УДК 004.912

ОПРЕДЕЛЕНИЕ АВТОРСТВА ТЕКСТОВ НА ОСНОВЕ ПОДХОДА N-ГРАММ

[1]Леонова А.В., [2]Леонова И.В.

[1]Санкт-Петербургский государственный университет, Санкт-Петербург, e-mail: aniliar@mail.ru;
[2]Краснодарское высшее военное авиационное училище летчиков имени Героя Советского Союза А.К. Серова, Краснодар, e-mail: irileonova@mail.ru

Статья знакомит с исследованием в области автоматизации определения авторства текстов. Выделяются и описываются характерные особенности методов и подходов решения задачи атрибуции. Эти методы делятся на два принципиально разных подхода: экспертный и формальный. При реализации экспертного подхода исследование текста осуществляется экспертом-лингвистом и во многом зависит от информации об авторе, которая содержится в самом тексте, а также от уровня квалификации специалиста. Поэтому он носит субъективный характер. При формальном подходе к решению задачи атрибуции используется статистический анализ и машинное обучение. В связи с этим формальный подход имеет объективный характер, что обусловливает актуальность его развития на базе современных информационных технологий. Актуальность статьи состоит в предложенном методе решения современной задачи определения авторства текстов, в котором особое внимание акцентируется на изучении подхода N-грамм. Цель исследования заключается в проведении детального анализа сценария полной кластеризации авторства. Методом исследования стал подход N-грамм, который имеет большую область применения: от теоретической математики до музыки. Используемый подход является простым, эффективным, не требующим сложной предварительной обработки входного текста и допускающим орфографические, грамматические ошибки. Автором выбрано восемь текстов на русском языке, которые подготовлены к использованию и преобразованы согласно условиям поставленной задачи, а также представлен алгоритм и анонсирована компьютерная программа, позволяющая автоматизировать решение задачи атрибуции текстов. Проведенное исследование позволило разработать авторские профили на основе N-грамм, создать коллекцию известных писателей, автоматизировать процедуру определения авторства текстов. Апробация разработанной компьютерной программы атрибуции текстов показала положительный результат.

Ключевые слова: алгоритм, компьютерная программа, кластеризация, классификация, подход N-грамм, задача атрибуции

DETERMINATION OF THE N-GRAM-BASED TEXT AUTHORSHIP

[1]Leonova A.V., [2]Leonova I.V.

[1]Saint Petersburg State University, Saint Petersburg, e-mail: aniliar@mail.ru;
[2]Krasnodar Air Force Academy named after the Hero of the Soviet Union A.K. Serov, Krasnodar, e-mail: irileonova@mail.ru

The article introduces the research in the field of automation of determining the authorship of texts. Allocated and describes the characteristics of the methods and approaches of solving the problem of attribution. These methods are divided into two fundamentally different approaches: expert and formal. In the implementation of the expert approach, the study of the text is carried out by an expert linguist and largely depends on the information about the author, which is contained in the text itself, as well as on the level of qualification of the specialist. Therefore, it is subjective. The formal approach to the attribution problem uses statistical analysis and machine learning. In this regard, the formal approach is objective, which determines the relevance of its development on the basis of modern information technologies. The relevance of the article is in the proposed method of solving the modern problem of determining the authorship of texts, which focuses on the study of the approach of N-grams. The purpose of the study is to conduct a detailed analysis of the scenario of full clustering of authorship. The method of research was the approach of N-grams, which has a large field of application: from theoretical mathematics to music. The approach used is simple, effective, does not require complex pre-processing of the input text and allows spelling, grammatical errors. The author has chosen eight texts in Russian, which are prepared for use and converted according to the conditions of the task, as well as the algorithm and announced a computer program that allows you to automate the solution of the problem of attribution of texts. The conducted research allowed to develop author's profiles on the basis of N-grams, to create a collection of famous writers, to automate the procedure of determining the authorship of texts. Testing of the developed computer program of text attribution showed a positive result.

