ИЗВЕСТИЯ АКАДЕМИИ НАУК РЕСПУБЛИКИ ТАДЖИКИСТАН
ОТДЕЛЕНИЕ ФИЗИКО-МАТЕМАТИЧЕСКИХ, ХИМИЧЕСКИХ,
ГЕОЛОГИЧЕСКИХ И ТЕХНИЧЕСКИХ НАУК
№3 (180), 2020 г.

ИНФОРМАТИКА

УДК 81’322::811.222.8::519.25

З.Д.УСМАНОВ, А.А.КОСИМОВ*

О ВЛИЯНИИ ЦИФРОВОГО ПОРТРЕТА ТЕКСТА НА РАСПОЗНАВАНИЕ АВТОРА ПРОИЗВЕДЕНИЯ

Институт математики им. А.Джураева АН Республики Таджикистан
****Худжандский политехнический институт***
Таджикского технического университета им. акад. М.С.Осими
Поступила в редакцию 26.10.2020 г.

На примере модельной коллекции, количественные описания произведений которой основываются на различных вариантах упорядочения буквенных n-грамм ($n = 1, 2, 3$) с пробелами, выявляются особенности применения γ -классификатора при распознавания автора текста.

Ключевые слова: текст, *n*-грамма, γ-классификатор.

Согласно Рудману[1] современный исследователь может использовать около тысячи разнообразных признаков текста и каждому сопоставлять свой определенный цифровой портрет, формирующий количественный образ текста. В дальнейшем нас будет интересовать специфические широко используемые в γ-классификаторах [1, 2] портреты на основе распределения частотностей элементов текста.

Поясним некоторые понятия, используемые в статье [3].

Определение 1. ***Алфавит*** **–** *упорядоченное множество элементов текста.*

Примерами элементов текста могут служить буквы алфавита естественного языка, буквенные *n*-граммы и слоги, упорядоченные по алфавиту, длины слов и предложений, упорядоченные по возрастанию или убыванию длин и т.д.

Адрес для корреспонденции: *Усманов Зафар Джураевич. 734063, Республика Таджикистан, г. Душанбе, пр. Айни, д. 299/1, Институт математики АН РТ. E-mail: zafar-usmanov@rambler.ru*

[1] Rudman J. The state of authorship attribution studies: Some problems and solutions //Computers and Humanities. – 1998. – Vol.31. – p. 351-365.

Определение 2. ***Цифровым портретом*** (ЦП) ***текста*** *назовём распределение частотности элементов алфавита.*

Примерами ЦП текста являются распределения частотностей символьных, буквенных и словоформных n-грамм, длин слов и предложений и т.д.

В настоящей статье на примерах модельных коллекций текстов устанавливаются особенности ЦП и γ-классификатора в зависимости от упорядочения алфавитных элементов.

1. Состав модельной коллекции текстов, заимствованной из [4], представлен следующими произведениями

классической поэзии:

– А.Рӯдакӣ “Адабиёти пароканда” и “Қасоид”;

– А.Фирдавсӣ “Достони Рустам ва Сӯҳроб” и “Достони Бежан бо Манижа”;

– С.Шерозӣ “Ғазалиёт, қисми 1” и “Ғазалиёт, қисми 2”;

– Ҳ.Шерозӣ “Ғазалиёт, қисми 1” и “Ғазалиёт, қисми 2”;

– Ҷ.Румӣ “Маснавии Маънавӣ, Дафтари Аввал” и “Маснавии Маънавӣ, Дафтари Дуввум”;

современной поэзии:

– А.Суруш “Дафтари 1” и “Дафтари 2”;

– А.Шукӯҳӣ “Баргҳои тиллоӣ” и “Шохаи райҳон”;

– Г.Сафиева “Офтоб дар соя” и “Шӯъла дар санг”;

– И.Фарзона “101-Ғазал” и “Мӯҳри гули мино”;

– М.Турсунзода “Қиссаи Ҳиндустон” и “Ҳасани аробакаш” ;

современной прозы:

– А.Зоҳир “Бозгашт” и “Завол”;

– Г.Муҳаммадиева “Бӯи модар” и “Сафинаи муҳаббат”;

– М.Шакурӣ “Садри Бухоро” и “Хуросон аст ин ҷо”;

– С.Турсун “Нисфирӯзӣ” и “Повести Камони Рустам”;

– С.Айнӣ “Дохунда” и “Марги судхӯр”.

Таким образом, модельная коллекция составлена из 3-х частей: классической и современной поэзиями и современной прозой. В каждой части по 5 авторов, от каждого автора по 2 произведения.

