Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

Бурак
Александр Анатольевич

Определение автора текста

АВТОРЕФЕРАТ
на соискателя академической степени
магистра технических наук

по специальности 1 - 40 80 02 Системный анализ, управление и обработка информации

Научный руководитель
Герман О.В.
к.т.н.

Библиотека БГУИР

Минск 2016

КРАТКОЕ ВВЕДЕНИЕ

Задачи определения автора текста актуальны на сегодняшний день. Огромное количество текстовых данных создается каждую секунду в каждом уголке мира. Вопрос определения автора литературного текста присутствуют в области интересов как правозащитников, литературоведов, службы безопасности стран, так и математиков, и программистов. Огромный интерес вызывает исследование самого творческого процесса написания текста авторами, если рассматривать этот процесс в качестве алгоритма.

Каждый человек, который прочел за свою жизнь большое количество книг, замечал интересный факт о том, что любой писатель обладает уникальным стилем, индивидуальными чертами языка. Мы с легкостью может отличить произведения известных нам авторов по характерному стилю написания произведения, не подозревая как это у нас получается. Каждый автор употребляет свойственные ему части речи, стилистические обороты, определенные слова для описания конкретных ситуаций.

Все эти различия в текстах проявляются автоматически, неосознанно, под внешними и внутренними лингвистическими влияниями на автора. Автор в своих произведениях придерживается определенной уникальной манеры письма, которая использовать статистические методы для определения принадлежащих ему текстов. И если найти в чем заключается эта уникальность для каждого автора (его речевой портрет), то решается проблема идентификации автора текста.

Предполагая, что каждый авторский текст уникален, мы сможем используя статистику определить стилистику автора, манеры письма для каждого автора на основе частот повторения определенных символов, которые использовал писатель при создании произведения. Для идентификации автора произведений использовался метод близости частот встречаемости букв, пар букв (диграмм), триграмм (три буквы), слов русского языка. Предположение заключается в уникальности частоты использование определенных символов, слов определенным автором.

Библиотека БГУИР

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования.

Целью настоящей диссертационной работы применение статистических методов в сравнительных исследованиях текстов для идентификации автора текста.

Достижение данной цели потребовало решения следующих задач:

1. Изучить существующие стандарты и подходы для идентификации автора текста.
2. Изучить синтаксис и семантику существующих алгоритмов и программных продуктов сравнительного анализа текста.
3. Спроектировать и разработать библиотеку для идентификации автора текста на больших и малых фрагментах текста.
4. Сравнить показатели производительности разработанной библиотеки с существующими аналогами.

Объектом исследования является алгоритм идентификации автора текста. Предметом исследования являются методы работы методы работы и алгоритмы статистического анализа текста.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии научно-техническими заданиями и планами работ кафедры «Информационных технологий автоматизированных систем» по теме «Идентификация автора текста».

Апробация результатов диссертации

Основные положения диссертации были представлены на международной научно-практической конференции «Наука и образование третьего тысячелетия» (Москва, Россия, 2015).

Личный вклад магистранта

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя О. В. Герман, заключается в формулировке целей и задач исследования.

Библиотека БГУИР

Опубликованность результатов диссертации

По теме диссертации опубликована работа в сборнике трудов и материалов международной конференции "Наука и образование третьего тысячелетия" в журнале "Альманах мировой науки" (ISSN 2412-8597, DOI постатейно, РИНЦ) - свидетельство о регистрации СМИ ЭЛ № ФС 77 - 63156 от 01.10.2015 г

Актуальность темы

Развитие технологий в области обработки текстовой информации происходит очень быстрыми темпами.

При решении задачи установления авторства текста (или задачи атрибуции) неизбежно приходится обращаться к экспертам. Эксперты могут идентифицировать автора неизвестного текста или определить принадлежность произведения другому автору при помощи характерных языковых особенностей и стилистических приемов. Несомненно, подобные исследования трудоемки, однако задача установления авторства текстов возникает в различных областях и представляет интерес для филологов, литературоведов, юристов, криминалистов, историков. Поэтому встает вопрос о создании формальных методов ее решения. В настоящее время для атрибуции текстов применяются подходы из теории распознавания образов, математической статистики и теории вероятностей, алгоритмы нейронных сетей и кластерного анализа и многие другие.