Keywords: algorithm, computer program, clusterization, classification, N-gram approach, attribution problem

Существует несколько методов установления авторства текстов. Они делятся на два принципиально разных подхода: экспертный и формальный. При реализации экспертного подхода исследование текста осуществляется экспертом-лингвистом, который изучает прямые указания авторства (если таковые есть в тексте), к которым относятся собственное имя, прозвище, псевдоним, геральдика, генеалогия. Если в тексте нет прямых указаний на автора, то в этом случае исследуют косвенные данные, такие как общепринятые обращения к лицам, принадлежащим к разным социальным слоям, историю государственных учреждений, государственные и церковные награды

(данные фалеристики), печати различных учреждений (сведения сфагистики) [1].

Большой вклад в развитие теории авторского стиля внес В.В. Виноградов, изучая особенности использования фразеологических оборотов, частоты применения терминов, особых слов и выражений в процессе написания текстов. Согласно этой теории авторство устанавливалось в первую очередь на основе фразеологических и лексических особенностей, а во вторую – на грамматических [2]. Однако этот метод не вполне решает задачу атрибуции, так как автор текстового источника мог подражать другому писателю, или фразеологические и лексические признаки были присущи группе писателей одного жанра, или текст содержал большое количество цитат, что лишало его индивидуальности.

Формальный подход при решении задачи атрибуции базируется на методах идентификации, которые в свою очередь делятся на две большие группы: статистический анализ и машинное обучение. Статистический анализ подразделяется на одномерный, который включает в себя критерии Стьюдента, хи-квадрат Пирсона, двусторонний критерий Фишера, QSUM и многомерный, представленный критериями Колмогорова – Смирнова, хи-квадратом Пирсона для распределений, линейным дискретным анализом, методом главных компонент, энтропийным подходом, марковскими цепями, сложностным подходом, задачами кластеризации и классификации [3].

Машинное обучение базируется на таких методах, как нейронные сети, машина опорных векторов, метод *k* ближайших соседей, генетические алгоритмы, деревья решений, байесовский классификатор. Развитие методов идентификации атрибуции текстов имеет важное значение, так как они позволяют упростить процесс определения автора текстов, который является сложным, кропотливым и времятратным. Развитие информационных технологий позволило сократить времятраты на процесс анализа текста и его последующей атрибуции.

Стремительный рост информации, представленной и передаваемой в цифровом виде, формирует интерес к задаче определения авторства текстов. Задача атрибуции (установления авторства текстов) встречается в различных областях и представляет интерес для филологов, литературоведов, юристов, криминалистов, историков. Поэтому возникла потребность в создании формальных методов ее решения. Экспертный анализ авторского стиля является трудоемким процессом, поэтому в данной работе рассматривается подход, позволяющий автоматизировать анализ информации и ее атрибуции.

В данной работе проводится детальный анализ сценария полной кластеризации авторства, метода N-грамм, с последующим определением *k* различных авторов в коллекции ровно к одной группе кластера.

Материалы и методы исследования

В настоящее время анализ авторского стиля возможен с помощью задач кластеризации и классификации текстов по различным категориям, проверки на плагиат, идентификации авторов. Для решения задач кластеризации большую популярность приобретают методы, основанные на моделях сжатия текстов, а также на обучающихся системах. К обучающимся системам относятся методы нейронных сетей, деревьев решений, машин опорных векторов [3].

N-граммы могут применяться в широкой области наук: теоретической математике, биологии, картографии, музыке, генетике, а также для кластеризации серии спутниковых снимков Земли из космоса, в компьютерном сжатии, для индексирования данных в поисковых системах [4].

Этот подход является относительно простым, но эффективным, допускающим орфографические и грамматические ошибки и не требующим сложной предварительной обработки входного текста. Иногда необходима только базовая фильтрация: удаление пробелов и / или знаков препинания.

