DOI 10.32460/ergo-6-2022-0006
УДК 004.415:81-139

УЛУЧШЕНИЕ КАЧЕСТВА СТИЛЕВОЙ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ НА ОСНОВЕ СТАТИСТИЧЕСКИХ ИНДЕКСОВ

Филимонов Виктор Валентинович, старший преподаватель,
Уральский федеральный университет
им. первого Президента России Б. Н. Ельцина
Живодеров Андрей Алексеевич,
кандидат физико-математических наук,
старший научный сотрудник,
Горбич Леонид Геннадьевич, научный сотрудник,
Центральная научная библиотека Уральского отделения
Российской академии наук
Дерябина Екатерина Игоревна, студент,
Уральский федеральный университет
им. первого Президента России Б. Н. Ельцина

Для решения задачи автоматической стилевой классификации текстов были применены методы дискриминантного анализа. В качестве возможных факторов классификации рассматривались индекс триграмм (ИТ), индекс биграмм (ИБ), их отношение (ИТ/ИБ), индекс сжимаемости текстов (Deflate), а также информационные индексы: соотношение порядка и хаоса в системе – так называемая R-функция (Rf), и функция развития (Df). Найдены оптимальные сочетания индексов для решения этой задачи. Удалось добиться значительного улучшения качества классификации текстов при одновременном уменьшении количества используемых индексов по сравнению с предыдущими работами.

Ключевые слова: автоматическая классификация текстов, дискриминантный анализ, n-грамма букв

IMPROVING THE QUALITY OF STYLISTIC CLASSIFICATION OF RUSSIAN-LANGUAGE TEXTS BASED ON STATISTICAL INDEXES

Filimonov Victor Valentinovich, senior lecturer,
Ural Federal University
Zhivodyorov Andrey Alekseevich, Candidate of Physical and
Mathematical Sciences, senior researcher,

Gorbich Leonid Gennad'evich, reseacher,
Central Scientific Library of the Ural Branch
of the Russian Academy of Sciences
Deryabina Ekaterina Igorevna, student,
Ural Federal University

To solve the problem of automatic stylistic classification of texts, methods of discriminant analysis were applied. The trigram index (TI), the bigram index (BI), their ratio (TI/BI), the text compressibility index (Deflate), as well as information indexes: the ratio of order and chaos in the system – the so-called R-function (Rf), and the development function (Df) were considered as possible classification factors. Optimal combinations of indexes for solving this problem are found. It was possible to achieve a significant improvement in the quality of text classification and reducing the number of used indexes in comparison with previous works.

Keywords: automatic text classification, discriminant analysis, character n-gram

Введение

Одним из ярких примеров успешного применения междисциплинарного подхода в научных исследованиях и, как следствие, в современных технологиях является использование методов компьютерной обработки для лингвистических задач. Компьютерные методы используются для решения проблем релевантности поиска в сети «Интернет» и других слабоструктурированных источниках данных при автоматическом формировании аннотаций для атрибуции и классификации текстов.

Для электронных библиотек одним из наиболее перспективных способов использования таких методов является автоматическая классификация произведений по жанрам и стилям. Процедура классификации позволяет отфильтровывать тексты, являющиеся непрофильными для данной библиотеки. В частности, для научной академической библиотеки значимым является выделение научных текстов из всего массива документов.

При классификации текстов (впрочем, как и при классификации любых других объектов) ключевыми являются два решения: выбор алгоритма классификации и выбор факторов классификации – свойств текста, по которым делается заключение о принадлежности текста к тому или иному классу. В качестве алгоритмов в других работах наиболее часто использовались: байесовская классификация, линейная и логистическая регрессии, методы k-ближайших сосе-

дей, деревьев решений и «случайного леса», ансамблевые методы, а также комбинации из этих методов. В последнее время с ростом вычислительных мощностей, появлением специализированных вычислительных устройств и созданием стандартных библиотек обработки все чаще используются методы нейронных сетей разных топологий. Сравнительный обзор методов классификации дается, например, в недавней работе К. Ковсари и др. [1].

