-кривым может быть установлен в 3,6, тогда правильно классифицируются 39 текстов (89 %). Однако, если, пользуясь визуальной оценкой, этот порог определить в 3,5, модель обеспечивает правильную классификацию 41 фрагмента текстов (93 %).

Отношение ИТГ/ИБГ может использоваться не только само по себе, но и в составе набора других факторов. Например, при дискриминантном анализе использование этого соотношения совместно с факторами Df, Rf и Deflate дает точность классификации 98 %.

Проверка моделей классификации

Качество полученных моделей классификации может быть оценено по проверочному набору текстов. Тексты научного стиля были взяты из работ А.Е. Ферсмана, не вошедших в обучающую выборку. Использовались также отрывки из научно-популярных книг других авторов. Эти книги входили в серию «Занимательная наука», издававшуюся в 20–30-х годах прошлого века, в этой же серии были изданы и произведения А.Е. Ферсмана. Авторы книг, входивших в серию, тоже были хорошо известны читателям: Д.О. Святский, Я.И. Перельман, В.В. Рюмин, С.П. Аржанов. Оценка проводилась на 12 текстах научного и 12 текстах научно-популярного стилей. Размер текстовых отрывков, как и в обучающей выборке, был близок к 200 тысячам печатных знаков.

Дискриминантная модель, основанная на пяти факторах, дает корректное распознавание всех текстов (100 %). Модель дискриминантного анализа с двумя параметрами показывает корректное разделение только 22 текстов (92 %).

Логит-модель с тремя факторами: при использовании порога вероятности, равного 0,634 (по ROC-кривым), корректно распознавались все 24 текста (100 %); при пороге 0,3 правильно распознавался 21 текст (88 %) из тестовой выборки.

Логит-модель с двумя факторами: при пороге вероятности, равном 0,603 (по ROC-кривым), правильное распознавание составляло 100 %. Если же принять пороговое значение за 0,4, корректно распознаются только 23 текста (96 %).

Для однофакторной модели, использующей соотношение ИТГ/ИБГ, при пороге в 3,6 (ROC-кривые) и 3,574 (дискриминантный анализ) распознавание составило также 100 %, а при пороге в 3,5, выставленном для максимально правильного распознавания обучающей выборки, на тестовой выборке получим лишь 23 корректно распознанных текста, или 96 %.

Таким образом, чрезмерное желание более точно подобрать параметры модели для обучающей выборки может привести к переобученности и, как следствие, к худшим показателям распознавания на тестовой выборке.

Заключение

В результате удалось доказать возможность использования формально-статистических индексов для задачи различения научных и научно-популярных текстов (на примере трудов А.Е. Ферсмана), оптимизировать набор этих индексов и улучшить качество классификации. Показана также перспективность предложенного подхода для классификации трудов других авторов. Предложен комплексный индекс ИТГ/ИБГ, использование которого существенно улучшает классификацию и упрощает модели.

Литература

1. Филимонов В.В., Живодеров А.А., Горбич Л.Г. Экспрессия и упорядоченность в письменной речи // Изв. УрФУ. Сер. 1: Проблемы образования, науки и культуры. 2012. Т. 104. № 3. С. 313–319.

2. Filimonov V.V., Zhivodyorov A.A., Amieva A.M., Pykhova E.D. A sufficient set of statistical parameters for the classification of Russian-language texts. AIP Conf. Proc., 2018, vol. 2015, no. 1. DOI: 10.1063/1.5055095.

3. Filimonov V.V., Zhivodyorov A.A., Chernykh Y.A., Gorbich L.G. Application of information parameters for the classification of Russian-language texts. AIP Conf. Proc., 2019, vol. 2174, no. 1. DOI: 10.1063/1.5134274.

4. Епрев А.С. Автоматическая классификация текстовых документов // МСиМ. 2010. № 21. С. 65–81.

5. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы. 2017. № 1. С. 85–99. DOI: 10.15827/0236-235X.117.085-099.

6. Марков А.А. Примѣръ статистическаго изслѣдованія надъ текстомъ «Евгенія Онѣгина», иллюстрирующій связь испытаній въ цѣпь // Извѣстія Имп. Академіи Наукъ. 1913. Сер. VI. Т. 7. № 3. С. 153–162.

7. Андреев Н.Е. Статистические и комбинаторные методы в теоретической и прикладной лингвистике. Л.: Наука, 1967. 403 с.

8. Головин Б.Н. Язык и статистика. М.: Просвещение, 1970. 190 с.

9. Хмелев Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестн. МГУ. Сер. 9: Филология. 2000. № 2. С. 115–126.

10. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. М.: URSS, 2011. 300 с.

11. Вяткин В.Б. Хаос и порядок дискретных систем в свете синергетической теории информации // Научный журнал КубГАУ. 2009. № 47. URL: http://ej.kubagro.ru/2009/03/pdf/08.pdf (дата обращения: 10.09.2020).

12. Вяткин В.Б. Характеристическая длина текста // Информатика: проблемы, методология, технологии: матер. XIV междунар. науч.-методолог. конф. 2014. Т. 1. С. 263–266.

13. Vyatkin V.B. A synergetic theory of information. Information, 2019, vol. 10, no. 4, pp. 142. DOI: 10.3390/info10040142.