2. Примеры текстовых элементов и их алфавитов. При изложении данного вопроса ограничимся рассмотрением простейших случаев, когда в качестве элементов текста выбираются n-граммы ($n = 1, 2, 3$) с пробелами.

Для униграмм ($n = 1$) естественных языков существующие алфавиты уже являются отсортированными в определенном порядке конечными множествами букв (также и с учётом пробела). Лексикографический порядок, аналогичный алфавитной сортировке, алфавитизирует также n-граммы ($n \geq 2$) и более сложные буквенно-символьные комбинации. Однако в дополнение к сказанному отметим, что такие комбинации, упорядоченные каким-либо другим способом, будут также называться алфавитными элементами текста. Как будет отмечено в п. 4, расстояния между цифровыми портретами текстов зависят от порядка элементов алфавита и потому не ясно, какому из допустимых алфавитов следует отдать предпочтение.

3. Цифровой портрет текстов и расстояния между ними. После выбора фиксированного алфавита цифровой портрет текста T удобно представлять в табличной форме:

$$\begin{array}{ll} \bar{N}: & 1 \quad 2 \quad . \; . \; . \quad m \\ P: & p_1 \quad p_2 \; . \; . \; . \quad p_m \end{array} \tag{1}$$

в которой m - число элементов алфавита, строка $\bar{N}$ указывает номера упорядоченных элементов алфавита, а строка P −их относительные частоты встречаемости в T, причём

$$\sum_{k=1}^{m} p_k = 1.$$

Цифровой портрет можно задавать также дискретной функцией

$$F(s) = \sum_{k=1}^{s} p_k \;\; (s = 1, \cdots, m),$$

характеризующей распределение в тексте частот встречаемости элементов алфавита.

Определение 3. *Расстоянием между двумя текстами называется расстояние между их цифровыми портретами, отнесенными к единому алфавиту.*

Пусть T_1, T_2 –произвольная пара текстов из коллекции $\mathbb{T}$ и

$$F^{(\alpha)}(s) \; = \sum_{k=1}^{s} p_k^{(\alpha)} - \tag{2}$$

соответствующие им дискретные функции, $\alpha = 1, 2$ и $s = \; 1, \cdots, m$.

Определение 4. *Расстоянием между текстами T_1 и T_2 называется положительное число $\rho\,(T_1\,, T_2)$, определяемое по формуле*

$$\rho\,(T_1\,, T_2) = \sqrt{m/2}\; max_s \left| F^{(1)}(s) - \; F^{(2)}(s) \right|, \tag{3}$$

то есть расстояние между двумя текстами вычисляется как максимальное расстояние по оси ординат между их дискретными функциями $F^{(1)}(s)$ и $F^{(2)}(s)$, помноженное на весовой коэффициент $\sqrt{m/2}$.

Замечание. Условие $\rho\,(T_1\,, T_2) = 0$ означает тождество цифровых портретов текстов, то есть ЦПT_1 = ЦПT_2, но не $T_1 = \; T_2$, то есть идентичность текстов.

4. Обработка данных коллекционного материала, представленного в п.1, состояла из 3 этапов.

Этап 1. Использование для всех произведений трёх частей коллекции трёх типов текстовых элементов:

- униграмм с учетом пробела (в таджикском языке 35 букв алфавита, потому общее число униграмм - 36):
- биграмм с учетом пробела (общее число таковых - $36^2 = 1296$):
- триграмм с учетом пробела (число таковых - $36^3 = 46656$).

Множества n-грамм ($n = 1, 2, 3$) в зависимости от упорядочения своих элементов рассматриваются в 4-х вариантах:

1) элементы располагаются в алфавитном порядке с пробелом в качестве последнего элемента алфавита (обозначается как ABC)[2];
2) элементы располагаются в порядке, обратном алфавитному с пробелом в качестве первого элемента алфавита (обозначается как CBA)[3];
3) элементы располагаются в порядке убывания их частотности в тексте (обозначается символом " ↘");
4) элементы располагаются в порядке возрастания их частотности в тексте (обозначается символом " ↗ ").

Этап 2. Для каждого из 4-х вариантов упорядочения n-грамм ($n = 1, 2, 3$) путём автоматической обработки формируются в табличном виде (1) цифровые портреты всех произведений коллекции и затем по формулам (2) и (3) вычисляются расстояния между парами текстов на таджикском языке по отдельности из классической поэзии, современной поэзии и современной прозы. Из-за большого количества расстояний (таковых 135 = 3 х 45) мы не приводим итоговых результатов, однако обращаем внимание на тот факт, что расстояния, вычисляемые между любыми двумя текстами для различных вариантов расположения алфавитных элементов, оказываются в общем случае различными. В этом можно убедиться на простых примерах.

