УДК 004.93.12

ОБЗОР ПОДХОДОВ К ОПРЕДЕЛЕНИЮ АВТОРСТВА СПОРНЫХ ТЕКСТОВ ФОРМАЛЬНЫМИ МЕТОДАМИ

Чащин С.В.
Белорусский национальный технический университет, Минск, Беларусь,
SergtoUn@gmail.com

Исторический обзор подходов, эффективность большей части которых была оспорена практической работой, приведена во многих работах, в том числе [1]. Вместе с тем, ввиду того, что подавляющее большинство этих работ написаны иностранными авторами на иностранных языках, представляется целесообразным дать краткое описание проб и ошибок, допущенных исследователями на пути решения задач атрибуции текста, а также выработанные, в результате, успешные подходы.

Научный подход к атрибуции текста был заложен в конце 19 века в работах Mendenhall [1] изучавшего произведения Марлоу, Бэкона и Шекспира, а также Mascol [2, 85; 3], исследовавшего новозаветные Евангелия. Оба автора пытались выявить некоторый уникальный признак, который позволил бы однозначно определить принадлежность анонимного или псевдонимного текста известному автору. Для решения задачи авторы строили график отношения длины слова и частоты появления слова в тексте. Вопросы об авторстве как исследованных Mendenhall произведений, так и исследованных Mascol Евангелий до сих пор остаются нерешёнными[1], однако эти работы заложили основу статистических методов в атрибуции текста.

До начала 60-х годов основной целью исследователей был поиск некоторого уникального признака, позволяющего однозначно идентифицировать автора. Так, в 1944 году Yule [1] в качестве подобного маркера предложил длину предложения, хотя и указывал на невысокую надёжность данного метода. Предложенная им характеристика K является мерой словарного запаса автора текста, и не позволяет однозначно идентифицировать самого автора.

Новый подход к установлению автора текста был предложен Mosteller и Wallace [1] в 1964 году. Применив инновационный на тот момент наивный байесовский классификатор и используя в качестве признаков служебные слова, авторы смогли с высокой точностью (до 88%) идентифицировать авторов каждого из т.н. Федералистских текстов – псевдонимных статей, касающихся американской конституции, опубликованных в 1787-1788 гг. тремя разными авторами (Джей, Гамильтон и Мэдисон) под псевдонимом «Publius» («Публий»). Данная работа стала классическим примером применения количественных методов в анализе текстов и первой работой по применению многомерного анализа текста.

В 1971 году методом Mosteller и Wallace удалось доказать (вкупе с другими доказательствами) невиновность осуждённого за убийство полицейского и повешенного в 1953 году Дерека Бентли (реабилитирован в 1993 году). Среди главных доводов в пользу его невиновности были относительные частоты употребления слова then (десятикратно превосходила среднюю по Корпусу записей разговорного английского языка Банка Англии) и словосочетания I then (почти тысячекратно превосходила среднюю по указанному выше Корпусу) в документах, предъявленных суду в 1952 году в качестве признательных показаний Бентли. При этом, количественные значения и того, и другого

---

[1] В 1987-1990 гг. Пенн Лири, применяя средства компьютерной техники, провёл исследование, как он указывал, «шифров в текстах произведений Шекспира» (опубликовано в [4]), по результатам которого был сделан вывод, что значительная часть драм Шекспира была написана Бэконом. В последующем, использованный Пенном метод был раскритикован и отвергнут, т.к., согласно методике, Бэкон был также автором, как минимум, «Моби Дика», «Галльской войны» и Библии [5, с.565]. Спор об авторстве отдельных произведений, приписываемых Шекспиру, до сих пор не окончен.

параметров соответствовали частотам их употребления в показаниях полицейских. Более того, именно такие значения частоты встречаемости этих признаков были определены экспертом-лингвистом как уникальные признаки письменной речи сотрудников полиции.

Схожие подходы были предложены Burrows [1], применившим метод главных компонент к частотам слов. Идея заключается в проецировании многомерных векторов частот слов в сравнительных текстах в двумерное пространство. В результате, при чётком визуальном разделении спроецированных векторов разных авторов автор неизвестного текста определяется как тот, к чьим текстам ближе всего расположен спорный документ.