Для решения задачи кластеризации авторских текстов с применением подхода N-грамм возьмем N равным 4 и рассмотрим распределение комбинаций из этого числа букв. Проанализируем результат на подготовленном наборе данных. В коллекции должно быть задано некоторое количество (k) различных авторов и каждый документ должен быть соотнесен ровно к одной группе. Для идентификации авторского текста предполагается, что текст отображает индивидуальный стилевой профиль автора. Пример четырехграммового профиля представлен на рисунке, где N возьмем равным 4.

"каратооркара"

ai:	fi:
кара	2/9
арат	1/9
рато	1/9
атоо	1/9
тоор	1/9
оорк	1/9
орка	1/9
ркар	1/9

Четырехграмм

Профиль D текста определялся как множество пар $\{(a_1, f_1), (a_2, f_2), (a_3, f_3), ...\}$, где f_i-нормированная частота встречаемости N-грамм a_i в тексте. Кроме того, тексты, авторство которых определяется четко, образуют профиль своего автора все вместе. Для сопоставления текстов друг с другом требуется числовая характеристика, которая отображала бы связь между произведениями одного и того же автора. В данной работе рассмотрена метрика L_1, которая вычисляется по формуле

$$d_1 = \|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i| (1),$$

где $p = (p_1, p_2, \ldots, p_n)$ и $q = (q_1, q_2, \ldots, q_n)$ – векторы.

Таким образом, следующий шаг исследования заключался в определении расстояния между текстовым профилем и профилем любого автора. Автор, расстояние к профилю которого является наименьшим, считался создателем данного текста.

Ниже приведен алгоритм классификации текстов, основанный на подходе N-грамм [5–7].

Алгоритм

T – набор неопознанных текстов
t – набор текстов известного авторства
A – набор авторов
tr(a) возвращает все тексты автора a из множества t

for a∈A ***do***
построение авторского профиля Da, где Da = D конкатенация всех текстов в tr(a)
End for

for x∈T***do***
построение профиля Dx
a* <– argmin_a (distance(Da, Dx))
end for

Набор данных, используемых в исследовании

Автор	Произведения	Минимальное количество символов	Максимальное количество символов
Кир Булычев	– Чудеса в Гусляре – Вирусы не отстирываются – Монументы Марса – Театр Теней – 3 – Город без памяти (Алиса Селезнева)	332 069	540 930
Аркадий и Борис Стругацкие	– Малыш – Пикник на обочине – Трудно быть богом – Сказка о Тройке – Понедельник начинается в субботу	213 711	310 178
Сергей Лукьяненко	– Ночной дозор – Лорд с планеты Земля – Дневной дозор – Мальчик и тьма – Тринадцатый город	283 485	530 243
Оксана Панкеева	– Распутья. Добрые люди – О пользе проклятий – Поспорить с судьбой – Путь, выбирающий нас – Дороги и сны	424 787	600 851
Александр Тюрин	– Зона посещения – Правда о Николае I – Клетка для буйных – Вооруженное восстание животных – Война и мир Ивана Грозного	231 467	474 814
Наталья Щерба	– Часограмма – Быть ведьмой – Часовой ключ – Двуликий мир – Часовое сердце	419 140	515 824
Сергей Щеглов	– Часовой Армагеддона – Разводящий апокалипсиса – Начальник судного дня – Банной горы хозяин – Дипломат особого назначение	446 436	598 538

Представленный подход определялся следующими параметрами: N – длина комбинации символов; L – количество наиболее встречаемых N-грамм, T/t – соотношение обучающих к тестовым текстам авторов, Size – длина текста – количество символов без учета пробелов, знаков препинаний.

Для этого подхода наиболее существенными являлись параметры N и L, в то время как T/t и Size носили более общий характер оценки алгоритма.

В данной работе взяты следующие данные: N = 4, L = 700, T/t = 4/1, 180000 < Size < 590000 и подготовлен набор данных, который включал тексты на русском языке восьми писателей: Кира Булычева, братьев Аркадия и Бориса Стругацких, Сергея Лукьяненко, Оксаны Панкеевой, Александра Тюрина, Натальи Щерба, Сергея Щеглова. Использованные данные представлены в таблице.

