

Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Базовая кафедра вычислительных и информационных технологий

УТВЕРЖДАЮ

Заведующий кафедрой

_____ / В.В. Шайдуров

«___» _____ 2019г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

ИССЛЕДОВАНИЕ МЕТОДОВ РЕШЕНИЯ ЗАДАЧИ ИДЕНТИФИКАЦИИ АВТОРСТВА ТЕКСТОВ

Направление 02.04.01 Математика и компьютерные науки

Магистерская программа 02.04.01.02 Вычислительная математика

Научный руководитель
кандидат физико-математических наук,
доцент

_____ / И.В. Баранова

Выпускник

_____ / А.В. Брестер

Красноярск 2019

АННОТАЦИЯ

Целью работы является исследование задачи идентификации авторства текста и основных методов ее решения.

Для решения задачи были выбраны метод идентификации авторства текста по распределению частот буквосочетаний, метод Хмелева, частотный анализ текста и метод «мешка слов».

В результате работы было разработано программное обеспечение, реализующее работу выбранных методов, произведено исследование алгоритмов по точности их работы, решена практическая задача идентификации авторства текстов и выполнен анализ полученных результатов.

Ключевые слова: кластерный анализ, методы идентификации текстов, метод Хмелева, частотный анализ, мешок слов.

ANNOTATION

The purpose of the work is to study the problem of identification of authorship of the text and the main methods of its solution.

For the solution of the problem the method of identification of authorship of the text by distribution of frequencies of letter combinations, the method of Hmelev, frequency analysis of the text and the method of "bag of words" were chosen.

As a result of the work, the software implementing the work of the chosen methods was developed, the study of algorithms on the accuracy of their work was made, the practical task of identification of authorship of texts was solved and the analysis of the obtained results was performed.

Key words: cluster analysis, methods of text identification, hop method, frequency analysis, bag of words.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
Объект исследования	4
Предмет исследования.....	4
Цель работы	5
Теоретическая и практическая ценность	5
1.Задача идентификации авторства текстов	6
1.1 Постановка задачи классификации текстов	6
1.2 Постановка задачи идентификации авторства текстов.....	7
2. Обзор формальных методов идентификации авторства текста	10
3. Основные методы решения задачи.....	13
3.1 Идентификация авторства по распределению частот буквосочетаний .	13
3.2. Метод Хмелева	15
3.3. Частотный анализ текста.....	16
3.3. Метод «мешка слов».....	17
4.Решение задачи идентификации авторства текстов	19
4.4 Вычислительный эксперимент	22
4.5 Сравнение представленных алгоритмов.....	26
5.Описание программного модуля	30
5.1 Программный модуль методов распределения частот буквосочетаний, Хмелева и частотного анализа текста	31
5.2 Программный модуль метода «мешка слов»	32
ЗАКЛЮЧЕНИЕ	34
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	35
ПРИЛОЖЕНИЕ А	37

ВВЕДЕНИЕ

В настоящее время в связи с повсеместным внедрением информационных технологий и интернета значительно увеличилась доля электронных документов и электронного документооборота. Многие предприятия, учреждения, частные компании, создатели художественных и научных работ используют в своей работе электронные текстовые документы, причём авторы данных произведений могут быть неизвестны или указаны неверно, в связи с чем становится актуальной задача идентификации авторства текстов. Также данная задача является важной в исторических, лингвистических и криминалистических исследованиях.

Общие методы решения задачи идентификации авторства текста разделяют на два вида: *экспертные* и *формальные*. *Экспертные методы* требуют непосредственного участия эксперта – сведущего лица, обладающего специальными познаниями в области науки, искусства, техники и ремесла, а именно: знания об условиях и закономерностях речевого поведения человека, обуславливающих индивидуальность, динамическую устойчивость. *Формальные методы* основаны на представлении текста, после предварительной обработки, в числовом виде и сравнении полученных числовых характеристик. В данных методах используются приемы из теории вероятностей, математической статистики, машинного обучения и кластерного анализа. В данной работе будут рассматриваться только основные *формальные* методы решения задачи идентификации авторства текста.

Перед применением алгоритмов каждый из текстов должен быть подвергнут предварительной обработке, это необходимо для ускорения и улучшения работы методов.

В первой главе работы приводятся основные понятия и постановка задачи идентификации авторства текста. Рассматриваются алгоритмы предварительной обработки текста.

Во второй главе работы рассматриваются формальные методы решения задачи идентификации авторства текста. Дается их описание и рассматриваются основные принципы их работы

В третьей, четвертой, пятой и шестой главах работы рассматриваются алгоритмы работы выбранных методов: метода идентификации авторства текста по распределению частот буквосочетаний, метода Хмелева, метода частотного анализа и метода «мешка слов». Дается подробное описание данных алгоритмов, и указываются особенности их работы.

В седьмой главе работы решается практическая задача идентификации авторства текстов, проводится сравнение точности работы предложенных алгоритмов. Выполняется серия численных экспериментов, позволяющих оценить зависимость точности идентификации от объема идентифицируемого текста,

Применяются различные меры расстояний в работе метода «мешка слов» и сравнивается качество идентификации для каждой меры.

В последней главе диссертации приводится описание разработанного программного модуля, и демонстрируются окна каждого из рассмотренных методов.

Объект исследования

В качестве объекта исследования в представленной диссертации выступают методы решения задачи идентификации авторства текстов.

Предмет исследования

Предметом исследования в данной работе выступает задача идентификации авторства текстов.

Цель работы

Целью магистерской диссертации является исследование основных методов решения задачи идентификации авторства текстов.

