Отзыв директора Института математики им. А.Джураева Национальной академии наук Таджикистана о работе Косимова Абдунаби Абдурауфовича «Статистические закономерности распознавания однородности текстов с помощью γ-классификатора (Қонуниятҳои омории шинохти якҷинсагии матн бо истифода аз γ-таснифгар)», представляемой на соискание учёной степени доктора технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Проблема распознавания текста возникла одновременно с появлением письменности. Длительное время она ограничивалась, в основном, определением личности писаря по своеобразию его почерка. В дальнейшем, с наступлением эры книгопечатания, в проблеме актуализировалось новое направление – распознавания однородности печатной продукции, что в настоящее время и составило основное содержание всей проблемы.

Цель работы – алгоритмизировать процесс распознавания однородности текстов и реализовать его в виде компьютерного программного комплекса.

Задачи исследования. Для достижения цели решаются следующие задачи:

1) сформировать две электронные коллекции текстов, из которых первая предназначена для предварительного тестирования, а вторая – для оценки перспективности применения γ-классификатора;

2) исследовать цифрового портрета текста (ЦПТ) для распознавания автора текста;

3) установить статистическую эффективность применения γ-классификатора для распознавания авторов произведений;

4) определить минимальный размер незнакомого текста, пригодного для распознавания его автора;

5) исследовать эффективность применения высокочастотных элементов ЦПТ для идентификации автора текста;

6) установить статистическую эффективность применения γ-классификатора и исследования пригодности ЦП на основе распределения частотности различных алфавитных элементов текста для распознавания других признаков однородности, таких как тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений, шифры научных работ и т.д.;

7) исследовать статистические закономерности распознавания однородных текстов на корпусах художественных литературных произведений;

8) определить эффективность применения γ-классификатора для

атрибуции искусственно сгенерированных произведений авторов;

9) исследовать влияние порядка ЦП текста на распознавание однородности произведения с помощью γ-классификатора;

10) спроектировать и реализовать компьютерный программный комплекс для распознавания (идентификации) однородности текста на основе различных ЦП текста и γ-классификатора.

Объект исследования – корпус печатных текстов и его характеристики на разных языках.

Предмет исследования – распознавание однородности произведения на основе γ-классификатора (математической триады) и частотности различных характеристик текста.

Научная новизна диссертации состоит в следующем:

- исследована информативность нетрадиционных признаков на предмет количественного описания таджикских текстов;

- установлена статистическая эффективность π математической модели опознавания авторов произведений таджикской классической поэзии (π = 1.00) на основе триграмм, современной поэзии (π = 0.98) с помощью униграмм и современной прозы (π = 0.96) на основе распределения длин предложений (в словах);

- установлена 100%-ная статистическая эффективность путем применения метрического γ-классификатора и метода ближайшего (по расстоянию) соседа идентифицировать авторов произведений – убывающих по размерам последовательности текстовых фрагментов от величины в 7000 слов (40000 символов) вплоть до 20 слов (100 символов);

- для целей существенного сокращения объёма вычислительных процедур установлена возможность эффективного использования не всех, а только высокочастотных элементов ЦП текстов;

- установлена статистическая эффективность применения γ-классификатора и исследована пригодность ЦП на основе распределения частотности различных алфавитных элементов текста для распознавания других признаков однородности, таких как тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений и шифры научных работ;

- исследованы статистические закономерности опознавания авторов и языков произведений на корпусах художественных литературных произведений с помощью γ-классификатора;

- γ-классификатор и метод ближайшего соседа были протестированы на случайных выборках текстов, распознаются с достаточно высокой точностью признаки однородности произведений различных модельных коллекций и корпусов;

- установлена эффективность применения γ-классификатора для

атрибуции искусственно сгенерированных поэм «Шахнаме» А. Фирдоуси по обучению рекуррентных нейронных сетей LSTM (Long short-term memory);

- исследовано влияние порядка ЦП текста на распознавание однородности произведения с помощью γ-классификатора;

- впервые в Таджикистане создан объектно-ориентированный компьютерный программный комплекс распознавания (идентификации) однородности текста на основе различных ЦП текста и γ-классификатора среди сколь угодно большого числа текстов.

Теоретическая значимость работы состоит в том, что в ней экспериментально опробован новый метод классификации дискретных случайных величин и установлена эффективность его применения для целей распознавания авторства и для самых разных типов «однородностей» произведений художественной литературы для любых естественных языков на основе различных ЦП текста.

Практическая ценность работы состоит в том, что она нацелена на применение созданного в ней компьютерного программного комплекса в государственной административной деятельности для автоматизации процесса обработки текстовой информации, в сфере криминалистики для установления авторства анонимных текстов, в области образования и науки для обнаружения плагиата в курсовых и дипломных проектах, а также в представленных к защите кандидатских и докторских диссертациях.

Положения, выносимые на защиту: экспериментальное доказательство эффективности применения γ-классификатора с помощью различных ЦП текста для распознавания однородности текстовой информации.

Достоверность и обоснованность полученных результатов подтверждены сериями вычислительных экспериментов, в которых посредством γ-классификатора и метода ближайшего соседа распознаются с достаточно высокой точностью самых разных типов «однородностей» произведения различных модельных коллекций и корпусов.

Соответствие диссертации паспорту научной специальности. Содержание исследования данной диссертации соответствует пунктам 1, 3, 4, 5 и 7 по специальности 05.13.11 - «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»:

– модели, методы и алгоритмы проектирования и анализа программ и программных систем, их эквивалентных преобразований, верификации и тестирования;

– модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем;

– системы управления базами данных и знаний;

– программные системы символьных вычислений;

– человеко-машинные интерфейсы; модели, методы, алгоритмы и программные средства машинной графики, визуализации, обработки изображений, систем виртуальной реальности, мультимедийного общения.

Публикации по теме диссертации. По теме диссертации опубликовано 73 работы, из них 34 (14 без соавторов) статьи в журналах из перечня, рекомендованных ВАК при Президенте Республики Таджикистан, 30 докладов в сборниках трудов и международных конференций, две монографии и два учебных пособия, а также пять баз данных и программ для ЭВМ, зарегистрированных в качестве объектов интеллектуальной собственности.

Структура диссертации и объём. Диссертация состоит из введения, шести глав, заключения и приложений. Библиографический список включает 397 наименований. Основная часть диссертации изложена на 271 странице. Диссертация содержит 9 рисунков и 107 таблиц.

Как директор подтверждаю весомый вклад диссертанта и считаю, что за прошедшее время А.А.Косимов существенно повысил свою научную квалификацию, поднявшись до уровня самостоятельно мыслящего, инициативного исследователя, способного выдвигать перспективные научные проекты, указывать пути решения поставленных задач, руководить подготовкой молодых специалистов.

По моему глубокому убеждению, работа А.А.Косимова отвечает всем требованиям ВАК как в теоретическом отношении, так и практической направленности и вполне готова к представлению в качестве научного доклада для государственной итоговой аттестации на предмет присуждения ему учёной степени доктора технических наук по специальностям 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Доктор физико-математических наук,
академик НАН Таджикистана, Рахмонов З.Х.

Подпись академика З.Х. Рахмонова удостоверяю.
Начальник ОК Института
математики им. А.Джураева Маллаева М.Р.