Этап 3. Настройка γ-классификатора – алгоритма, зависящего от одного вещественного параметра γ и устанавливающего в пределах модельной коллекции соответствие между текстами и их авторами. Существо настройки заключается в определении такого значения γ, при котором произведения одного автора "γ-однородны", а разных авторов – "γ-неоднородны". Однородность всех текстов одного автора в рамках математической модели означает справедливость неравенства

$$\rho\,(T_1\,, T_2) \leq\ \gamma, \tag{4}$$

а неоднородность любых двух текстов разных авторов – справедливость неравенства

$$\rho\,(T_1\,, T_2) > \gamma. \tag{5}$$

Ошибки в настройке γ-классификатора выявляются в случае, когда для каких-то пар текстов одного и того же автора вместо неравенства (4) имеет место неравенство (5), а также в случае, когда какие-то два произведения двух различных авторов удовлетворяют неравенству (4) вместо того, чтобы выполнялось неравенство (5).

Суммарное количество $\tau = \tau(\gamma)$ допущенных ошибок одновременно в двух случаях позволяет подсчитать величину π эффективности γ-классификатора при распознавании авторов текста по формуле

$$\pi = 1 -\ \tau(\gamma)/L\,, \tag{6}$$

[2] Для биграмм и триграмм – с двумя и тремя пробелами в конце.

[3] Для биграмм и триграмм – с двумя и тремя пробелами в начале.

где $L = 45$ – число взаимных расстояний между всеми парами произведений из классической и современных поэзий, а также из современной прозы. Детальное описание алгоритма для нахождения оптимального значения γ, при котором π принимает максимальное значение, содержится в статьях [2, 3].

Итоги применения трёх этапов автоматической обработки модельной коллекции текстов показаны в табл. 1-3, соответственно для 3-х частей коллекции.

Таблица 1

Значения π и γ для произведений классической поэзии

Элементы текста	Число элементов алфавита	Порядок элементов алфавита	π-эффективность распознавания автора	Оптимальный γ-полуинтервал
уни-граммы	36	А В С	0,98	[0.0354; 0.0447)
		С В А	0,98	[0.0354; 0.0447)
		по ↘	0,98	[0.0337; 0.0342)
		по ↗	0,98	[0.0337; 0.0342)
би- граммы	1296	А В С	0,98	[0.2987; 0.3551)
		С В А	0,98	[0.2987; 0.3551)
		по ↘	0,96	[0.2065; 0.2212)
		по ↗	0,96	[0.2065; 0.2212)
три- граммы	46656	А В С	1,00	[2.1630; 2.1648)
		С В А	1,00	[2.1630; 2.1648)
		по ↘	0,96	[1.2426; 1.4051)
		по ↗	0,96	[1.2426; 1.4051)

В этой таблице также как и в двух последующих, в столбце третьем для описания порядка следования алфавитных элементов приняты обозначения, введенные в п. 4, этап 1.

Таблица 2

Значения π и γ для произведений современной поэзии

Элементы текста	Число элементов алфавита	Порядок элементов алфавита	π-эффективность распознавания автора	Оптимальный γ-полуинтервал
уни-граммы	36	А В С	0,98	[0.0268; 0.0423)
		С В А	0,98	[0.0268; 0.0423)
		по ↘	0,98	[0.0384; 0.0415)
		по ↗	0,98	[0.0384; 0.0415)
би- граммы	1296	А В С	0,98	[0.2318; 0.2816)
		С В А	0,98	[0.2318; 0.2816)
		по ↘	0,98	[0.2484; 0.2745)
		по ↗	0,98	[0.2484; 0.2745)
три- граммы	46656	А В С	0,98	[1.3885; 1.7054)
		С В А	0,98	[1.3885; 1.7054)
		по ↘	0,98	[1.5556; 1.6453)
		по ↗	0,98	[1.5556; 1.6453)

Таблица 3

Значения π и γ для произведений современной прозы

Элементы текста	Число элементов алфавита	Порядок элементов алфавита	π-эффективность распознавания автора	Оптимальный γ-полуинтервал
уни-граммы	36	А В С	0,96	[0.0285; 0.0336)
		С В А	0,96	[0.0285; 0.0336)
		по ↘	0,91	[0.0165; 0.0236)
		по ↗	0,91	[0.0165; 0.0236)
би- граммы	1296	А В С	0,93	[0.2216; 0.2272)
		С В А	0,93	[0.2216; 0.2272)
		по ↘	0,91	[0.2386; 0.2568)
		по ↗	0,91	[0.2386; 0.2568)
три- граммы	46656	А В С	0,96	[1.3379; 1.3412)
		С В А	0,96	[1.3379; 1.3412)
		по ↘	0,91	[0.7450; 1.3704)
		по ↗	0,91	[0.7450; 1.3704)