В качестве факторов классификации используются такие показатели, как частотности n-грамм букв и n-грамм слов, статистика употребления частей речи и редких слов, а также другие, более сложные характеристики текстов. Сравнительный анализ сочетаний алгоритмов и факторов классификации для решения задачи классификации текстов по жанру можно найти, к примеру, в статье А. Онана [2].

В нашей статье рассмотрены методика и результаты совместного применения метода линейного дискриминантного анализа и метода статистических индексов, ранее уже использовавшиеся авторами [3–11] для распознавания стилистической принадлежности текстов.

Тексты для анализа

В качестве материала для исследования были выбраны тексты разных стилей, изначально написанные на русском языке. Размер произведений варьировался от 8 898 до 3 130 234 символов. Всего в расчетах было использовано 375 текстов, из них: поэзия – 66 текстов (ПО), художественная проза – 62 текста (ПР), публицистика – 47 текстов (ПУ), научно-популярные – 58 текстов (НП), научные – 84 текста (Н), официально-деловые – 58 текстов (ОД). Размер текстов был ограничен снизу величиной в 7000 символов, поскольку в противном случае длина текстов была бы недостаточной для статистического анализа. Сохранялась авторская орфография исходных текстов.

Используемые индексы

В работе использовались семь видов статистических индексов, относящихся к трем разным направлениям в индексной классификации текстов. Это индексы биграмм и триграмм гласных букв (ИБ, ИТ), а также их отношение ИТ/ИБ: 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(1):

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(p_i^{theor} - p_i^{emp}\right)^2}{p_i^{theor}}, \qquad (1)$$

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.

Применялись также индексы энтропии, предложенные в работах В. Б. Вяткина [12-14] и получаемые исходя из представлений о хаосе и упорядоченности в текстах. 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 синтропии (I_Σ) и энтропии отражения (S). Формулы этих разновидностей синергетической информации (2–4) имеют вид:

$$I_0 = log_2 M \quad , \qquad (2)$$

$$I_\Sigma = \sum_{i=1}^{N} \frac{m_i}{M} log_2 m_i \quad , \qquad (3)$$

$$S = -\sum_{i=1}^{N} \frac{m_i}{M} log_2 \frac{m_i}{M} \quad , \qquad (4)$$

где M – общее количество элементов в составе системы, N – число частей системы, m_i –количество элементов в i-й части. В настоящем исследовании расчет аддитивной синтропии и энтропии отражения производился методом «скользящего окна» (фрейма) длиной 416 букв, таким образом, в нашем случае: M – количество букв во фрейме, N – количество букв в алфавите, m_i – число появлений каждой буквы во фрейме.

В качестве факторов классификации в работе использовались R-функция (Rf) и функция развития (Df). R-функция вычислялась как отношение аддитивной синтропии к энтропии отражения (5):

$$Rf = \frac{I_\Sigma}{S} \quad (5).$$

Функция развития определялась с помощью формулы (6):

$$Df = I_\Sigma S \quad (6).$$

В качестве фактора классификации использовался также индекс сжимаемости текстов: отношение размера файла в байтах после сжатия стандартным алгоритмом (Deflate) к исходному размеру файла.

Отметим, что все использованные индексы являются формально-статистическими, то есть они не связанны напрямую со смысловыми характеристиками текстов, а значит, хорошо применимы для автоматической обработки.

Все индексы рассчитывались с применением информационной системы Corpus, разработанной в ЦНБ УрО РАН.

Результаты расчетов

При классификации методом дискриминантного анализа существует возможность из многих предлагаемых факторов классификации выбрать такие наборы, которые обеспечивают максимальное разделение классов. В качестве таких наборов-кандидатов были получены следующие:

1) ИТ/ИБ, Df, Deflate.
2) ИТ, Df, ИБ, Deflate.
3) ИТ, Df, ИТ/ИБ, Deflate.

Отметим, что фактор Rf не вошел ни в один из наборов.

Уровни статистической значимости для этих наборов индексов при классификации текстов приведены в таблицах 1–3.