Результаты исследования и их обсуждение

На первом этапе проведенного исследования выбраны тексты советских, российских писателей и преобразованы под условия задачи установления авторства. Для этого все символы, кроме буквенных, удалены, а все буквенные – переведены в нижний регистр. На втором этапе исследования создавался профиль авторов посредством формирования множество пар N-грамм и частоты их встречаемости в документах (ai, fi). Из полностью построенных профилей авторов Dai выбиралось L количество наиболее встречаемых N-грамм для сравнения с профилем неизвестного автора Dx, построенного по такому же принципу. После чего проводилось нормирование значения частоты встречаемости N-грамм fi. Для этого вычислялась сумма ненормированных fi и выполнялась нормировка по следующей формуле:

$$f_{i_norm} = \frac{f_{i_{non}-norm}}{\sum_1^L f_{i_{non}-norm}},$$

где $fi_{\text{non-norm}}$ – ненормированные величины частотности N-грамма ai, L – количество наиболее встречаемых N-грамм. Затем необходимо было произвести оценку расстояния между двумя профилями с помощью суммы «расстояний» по каждому N-грамму профиля Dx. Если в профиле Da не содержалось N-грамма из Dx, то значение частотности fi принималось равным нулю. Идентификация текста наступала при условии наименьшего расстояния между профилями Da и Dx.

Разработанный алгоритм и коллекция текстовых профилей авторов положены в основу компьютерной программы, написанной на языке C#, которая позволила автоматически проводить экспертный анализ текстов и определять создателя неизвестного текста. На разработанном наборе данных результаты работы алгоритма с оценкой L1 нормы показало хороший результат точности 83–98 %, которая рассчитана как процент правильно определенных авторов. При атрибуции текста в процессе построения профиля автора на основе всего текста, а не фрагмента показатель точности выше. Полученный высокий результат точности, возможно, связан с несбалансированностью тестовой коллекции по длине, которая содержит небольшое количество авторов и длинные тексты, и с тем, что обучение проводится на большом количестве документов и позволяет создавать отличные (контрастные/ярко характерные) авторские профили.

Выводы

Проведенное исследование позволило разработать авторские профили на основе N-грамм, создав коллекцию известных авторов, написать компьютерную программу, которая автоматизирует процедуру определения авторства текстов. Апробация разработанной компьютерной программы определения авторства показала положительный результат на таком наборе данных: N = 4; L = 700; 180000 < Size < 590000. С помощью метрики L1 на несбалансированных наборах данных, которые содержали достаточно длинные тексты, точность определения авторства составила 83–98 %. Проведенное исследование показало, что L1 норма может быть успешно использована для решения задачи атрибуции текстов на основе подхода N-грамм.

Список литературы

1. Лингвистическая безопасность речевой коммуникации // ГЛЭДИС. 2004 [Электронный ресурс]. URL: http://www.rusexpert.ru/magazine/034.htm (дата обращения: 22.11.2018).

2. Виноградов В.В. Лингвистические основы научной критики текста // Вопросы языкознания. 1958. № 2. С. 3.

3. Батура Т.В. Формальные методы определения авторства текстов // Вестник НГУ. Серия: Информационные технологии. 2012. Т. 10. вып. 4. С. 81–94.

4. Diurdeva P., Mikhailova E., Shalymov D. Writer identification based on letter frequency distribution. Open Innovations Association (FRUCT), 2016 19th Conference of. IEEE, 2016. P. 24–30.

5. Stamatatos E.A. Survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology. 2009. Т. 60. № 3. P. 538–556.

6. Борисов Л.А., Орлов Ю.Н., Осминин К.П. Идентификация автора текста по распределению частот буквосочетаний // Препринты Института прикладной математики им. М.В. Келдыша РАН. 2013. № 27. С. 27–26.

7. Суркова А.С. Идентификация авторства текстов на основе информационных портретов // Вестник Нижегородского университета им. Н.И. Лобачевского. 2014. № 3 (1). С. 145–149.