14. Green D.M., Swets J.A. Signal Detection Theory and Psychophysics. NY: John Wiley and Sons Inc. Publ., 1966, 260 p.

15. Powers D.M.W. The problem with kappa. Proc. EACL, 2012, pp. 345–355.

Software & Systems
DOI: 10.15827/0236-235X.132.720-725

Received 29.09.20
2020, vol. 33, no. 4, pp. 720–725

Using statistical indexes to distinguish between scientific and popular science texts on the example of the works of A.E. Fersman

L.G. Gorbich [1], *Research Associate, glg@cbibl.uran.ru*
A.A. Zhivoderov [1], *Ph.D. (Physics and Mathematics), Senior Researcher, csl@cbibl.uran.ru*

[1] *Central Scientific Library of the Ural Branch of the Russian Academy of Sciences, Ekaterinburg, 620137, Russian Federation*

Abstract. With the development of information technology and information systems, the problem of developing methods for machine attribution of texts has become more relevant. These techniques can be used to automatically search for texts of the required genre and style, and establish authorship using computer technology.

The development of our methodology was based on the hypothesis that there are structural features of the text that allow it to be attributed to a certain genre or author without taking into account the semantic content, based on the calculation of purely quantitative values of certain parameters and indices. The authors of this paper, along with other researchers, have been developing such indices and forming an optimal set of them for a number of years, and have achieved some success in this. In particular, a set of indexes was formed that allows one to correctly classify texts of different authors by genre with a probability of up to 86 %.

To solve the problem of automatic classification of scientific and popular science texts, the authors applied and improved a set of statistical indexes that they had previously developed for attributing other styles. The research material was based on the works of academician A.E. Fersman. One of the features of this author is the style duality – the presence of a large number of scientific and popular scientific texts belonging to him, which created a unique opportunity to try to solve the problem of automatic classification of text styles belonging to one author. In the course of the work, it was shown that the sample averages of statistical indices for texts of the two styles differ significantly. Using the methods of discriminant analysis, logistic regression, and ROC-curves, the authors demonstrated the possibility of automatic classification of texts of two styles and, by optimizing the set of indexes used, achieved a significant improvement in the quality of classification. A new statistical index is also proposed that allows minimizing computational costs and successfully (up to 100 % accuracy) solving the problem of classification of scientific and popular science texts, even when using it as the only factor. The results of the study were checked for texts by other authors.

Keywords: text style, automatic text classification, statistical index, discriminant analysis, logistic regression, ROC-curve.

References

1. Filimonov V.V., Zhivodyorov A.A., Gorbich L.G. Expression and order in written speech. *Izvestia URFU. Ser. 1: the Problems of Education, Science and Culture*, 2012, vol. 104, no. 3, pp. 313–319 (in Russ.).

2. Filimonov V.V., Zhivodyorov A.A., Amieva A.M., Pykhova E.D. A sufficient set of statistical parameters for the classification of Russian-language texts. *AIP Conf. Proc*., 2018, vol. 2015, no. 1. DOI: 10.1063/1.5055095.

3. Filimonov V.V., Zhivodyorov A.A., Chernykh Y.A., Gorbich L.G. Application of information parameters for the classification of Russian-language texts. *AIP Conf. Proc*., 2019, vol. 2174, no. 1. DOI: 10.1063/1.5134274.

4. Eprev A.S. Automatic classification of text documents. *Mathematical Structures and Modeling*, 2010, no. 21, pp. 65–81 (in Russ.).

5. Batura T.V. Automatic text classification methods. *Software & Systems*, 2017, no. 1, pp. 85–99 (in Russ.). DOI: 10.15827/0236-235X.117.085-099.

6. Markov A.A. An example of a statistical study of the text ″Eugene Onegin″ illustrating the connection of tests in a chain. *News of the Imperial Academy of Sciences*, ser. VI, vol. 7, no. 3, pp. 153–162 (in Russ.).

7. Andreev N.E. *Statistical and Combinatorial Methods in Theoretical and Applied Linguistics*. Leningrad, 1967, 407 p. (in Russ.).

8. Golovin B.N. *Language and Statistics*. Moscow, 1970, 190 p. (in Russ.).

9. Khmelev D.V Recognizing the author of the text using A.A. Markov chains. *Moscow State Univ. Bulletin. Ser. 9. Philology*, 2000, no. 2, pp. 115–126 (in Russ.).

10. Orlov Yu.N., Osminin K.P. *Methods of Statistical Analysis of Literary Texts*. 2011, 300 p. (in Russ.).

11. Vyatkin V.B. Chaos and order of discrete systems in the light of the synergetic theory of information. *Polythematic Online Scientific Journal of KubSAU*, 2009, no. 47. Available at: http://ej.kubagro.ru/2009/03/pdf/08.pdf (accessed September 10, 2020) (in Russ.).

12. Vyatkin V.B. Characteristic text length. *Proc. XIV Intern. Sci. and Method. Conf. of Informatics: Problems, Methodology, Technology*, 2014, vol. 1, pp. 263–266 (in Russ.).

13. Vyatkin V.B. A synergetic theory of information. *Information*, 2019, vol. 10, no. 4, pp. 142. DOI: 10.3390/info10040142.

14. Green D.M., Swets J.A. *Signal Detection Theory and Psychophysics*. NY: John Wiley and Sons Inc. Publ., 1966, 260 p.

15. Powers D.M.W. The problem with kappa. *Proc. EACL*, 2012, pp. 345–355.

Для цитирования

Горбич Л.Г., Живодеров А.А. Использование статистических индексов для различения научных и научно-популярных текстов на примере трудов А.Е. Ферсмана // Программные продукты и системы. 2020. Т. 33. № 4. С. 720–725. DOI: 10.15827/0236-235X.132.720-725.

For citation

Gorbich L.G., Zhivoderov A.A. Using statistical indexes to distinguish between scientific and popular science texts on the example of the works of A.E. Fersman. *Software & Systems*, 2020, vol. 33, no. 4, pp. 720–725 (in Russ.). DOI: 10.15827/0236-235X.132.720-725.