Таблица 1

Результаты расчета для набора индексов № 1

	Критерий Фишера (6.371)	**p-значение**
ИТ/ИБ	12.11	$< 10^{-6}$
Df	9.99	$< 10^{-6}$
Deflate	110.02	$< 10^{-6}$

Таблица 2

Результаты расчета для набора индексов № 2

	Критерий Фишера (6.371)	**p-значение**
Df	9.27	$< 10^{-6}$
Deflate	35.43	$< 10^{-6}$
ИТ	5.90	0.000007
ИБ	17.47	$< 10^{-6}$

Таблица 3

Результаты расчета для набора индексов № 3

	Критерий Фишера (6.370)	**p-значение**
Deflate	34.18	$< 10^{-6}$
ИБ	37.17	$< 10^{-6}$
ИТ/ ИБ	11.00	$< 10^{-6}$
Df	9.58	$< 10^{-6}$

При итоговой классификации текстов методом дискриминантного анализа были получены результаты, представленные в таблицах 4–6.

Таблица 4

Результаты классификации по набору № 1

		Процент правильной классификации	**Классификация по модели (количество текстов)**					
			ПО	**ПР**	**ПУ**	**НП**	**Н**	**ОД**
Экспертная классификация	**ПО**	39.4	26	17	0	3	20	0
	ПР	87.1	4	54	2	2	0	0
	ПУ	15.4	7	22	8	5	10	0
	НП	37.7	10	9	8	20	6	0
	Н	76.6	8	3	4	3	59	0
	ОД	89.7	1	0	0	0	5	52
	Итог:	59.5	56	105	22	33	100	52

С набором факторов № 1 метод хорошо выделяет прозу, научный и официально-деловой стиль.

Таблица 5

Результаты классификации по набору № 2

		Процент правильной классификации	**Классификация по модели (количество текстов)**					
			ПО	**ПР**	**ПУ**	**НП**	**Н**	**ОД**
Экспертная классификация	**ПО**	74.2	49	17	0	0	0	0
	ПР	90.3	2	56	2	2	0	0
	ПУ	21.2	6	19	11	8	8	0
	НП	49.1	2	6	12	26	7	0
	Н	80.5	2	2	6	3	62	2
	ОД	89.7	0	0	0	1	5	52
	Итог:	69.6	61	100	31	40	82	54

С набором факторов № 2, метод распознает прозу, поэзию, научный и официально-деловой стиль. Общий процент распознания заметно выше, чем для набора № 1.

Таблица 6

Результаты классификации по набору № 3

		Процент правильной классификации	Классификация по модели (количество текстов)					
			ПО	ПР	ПУ	НП	Н	ОД
Экспертная классификация	ПО	75.8	50	16	0	0	0	0
	ПР	91.9	2	57	1	2	0	0
	ПУ	21.2	4	19	11	8	10	0
	НП	43.4	2	7	15	23	6	0
	Н	84.4	2	2	4	3	65	1
	ОД	89.7	0	0	0	1	5	52
	Итог:	70.1	60	101	31	37	86	53

С набором факторов № 3, метод хорошо распознает поэзию, научный и официально-деловой стиль. С высокой точностью распознает прозу. Общий процент распознания самый высокий по сравнению с наборами № 1 и 2.

В таблице 7 приведены коэффициенты канонических дискриминантных функций для классификации текстов при помощи набора индексов № 3, обеспечивающего наилучшие результаты классификации.

Таблица 7

Коэффициенты канонических дискриминантных функций при классификации текстов с оптимальным набором индексов (набор № 3)

	ПО	ПР	ПУ	НП	Н	ОД
ИБ	-125.95	-74.49	-34.42	-21.75	16.74	120.16
ИТ/ИБ	8.78	7.04	7.53	7.38	8.71	9.18
Df	1.58	1.19	1.28	1.49	1.26	2.11
Deflate	-9.30	5.11	-5.82	-3.91	-10.53	-81.25
Константа	-30.83	-23.10	-24.92	-29.73	-29.96	-43.18

Благодаря приведенным коэффициентам дискриминантных функций можно значительно упростить и ускорить дальнейшие расчеты, не прибегая повторно к методу дискриминантного анализа.