5. Заключение. Из представленных в 4-х и 5-х колонках результатов вычислений напрашиваются следующие выводы:

1) наивысшее значение $\pi = 1$ коэффициента эффективности распознавания автора текста реализуется для произведений классической поэзии на триграммах, упорядоченных как по АВС, так и по СВА;
2) значения коэффициентов π эффективности на основе порядков АВС и СВА расположения n-грамм ($n = 1, 2, 3$) равны;
3) значения коэффициентов π эффективности на основе порядков расположения n-грамм ($n = 1, 2, 3$) по убыванию (↘) или возрастанию (↗) также равны;
4) значение коэффициента π эффективности на основе порядка АВС и СВА расположения n-грамм ($n = 1, 2, 3$) не ниже значения, основанного на порядке расположения n-грамм ($n = 1, 2, 3$) по убыванию (↘) или возрастанию (↗);
5) коэффициент π эффективности распознавания автора произведений современной поэзии как для любых n-грамм ($n = 1, 2, 3$), так и для всех вариантов их упорядочения, определяется значением 0.98;
6) коэффициенты π для произведений современной прозы несколько ниже аналогичных значений для произведений классической и современной поэзии;
7) полуинтервалы оптимальных значений γ для двух противоположных порядков расположения n-грамм ($n = 1, 2, 3$) одинаковы.

Из огромного количества всевозможных вариантов упорядоченного расположения элементов текста были рассмотрены только четыре: два из них - связанных с алфавитным порядком и два других – с учётом частотности элементов. Именно в этих двух случаях, прямого и обратного порядков упорядочения элементов, расстояния между любыми парами произведений оказывались равными, вследствие чего равными оказывались коэффициенты π эффективности γ-классификатора (см. п.п. 2 и 3 заключения), а также и полуинтервалы оптимальных значений γ (см. п. 7 заключения). Другие допустимые варианты нуждаются в специальном исследовании.

ЛИТЕРАТУРА

1. Усманов З.Д. Классификатор дискретных случайных величин. – ДАН РТ, 2017, т. 60, № 7-8, с. 291-300.
2. Усманов З.Д. Алгоритм настройки кластеризатора дискретных случайных величин. – ДАН РТ, 2017, т. 60, № 9, с. 392-397.
3. Усманов З.Д. Оценка эффективности применения γ-классификатора для атрибуции печатного текста. – ДАН РТ, 2020, т. 63, № 3-4, с.172-179
4. Косимов А.А., Бахтеев К.С. О распознавании автора текста на основе частотности длин предложений. – ДАН РТ, 2020, т. 63, № 3-4, с. 180-186

З.Ҷ.УСМОНОВ, А.А.ҚОСИМОВ*

ОИД БА ТАЪСИРИ СИМОӢ РАҚАМИИ МАТН БАРОИ МУАЙЯНКУНИИ МУАЛЛИФИ АСАРҲО

Институти математика ба номи А.Ҷӯраев
Академияи илмҳои Ҷумҳурии Тоҷикистон
****Донишкадаи политехникии Донишгоҳи техникии Тоҷикистон***
ба номи академик М.С.Осими дар ш. Хуҷанд

Дар мисоли амсилаи маҷмӯъ, тавсифи миқдории асарҳо, ки дар вариантҳои гуногун ба тартиб овардашудаи n-грамм (n = 1,2,3)-ҳои ҳарфӣ бо фосила асос ёфтаанд, хусусиятҳои истифодаи γ-таснифкунанда ҳангоми шинохтӣ муаллифи матн ошкор гардид.

Калимаҳои калидӣ: матн, n-грамма, γ-таснифгар.

Z.D.USMANOV, A.A.KOSIMOV*

ABOUT THE INFLUENCE OF THE DIGITAL TEXT PORTRAIT ON THE RECOGNITION OF THE AUTHOR OF THE WORKS

A.Dzhuraev Institute of Mathematics, Academy of Sciences of Republic of Tajikistan
****Khujand's Polytechnic Institute of Tajik Technical University***

By the example of a model collection, the quantitative descriptions of the works of which are based on various variants of ordering alphabetic n-grams (n = 1,2,3) with spaces, features of the use of the γ-classifier in recognizing the author of the text are revealed.

Key words: text, n-gram, γ-classifier.