Для достижения указанной цели в диссертационной работе были поставлены и решены следующие задачи:

- изучить постановку задачи идентификации авторства текстов;
- изучить основные алгоритмы решения задачи идентификации авторства текстов: идентификация по распределению частот буквосочетаний, метод Хмелева, частотный анализ текста и метод «мешка слов»;
- применить различные меры расстояния в работе метода «мешка слов»;
- разработать программное обеспечение, реализующее работу перечисленных методов;
- решить практическую задачу идентификации авторства текстов;
- выполнить сравнительный анализ полученных результатов.

Теоретическая и практическая ценность

Решение данной задачи может использоваться в ряде криминалистических экспертиз, в лингвистических исследованиях произведений с анонимным авторством, в исторических исследованиях различных документов и других задач анализа текстовых документов.

Задача идентификации авторства текстов является важной частью в решении задач классификации документов, например, в задачах автоматической каталогизации электронных документов, так как ручная обработка этих текстов имеет недостатки в виде больших временных затрат и риска ошибки по причине человеческого фактора.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Воронцов, К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования / К. В. Воронцов. – Москва: МГУ, 2007. – 18 с.
2. Дуда, Р. Распознавание образов и анализ сцен: пер. с англ. Г. Г. Вайнштейна, А. М. Васьковского / Р. Дуда, П. Харт; под ред. В.Л. Стефанюка. – Москва: Мир, 1976. – 502 с.
3. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. – 270 с.
4. Зенин, А. В. Анализ методов распознавания образов / А.В. Зенин // Молодой ученый, 2017. – №16. – С. 125-130.
5. Борисов, Л.А. Идентификация автора текста по распределению частот буквосочетаний / Л.А Борисов, Ю.Н. Орлов, К.П. Осминин // Прикладная информатика, 2013. – Т. 26, № 2. – С. 95-108.
6. Орлов, Ю.Н. Определение жанра и автора литературного произведения статистическими методами. / Ю.Н. Орлов, К.П. Осминин // Препринты ИПМ им.М.В.Келдыша, 2010. – № 27. – 26 с.
7. Орлов, Ю.Н. Методы статистического анализа литературных текстов / Ю.Н. Орлов, К.П. Осминин. – Москва: ЭдиториалУРСС / Книжный дом «ЛИБРОКОМ», 2012. – 312 с.
8. Батура, Т.В. Формальные методы определения авторства текстов / Т.В. Батура // Вестник НГУ. Серия Информационные технологии. – 2012. – Т. 10, № 4. – С. 23-28
9. Романов, А. С. Методика и программный комплекс для идентификации автора неизвестного текста: автореф. дис. ... канд. техн. наук : 05.13.18 / Романов Александр Сергеевич. – Томск, 2010. – 26 с.
10. Программный комплекс СМАЛТ / А. А. Рогов, Г. Б. Гурин, А. А. Котов, Ю. В. Сидоров, Т. Г. Суровцова // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды X Всерос. науч. конф. «RCDL'2008». – Дубна, 2008. – С. 155–160.

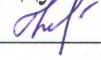
11. Марков, А. А. Об одном применении статистического метода / А.А. Марков // Известия Императорской Академии наук. Сер. 6. – 1916. –Т. 10, № 4. – С. 239–242.
12. Фоменко, В. П. Авторский инвариант русских литературных текстов / В. П. Фоменко, Т. Г. Фоменко. – М.: МГУ, 1995. – 422 с.
13. Хмелёв, Д. В. Распознавание автора текста с использованием цепей А. А. Маркова / Д.В. Хмелёв // Вестник МГУ. Сер. 9: Филология. – 2000. – № 2. С. 115–126.
14. Хмелёв, Д. В. Классификация и разметка текстов с использованием методов сжатия данных / Д.В. Хмелёв // Все о сжатии данных, изображений и видео. 2003. URL: <http://compression.ru/download/articles/classif/intro.html>
15. Кукушкина, О. В. Определение авторства текста с использованием буквенной и грамматической информации / О. В. Кукушкина, А. А. Поликарпов, Д. В. Хмелев // Проблемы передачи информации. – Москва: Наука, 2001. – Т. 37, № 2. – С. 96–108.
16. Шевелёв, О. Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. ... канд. техн. наук.: 05.13.18 / Шевелёв Олег Геннадьевич. – Томск, 2006. – 18 с.
17. Брестер, А.В. О задаче определения авторства текстов / А.В. Брестер // Электронный сборник материалов международной научно-технической конференции студентов, аспирантов и молодых учёных «Перспектив Свободный-2019». – Красноярск: СФУ, 2019.
18. Брестер, А.В. Methods for solving the problem of identifying the authorship of a text / А.В. Брестер // Электронный сборник материалов международной научно-технической конференции студентов, аспирантов и молодых учёных «Перспектив Свободный-2019». – Красноярск: СФУ, 2019.
19. Brester, A. Methods for solving the problem of identifying the authorship of a text / A. Brester // Proceedings of the XVII International FAMEMS Conf. and the III Workshop on Hilbert's Sixth Problem. – Krasnoyarsk: SFU, 2018. – pp. 111-114.

Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Базовая кафедра вычислительных и информационных технологий

УТВЕРЖДАЮ

Заведующий кафедрой

 / В.В. Шайдуров

«17» июня 2019г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

ИССЛЕДОВАНИЕ МЕТОДОВ РЕШЕНИЯ ЗАДАЧИ ИДЕНТИФИКАЦИИ АВТОРСТВА ТЕКСТОВ


Направление 02.04.01 Математика и компьютерные науки

Магистерская программа 02.04.01.02 Вычислительная математика

Научный руководитель
кандидат физико-математических наук,
доцент

 / И. В. Баранова

Выпускник

 17.06.19 / А. В. Брестер

Красноярск 2019