Скандальную известность в 90-е годы XX века получил метод QSUM (иначе называемый CUSUM), основанный на применении популярного в управлении качеством инструмента кумулятивных сумм. Главная идея подхода состоит в том, что отношения между отдельными авторскими инвариантами – например, длина предложения и количество служебных слов в тексте – являются постоянными на всё протяжении текста, и могут быть достаточной характеристикой автора текста. Для каждой статистики рассчитывается среднее значение, а отклонения от средних значений по каждому предложению суммируются накопительным итогом. Для проверки неизвестного текста, который, как правило, по объёму намного меньше текста с достоверным авторством, неизвестный текст вставляется в корпус известных. Изменения в графике кумулятивных сумм показывают, что неизвестный текст не относится к автору известного текста.

Простота данного метода обусловила его принятие в США в качестве стандарта в судах. Однако тщательные исследования показали, что авторы метода подгоняли исходные данные, а последующие исследования и расследования привели к отмене более 70% решений, принятых в американских судах на основе данного метода.

Изучение динамики служебных слов для определения плагиата проводилось и в СССР. Одной из наиболее известных является работа В.П.Фоменко и Т.Г.Фоменко, опубликованных в 1995 году А.Т.Фоменко в качестве приложения к работе по новой хронологии истории [6]. Использованный авторами метод близок к QSUM. Проанализировав ряд параметров (количество слов в предложении, процент служебных слов в тексте, количество слогов в слове, и др.) авторы предлагают в качестве дифференцирующего признака процент служебных слов в выборках текста. В самой работе не даётся точный количественный критерий, на основании которого можно делать вывод о принадлежности произведения данному автору. Как предлагают авторы, «Если для двух исследуемых произведений значения параметра 3 (процент служебных слов) разнятся больше, чем на единицу, то есть основания заподозрить различное авторство сравниваемых текстов». Чем больше разница инварианта, тем подозрения серьёзнее» [6]. Более того, как признаются сами авторы, метод может использоваться только на текстах крупного объёма. Определить что понималось авторами под «большим», «крупным» и «малым» объёмами из текста данной работы не удалось. Для целей дальнейшего исследования отметим следующее. Авторами брались выборки не менее 16.000 слов, шаг между ними составлял не менее 60 страниц, и в исследовании использовались только прозаические произведения. Применение несопоставимых единиц измерения объёмов произведения (количество страниц и слов) не позволяет определить минимальный размер произведения, который может анализироваться. В качестве признаков использовались отдельные предлоги, союзы и частицы (всего 54 служебных слова). Причины выборки именно данных служебных слов в исследовании не раскрываются. Средняя величина используемого параметра для различных авторов колеблется от 19,4% до 27,5%.

Данный подход применен авторами при исследовании проблемы авторства «Тихого Дона». На наш взгляд, выводы по результатам исследования (книги 1 и 2, части 1-5 и начало части 6 книги 3 «Тихого Дона» принадлежат перу Ф.Крюкова) не в полной мере согласуются с результатами исследования. Действительно, значения параметра в

произведениях Ф.Крюкова (21,11%[2]) ближе значениям параметра в книгах 1 и 2 «Тихого Дона» (19,55%) по сравнению с произведениями в раннего М.Шолохова (22,46%). Однако разница между значениями спорной части «Тихого Дона» и работами Ф.Крюкова (1,56% = 21,11 – 19,55) даже больше, чем между работами Ф.Крюкова и ранними произведениями М.Шолохова (1,35% = 22,46 – 21,11), что, исходя из предлагаемой авторами методики, позволяет также сомневаться и в авторстве Ф.Крюкова.

Ещё одно направление в применении математических моделей при атрибуции текста было положено Д.Хмелевым [7] в 90-х годах прошлого века. В ходе проведенных исследований был сделан вывод о том, что простейший подход с использованием цепей Маркова первого порядка показывает хорошие результаты на файлах большого объема и плохие по сравнению с другими методами на отрывках длиной в 2000 - 5000 символов. Этот метод был реализован в системе «Лингвоанализатор» (http://www.rusf.ru/books/analysis), – первого русскоязычного эксперимента по распознаванию автора текста. Детально данная система, а также проводимые с ней эксперименты описаны в работах [7; 8].