Заключение

В результате проведенных исследований удалось добиться значительного улучшения качества классификации текстов при одновременном уменьшении количества используемых индексов по сравнению с предыдущими работами.

Результаты классификации демонстрируют, что использованная методика позволяет успешно распознавать поэзию, научный и официально-деловой стиль. С самой высокой точностью распознается художественная проза. Сравнительно слабое разделение публицистических и научно-популярных текстов, по-видимому, связано с реальной близостью характеристик этих стилей, и их разделение может быть осуществлено посредством других факторов классификации, например, характерных слов.

Список источников

1. Text classification algorithms: A survey / K. Kowsari, K. J. Meimandi, M. Heidarysafa [et al.] // Information. – 2019. – Vol. 10, № 4. – P. 1–68. –URL : https://www.researchgate.net/publication/332463886_Text_Classification_Algorithms_A_Survey (дата обращения: 10.09.2020).

2. Onan A. An ensemble scheme based on language function analysis and feature engineering for text genre classification // Journal of Information Science. – 2018. – Vol. 44, № 1. – P. 28–47.

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

4. Filimonov V. V. Clustering of Russian-language texts using $\chi 2$ statistics / V. V. Filimonov, A. M. Amieva, A. P. Sergeev // Information: transmission, processing, perception. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2016). – Ekaterinburg, 2016. – P. 164–174.

5. Amieva A. M. Application of discriminant analysis to the classification of Russian-language text / A. M. Amieva, V. V. Filimonov, A. A. Zhivodyorov // Proceedings of the 4th international conference (Ekaterinburg, 7–9 December 2017). – Ekaterinburg, 2017. – P. 65–71.

6. Machine attribution of Russian-language texts: a review of methods / A. M. Amieva, A. A. Kramarenko, V. V. Filimonov, A. A. Zhivodyorov // New information technologies in education and science. Proceedings of the X International Scientific and Practical Conference (Ekaterinburg, 27 February-3 March, 2017). – Ekaterinburg : RGPPU, 2017. – P. 371–375.

7. Amieva A. M. Systematic differences statistical characteristics of texts of different genres / A. M. Amieva, V. V. Filimonov, A. A. Zhivodyorov // Information: transmission, processing, perception. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2018). – Ekaterinburg : UrFU, 2018. – P. 140–161.

8. Statistical description of Russian texts: parameters and factors / A. M. Amieva, V. V. Filimonov, A. A. Zhivodyorov, A. A. Kramarenko // Analysis of Images, Social Networks, and Texts. Proceedings of the international scientific-practical conference (Moscow, July 27–29, 2017). – Moscow : CEUR-WS, 2017. – P. 1–8.

9. A sufficient set of statistical parameters for the classification of Russian-language texts / V. V. Filimonov, A. M. Amieva, E. D. Pykhova, A. A. Zhivodyorov // AIP Conference Proceedings : Proceedings of the V International Young Researchers' Conference, Ekaterinburg, 14–18 May 2018 г. – Ekaterinburg, 2018. – P. 020022.

10. Application of information parameters for the classification of Russian-language texts / V. V. Filimonov, Y. A. Chernykh, A. A. Zhivodyorov, L. G. Gorbich // AIP Conference Proceedings : Proceedings of the VI International Young Researchers Conference Physics, Technologies and Innovation, PTI 2019 (Ekaterinburg, 20–23 May, 2019). – Ekaterinburg, 2019. – P. 020123.

11. Горбич Л. Г. Использование статистических индексов для различения научных и научно-популярных текстов на примере трудов А. Е. Ферсмана / Л. Г. Горбич, А. А. Живодеров // Программные продукты и системы. – 2020. – № 4. – С. 720–725.

12. 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).

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

14. Vyatkin V. B. A synergetic theory of information. – DOI: 10.3390/info10040142 // Information. – 2019. – Vol. 10, №. 4. – P. 142.

