ОТЗЫВ

научного руководителя о работе КОСИМОВА Абдунаби Абдурауфовича “Разработка основ автоматической системы распознавания автора незнакомого текста (на примере художественных произведений на таджикском языке)”, представляемой на соискание учёной степени кандидата технических наук по специальности 05.13.11 – “Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей”

Актуальность темы диссертации А.А.Косимова заключается в том, что она связана с вопросами *криминалистики*, заинтересованной в установлении авторства анонимных текстов, с проблемами обнаружения плагиата в сфере *образования и науки* и, что особенно важно, имеет непосредственное отношение к государственной административной деятельности, все более охватываемой процессом автоматизации, в которой видное место занимает автоматическая обработка текстовой информации.

Цель исследования – алгоритмизировать процесс распознавания авторства таджикских текстов и реализовать его в виде компьютерного программного комплекса.

Задачи исследования. Для достижения цели решены следующие задачи:

- исследовать информативность признаков, предназначаемых для распознавания автора текста;

- проверить на эффективность математические методы распознавания текстов;

- определить минимальный размер неизвестного текста, пригодного для распознавания его автора;

- исследовать эффективность применения высокочастотных элементов алфавита символьных *N*-грамм для идентификации автора текста;

- спроектировать и реализовать компьютерный программный комплекс для идентификации авторства неизвестного текста не меньше допустимого

размера с привлечения лишь высокоточных элементов алфавита символьных *N*-грамм.

Методы исследования. Для решения задач, указанных в рубрике “Цель работы", использовались методы математической статистики, вычислительного эксперимента, теории множеств, системного анализа, распознавания и объектно-ориентированного программирования для разработки программных средств.

Научная новизна. Основные результаты диссертации являются новыми и заключаются в следующем:

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

- установлена эффективность использования в качестве текстовых дескрипторов буквенных *N*-грамм (N = 1, 2, 3) с учётом пробела;

- установлена эффективность принципиально нового модификатора, способного распознавать с точностью до 96% автора тестового фрагмента размером вплоть до 625 слов (2800 символов) и с точностью не менее 84% автора тестового фрагмента размером даже до 150 слов (670 символов);

- для целей существенного сокращения объёма вычислительных процедур установлена возможность эффективного использования не всех, а только высокоточных элементов алфавита буквенных *N*–грамм (N=1,2,3);

- впервые в Таджикистане создан объектно-ориентированный компьютерный программный комплекс для идентификации авторства неизвестного текста среди сколь угодно большого числа предполагаемых авторов.

Теоретическая значимость работы состоит в том, что в ней экспериментально подтверждена эффективность применения нового метода классификации для целей распознавания авторства незнакомых печатных текстов для любых естественных языков с буквенным алфавитом.

Практическая ценность работы состоит в том, что разработанный в ней компьютерный программный комплекс применим *в государственной*

административной деятельности для автоматизации процесса обработка текстовой информации, *в сфере криминалистики* для установления авторства анонимных текстов, *в области образования и науки* для обнаружения плагиата.

Структура диссертации. Работа А.А.Косимова состоит из введения и 5 глав, заключения и списка литературы из 141 наименований. Материал введения преподносится стандартным образом: вначале дан обзор исследований, имеющих отношение к работе, затем в соответствии с требованиями ВАК представлены ответы на вопросы по достижениям диссертанта.

В **главе 1** тестируется ряд нетрадиционных количественных характеристик на предмет возможности их использования в качестве информативных признаков для распознавания автора искомого текста. Особо важный момент в этой главе заключается в том, что *распределения частотности униграммм и биграмм* в произведениях классической и современной поэзиях, а также в современной прозе таджикского языка *с позиции их корреляции статистически не различимы*. В то же время с помощью распределений частотностей буквенных триграмм удалось доказать статистически достоверные различия, как между произведениями, так и между творчествами различных авторов.

В **главе 2** устанавливается, что модификатор З.Д.Усманова выдает более точные результаты, нежели критерий Н.В.Смирнова, причём оба метода оказываются более приспособленными к идентификации авторов, чем метод сравнения пар текстов по частотностям *N*-грамм средствами корреляционного анализа.

В **главе 3** устанавливается возможность идентифицировать автора по фрагменту текста столь же успешно, как и по полному произведению, причём частотности символьных униграмм, биграмм и триграмм являются вполне приемлемыми признаками для количественного описания математического образа текста. Основной результат главы в том, что с

помощью классификатора З.Д.Усманова удаётся 100% -но определить авторов текстовых фрагментов размерами вплоть до 625 слов и с достаточно большой точностью текстов длиной в 300, 150 и даже 75 слов.

В **главе 4** устанавливается способность классификатора З.Д.Усманова решать проблему идентификации на основе не полного, а усечённого алфавита *N*-грамм, который представляется относительно небольшим списком высокочастотных *N*-грамм, а именно: 6-ю высокочастотными униграммами, 55-ю высокочастотными биграммами и 1000-ми высокочастотными триграммами.

В **главе 5** дается подробное описание программного комплекса **"TTA"** (tajik text author), предназначенного для распознавания автора незнакомого текста.

В **Заключении** выделяются основные результаты диссертации.

Публикации по теме диссертации. По теме диссертации А.А.Косимов опубликовал 17 статей. Из них - 14 наименований в изданиях, рекомендованных ВАК Республики Таджикистан. В соавторстве с научным руководителем - 7 статей.

Как научный руководитель подтверждаю весомый вклад диссертанта во все наши совместные исследования. 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 кандидата технических наук по специальностям 05.13.11 – "Математическое и программное обеспечение вычислительных машин,

комплексов и компьютерных сетей".

Научный руководитель УСМАНОВ Зафар Джураевич
доктор ф.-м.н., академик АН РТ, профессор

Дата: «____» _____ 2018 г. Подпись: _____

Подпись д.ф-м.н., Усманова З.Д. заверяю:
Начальник отдела кадров
Института математики АН РТ
имени академика А.Д. Джураева _____ Хотамова Р.Л.

«__» _____ 2018 г.