«Лингвоанализатор» представляет собой ресурс в глобальной сети Интернет. Основная задача системы – определить степень близости заданных пользователем исследуемых текстов индивидуальному стилю одного из авторских эталонов, определённых заранее и занесенных в систему.

В работе [9] показано, что последовательность символов текста не обладает свойствами простой цепи Маркова. Вместе с тем, моделирование языка с применением цепей Маркова стало крайне популярным в последнее время среди авторов, занимающихся исследованиями в области компьютерной лингвистики на различных языках, в т.ч. английском (см., например, [10, с.192-196]), арабском [11, с.474] и китайском ([12]) языках. Более того, цепи Маркова активно используются в настоящее время в системах распознавания образов, в частности, звучащей речи [13; 14, с.9-15] и текста [15].

Как продолжение развития подхода, использующего в качестве стилевых признаков бинарные буквосочетания, А. Н. Тимашев [16] предложил применять трехбуквенные сочетания. Был создан программный продукт для автоматического сравнения и классификации текстов по параметрам индивидуального авторского стиля под названием «Атрибутор» (электронный адрес программного продукта – http://www.textology.ru/web.htm).

База этой программы содержит произведения 103 авторов и использует экспертную обработку текстов. В эталонную выборку, на которой происходило обучение «Атрибутора», попали в основном романы и повести отечественных писателей XIX-XX вв. Пополнение шло за счет ресурсов известных электронных библиотек, наибольшее количество текстов было получено в библиотеке М. Мошкова. Выборка подбиралась таким образом, чтобы тексты разных писателей в максимальной степени различались друг от друга, а тексты одного писателя были максимально близки. Те случаи, когда известный писатель в какой-то период своего творчества резко менял стиль изложения, отсеивались.

Для обработки текста «Атрибутором» необходимо, чтобы его длина была не меньше 6 страниц. Ограничение на длину текста накладывается для того, чтобы избежать ошибок, связанных со сравнением статистически несопоставимых объектов. В обработку попадают все слова текста, за исключением имен собственных.

О.Шевелев в своём диссертационном исследовании [9] проводил анализ с использованием гипергеометрического критерия (двустороннего точного критерия Фишера) и критерия хи-квадрат по отдельным частотным признакам текстов, совокупности признаков, а также по их распределению. Автором также проведено сопоставление подхода Д.Хмелева и его модификаций, авторских вариантов метода Д.Хмелева, деревьев решений и нейронных сетей. В результате делается вывод о наименьшей эффективности метода деревьев решений для задач классификации текстов, а модифицированный метод

[2] Значения взяты из работы авторов и не перепроверялись.

Д.Хмелева и метод нейронных сетей в проведенном исследовании дал примерно равные результаты, и, в ряде случаев, достигал 100%. Согласно выводам О.Шевелева, критическое значение объема имеет порядок 30000-40000 символов или 5000-6000 слов, или 400-600 предложений. В данной работе основной акцент делается на методах автоматической кластеризации текстов. В своём диссертационном исследовании автор использовал три группы признаков: частоты появления пар букв, частоты появления 100 самых часто встречаемых словоформ из частотного словаря Шарова, частоты появления предложений с определенным числом слов. В [32] используется иной набор признаков – служебные слова по списку В.П.Фоменко и Т.Г.Фоменко.

В работе А.С.Романова [18] для атрибуции текста используется метод опорных векторов. Автор провёл обширное исследование с применением различных признаков: униграмм (определение n-грамм – униграмм, биграмм, триграмм и др. - будет дано ниже, см.п.1.3.2), биграмм, триграмм, 500 наиболее часто используемых биграмм, 500 наиболее часто используемых триграмм, контекстных слов, знаков пунктуации, частей речи, длины слова и др. В ходе исследования были сопоставлены такие параметры, как скорость обучения и точность построенных моделей в зависимости от архитектуры и объёмов выборки. Результаты классификации с использованием машин опорных векторов сравнивались с результатами классификации, полученными с использованием нейронных сетей - многослойного перцептрона и сетей каскадной корреляции. Автором сделан вывод о том, что метод опорных векторов является наиболее точным и наименее затратным по времени методом классификации.