References

1. Kowsari K., Meimandi K. J., Heidarysafa M. [et al.]. *Text classification algorithms: A survey.* Information, 2019, vol. 10, no. 4, pp. 1–68. URL : https://www.researchgate.net/publication/332463886_Text_Classification_Algorithms_A_Survey (accessed: 10.09.2020).

2. Onan A. *An ensemble scheme based on language function analysis and feature engineering for text genre classification*. Journal of Information Science, 2018, vol. 44, no. 1, pp. 28–47.

3. Filimonov V. V., Zhivodyorov A. A., Gorbich L. G. *Ekspressiya i uporyadochennost' v pis'mennoj rechi* [Expression and order in written speech]. *Izvestia UrFU. Series 1. The problems of education, science and culture,* 2012, no. 3 (104), pp. 313–319. (In Russ.).

4. Filimonov V. V., Amieva A. M., Sergeev A. P. *Clustering of Russian-language texts using χ2 statistics*. Information: transmission, processing, perception. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2016). Ekaterinburg, 2016, pp. 164–174.

5. Amieva A. M., Filimonov V. V., Zhivodyorov A. A. *Application of discriminant analysis to the classification of Russian-language text.* Proceedings of the 4th international conference (Ekaterinburg, 7–9 December, 2017). Ekaterinburg, 2017, pp. 65–71.

6. Amieva A. M., Kramarenko A. A., Filimonov V. V, Zhivodyorov A. A. *Machine attribution of Russian-language texts: a review of methods.* New information technologies in education and science. Proceedings of the X International Scientific and Practical Conference (Ekaterinburg, 27 February–3 March, 2017). Ekaterinburg: RGPPU, 2017, pp. 371–375.

7. Amieva A. M., Filimonov V. V., Zhivodyorov A. A. *Systematic differences statistical characteristics of texts of different genres.* Information: transmission, processing, perception. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2018). Ekaterinburg: UrFU, 2018, pp. 140–161.

8. Amieva A. M., Filimonov V. V., Zhivodyorov A. A., Kramarenko A. A. *Statistical description of Russian texts: parameters and factors.* Analysis of Images, Social Networks, and Texts.Proceedings of the international scientific-practical conference (Moscow, July 27–29, 2017). Moscow: CEUR-WS, 2017, pp. 1–8.

9. 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 Conference Proceedings. Proceedings of the V International Young Researchers Conference (Ekaterinburg, 14–18 May , 2018). Ekaterinburg, 2018, p. 020022.

10. Filimonov V. V., Zhivodyorov A. A., Chernykh Y. A., Gorbich L. G. *Application of information parameters for the classification of Russian-language texts*. AIP Conference Proceedings: Proceedings of the VI International Young Researchers Conference Physics, Technologies and Innovation, PTI 2019 (Ekaterinburg, 20–23 may 2019 r.). Ekaterinburg, 2019, p. 020123.

11. Gorbich L. G., Zhivoderov A. A. *Ispol'zovanie statisticheskih indeksov dlya razlicheniya nauchnyh i nauchno-populyarnyh tekstov na primere trudov A. E. Fersmana* [Using statistical indexes to distinguish between scientific and popular science texts on the example of the works

of A. E. Fersman]. *Programmnye produkty i sistemy,* 2020, vol. 33, no. 4, pp. 720–725 (in Russ.).

12. Vyatkin V. B. *Haos i poryadok diskretnyh sistem v svete sinergeticheskoj teorii informacii* [Chaos and order of discrete systems in the light of the synergetic theory of information]. *Nauchnyj zhurnal KubGAU*, 2009, no. 47. URL : http://ej.kubagro.ru/2009/03/pdf/08.pdf (accessed: 10.09.2020). (in Russ.).

13. Vyatkin V. B. *Harakteristicheskaya dlina teksta* [Characteristic text length]. *Informatika: problemy, metodologiya, tekhnologii: materialy XIV mezhdunar. nauch.-metodolog. konf.*, 2014, vol. 1, pp. 263–266. (in Russ.).

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