С развитием вычислительной техники появилась возможность реализовать методы, требующие огромных вычислений, чтобы облегчить работу экспертов. Существующие программные продукты позволяют учитывать и варьировать различные лингвостатистические параметры, разносторонне характеризующие текст. В статье приведен обзор различных формальных методов определения авторского стиля, предпринята попытка выявить их особенности и недостатки, сравнить программные продукты по атрибуции текстов, ориентированные на русский язык.

Попытка использовать нейронные сети для определения авторства текста показало отрицательный результат в плане затрат процессорного времени для получения результата. По этой причине большинство исследований в этой

Библиотека БГУИР

области перешло на статистические методы и алгоритмы идентификации текста, которые будут расмотрены в представленной работе.

Структура и объем диссертации

Работа состоит из следующих разделов: введение, обзор алгоритмов идентификации автора текста, исследование проблем производительности и способов оптимизации алгоритмов идентификации автора, проектирование и разработка высокопроизводительной библиотеки для идентификации автора текста. Диссертация изложена на 60 страницах машинописного текста, библиография включает в себя список использованной литературы, а также публикации автора.

Рекомендации по практическому применению результатов

1. Полученные результаты формируют теоретическую и практическую базу для задач задач идентификации автора текста.
2. Разработанный алгоритм и программный продукт может использоваться для получения вероятностных результатов идентификации текста.

Библиотека БГУИР

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы, сформулированы цель и задачи, определены объект и предмет исследования. Изложены методологические и теоретические основы диссертационного исследования, обоснованы его научная новизна и практическая значимость.

В первой главе - «Обзор алгоритмов идентификации автора текста**»** - дан обзор развития подходов распознавания автора текста, в частности охарактеризованы изменения, которые произошли с развитием вычислительных мощностей. Выделены их основные преимущества и недостатки.

Во второй главе - «Исследование проблем производительности и способов оптимизации алгоритмов расспознования автора текста » - были рассмотрены способы оптимизации существующих подходов.

В третьей главе – «Проектирование и разработка высокопроизводительной библиотеки для идентификации автора текста» производится разработка высокопроизводительной библиотеки и веб-сайта для идентификации автора текста.

Библиотека БГУИР

ЗАКЛЮЧЕНИЕ

Основным результатом данной работы является то, что использование грамматической информации в решении задачи определения действительного автора текста является не только осмысленным, но и достаточно эффективным, а в некоторых отношениях сопоставимым с использованием информации о встречаемости пар букв в тексте, как это было показано ранее в исследовании.

В данной работе для исследования уникальности частоты символов были взяты наиболее частые слова русского языка. Анализ уникальности частот произведен на основе 1000, 2000, 5000 наиболее частых русских слов, покрывающих по статистике 64.0708%, 76.5104%, 82.0604% текста соответственно. В работе сравнивается точность результатов, а также производительность (время идентификации автора текста) в зависимости от выбора символов для анализа частот и размеров текстов. Проведена проверка алгоритма на корректность работы на более чем 100 текстов русских авторов. В дальнейшем планируется сотрудничество с электронными библиотеками литературных источников для повышения количества текстов.

Перспективность использования грамматической информации в данной работе показана также и тем, что использование информации об одиночных обобщенных грамматических классах оказалось заметно более эффективным, чем использование информации об одиночных буквах.

То, что полученные в ходе исследования результаты с использованием различных единиц (букв и обобщенных грамматических классов) не противоречат друг другу, позволяет предположить, что в будущих развитых методиках определения авторства текста будут использоваться различные отображения текста, полученные на основе этих единиц для взаимной перепроверки результатов.

Полученные результаты свидетельствуют о перспективности использования данного метода для определения автора текста с определенной точностью и рациональной производительностью программного выполнения. Наибольшая точность достигается при использовании 5000 наиболее частых лем для текстов, превышающих 10 тысяч знаков.

Основные научные результаты диссертации.

1. Выявлены способы повышения производительности идентификации автора текста.
2. Разработана кроссбраузерная и кроссплатформенная библиотека и веб-система для распознавания автора.

Библиотека БГУИР

Рекомендации по практическому использованию результатов.

Разработанная библиотека может быть использована при написании простых и сложных приложений, для идентификации автора текста.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Бурак, А. А. Распознавание автора текста [Текст] / Бурак.А. А // Наука, образование, общество: тенденции и перспективы: материалы междунар. науч.–практ. конф. (Москва, 30 ноября 2015 г.). – Москва: ООО "АР-Консалт", 2015 N 2-1(2). – С. 93-94.

Библиотека БГУИР