К выводу о высокой результативности метода опорных векторов пришли также авторы зарубежного исследования [19]. Особенность исследования состоит в том, что в качестве исходных были избраны не художественные тексты, а статьи семи различных авторов в ежедневной газете Berliner Zeitung за трёхмесячный период (декабрь 1998 - февраль 1999 гг.). В качестве признаков использовались словоформы (т.е. обладающая признаками слова цепочка фонем, формально отличающаяся от другой), длина слов, метки частей речи (part-of-speech tags, POS-tags), комбинации служебных слов с метками частей речи, а также различные преобразования исходных признаков и различные ядра (линейное, квадратичное, кубическое и радиальной базисной функции). Исследование проводилось на текстах на немецком языке. В ходе исследования сравнивались результаты работы машины опорных векторов, деревьев решений и многослойного перцептрона, прецизионность метода опорных векторов оказалась наивысшей (по 5 авторам составила 100% вне зависимости от применяемых признаков их комбинаций). Специфичность колебалась в интервале от 21% до 92%.

Метод опорных векторов также применялся E.Stamatatos при исследовании текстов из ежедневной газеты Guardian [20]. В качестве признаков использовались слова и символьные триграммы. Автор провёл анализ возможности применения данных признаков на различных тематических (политика, общество, события в мире, события в Великобритании) и жанровых (статьи и рецензии на книги) выборках. В отличие от предыдущих авторов, E.Stamatatos был проведен ряд экспериментов по обучению модели на одной тематической и/или жанровой выборке, а тестированию – на другой. Максимальные результаты получены при использовании символьных триграмм. При обучении модели и её тестировании на одной и той же тематике на 6.000 наиболее часто встречающихся триграммах получена 100% точность. При обучении и тестировании на разных тематических выборках максимальная точность получена при 3.500 триграмм, при этом точность повышалась с ростом числа признаков до данной отметки, и далее резко сокращалась. Схожим образом была достигнута максимальна точность 77% при использовании 1.500 наиболее часто встречающихся слов. Высокая точность метода опорных векторов вместе с символьными триграммами проявилась даже при обучении и тестировании на выборках текстов разных жанров. Например, при обучении на выборке

текстов из темы «Политика» и тестировании на жанре «Рецензии на книги» точность достигала 80% (при количестве используемых признаков 9.000).

По результатам исследования более ранних работ, Koppel, Schler, Argamon [21] также делают вывод о том, что точность метода опорных векторов, как минимум, не ниже иных методов классификации.

В последние годы в компьютерной лингвистике всё чаще используются нейронные сети. Из публикаций наибольший интерес для целей нашего исследования представляет работа [22], в которой в качестве признаков анализируются различные словесные и символьные n-граммы, а в качестве метода применяется свёрточная нейронная сеть. В ходе исследования определялись авторы твитов в сети микрособщений Twitter. Изучались различные корпусы текстов с количеством авторов от 50 (после фильтрации ботов осталось 35 авторов) до 1000, с количеством твитов каждого автора от 50 до 1000. Максимальная точность – 76,1% – получена, как и следовало ожидать, на минимальном наборе авторов (50) при максимальном объёме твитов (1000).

Список литературы:

1. Juola, P. Authorship Attribution / P.Juola // Foundations and Trends in Information Retrieval. – Hanover, 2006. – №1(3),. С.233–334
2. Argamon S. The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning / S.Argamon [и др.] – Berlin: Springer-Verlag, 2010. – 339 с.
3. Rudman J. The State of Authorship Attribution Studies: Some Problems and Solutions / J.Rudman // Computers and the Humanities / Amsterdam: Kluwer Academic Publishers, 1998. – Vol.31. – С.351-365
4. Leary P. The Cryptographic Shakespeare [Электронный ресурс]. – Режим доступа: https://www.baconscipher.com/. – Дата доступа: 22.10.2020
5. Handbook of Natural Language Processing / R.Dale [и др.] – New York: Marcel Decker, Inc, 2000. – 999 с.
6. Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Приложение Кто был автором «Тихого Дона?» [Электронный ресурс]. – Режим доступа: http://chronologia.org/seven2_2/add3_2.html. – Дата доступа: 04.09.2020
7. Хмелёв, Д.В. Распознавание автора текста с использованием цепей А.А.Маркова / Д.В.Хмелёв / Вестник МГУ, сер.9: Филология // под ред.М.Л.Ремнёвой [Электронный ресурс]. – Москва, 2000. – №2. – С.115-126. – режим доступа: http://www.philol.msu.ru/~lex/khmelev/published/vestnik/vestnik2000win.html. – Дата доступа: 02.09.2020
8. Батура, Т.В. Формальные методы определения авторства текстов / Т.В.Батура // Вестник НГУ. Серия: Информационные технологии / под ред. А.М.Федотова. – Новосибирск, 2012. – Том 10, вып. 4. – С.81-94
9. Шевелёв О. Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. ... канд. техн. наук : 05.13.18 / О.Г.Шевелёв. – Томский гос. ун-т. – Томск, 2006. – 19 с.
10. Manning, Ch.D., Schütze, H. Foundations of Statistical Natural Language Processing / Cambridge: Massachusetts Institute of Technology, Second printing with corrections, 2000. – 704 с.
11. Altheneyan, A.S., Menai, Mohamed El Bachir. Naïve Bayes classifiers for authorship attribution of Arabic texts / Journal of King Saud University – Computer and Information Sciences// Ed. Ahmad Al Ghamdi. – Riyadh, 2014. – #26. – С.473-484
12. Jian-yun, N. On the Use of Words and N-gram for Chinese Information Retrieval / N.Jian-yun // Proceedings of the fifth international workshop on Information retrieval with Asian languages. – New York, USA: 2000. С.141-148

13. Rabiner L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition / L.R.Rabiner// Proceedings of the IEEE – Raleigh, North Carolina, USA, 1989. – vol.77, №2. – C.257-286
14. Fink Gernot A. Markov Models for Pattern Recognition: From Theory to Applications. Second Edition /Gernot A.Fink/ Springer-Verlag, London: 2008/2014. – 275 c.
15. Keselj V. N-gram-based Author Profiles For Authorship Attribution / V.Keselj [и др.]// PACLING'2003 / Pacific Association for Computational Linguistics. – Halifax, Canada: 2003. – C.255-264
16. Тимашев А. Н. Атрибутор // Текстология, ru [Электронный ресурс]. – 2002. – Режим доступа: http://www.textology.ru/atr_resum.html
17. Шевелев О.Г., Макаров А.Г., Поликарпов А.А., Поддубный В.В. Анализ количественных характеристик авторского стиля романа «Тихий Дон» и его соотношение с другими текстами М.А.Шолохова на основе иерархической кластеризации / Русский язык: исторические судьбы и современность. IV Международный конгресс исследователей русского языка. Москва, МГУ им.М.В.Ломоносова. Филологический факультет. 20-23 марта 2010 года. Труды и материалы. Сост.: М.Л.Ремнёва, А.А.Поликарпов. – М.:Изд-во МГУ, 2010. – с.523-524
18. Романов А. С. Методика и программный комплекс для идентификации автора неизвестного текста: автореф. дис. ... канд. техн. наук: 05.13.18 / А.С.Романов. – ТУСУР. – Томск, 2010. – 26 с.
19. Diederich J., Rindermann J., Leopold E., Paass, G. Authorship Attribution with Support Vector Machines / Applied Intelligence, Vol.19. Amsterdam: Kluwer Academic Publishers, 2003. -- с.109-103
20. Stamatatos E. On the robustness of authorship attribution based on character n-gram features / Journal of Law and Policy, vol.21(2). – Denver: Joh Wiley & Sons, 2013. – с.421-439
21. Koppel M., Schler J., Argamon S. Computational Methods in Authorship Attribution / Journal of the American Society for Information Science and Technology, vol.60(1). – Hoboken: Wiley-Blackwell, 2009. – с.9-26.
22. Shrestha P. и др. Convolutional Neural Networks for Authorship Atribution of Short Texts // P.Shrestha и др. / Proceedings of the 15-th Conference of the European Chapter of the Assosication of Computational Linguistics. – Valencia, April 3-7, 2017. – Vol. 2, Short Papers. – стр.669-674