ХУДЖАНДСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ
ТАДЖИКСКОГО ТЕХНИЧЕСКОГО УНИВЕРСИТЕТА
ИМЕНИ АКАДЕМИКА М.С. ОСИМИ

УДК 81’33+811.222.8+519.25

На правах рукописи

Косимов Абдунаби Абдурауфович

РАЗРАБОТКА ОСНОВ АВТОМАТИЧЕСКОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ АВТОРА НЕЗНАКОМОГО ТЕКСТА

(на примере художественных произведений на таджикском языке)

А В Т О Р Е Ф Е Р А Т

диссертации на соискание учёной степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Душанбе 2018

Научная работа выполнена в Худжандском политехническом институте Таджикского технического университета имени академика М.С. Осими

Научный руководитель: Усманов Зафар Джураевич, доктор физико-математических наук, академик АН РТ, профессор, заведующий отделом математического моделирования Института математики АН РТ

Официальные оппоненты: Пруцков Александр Викторович, доктор технических наук, Федеральное государственное бюджетное образовательное учреждение высшего образования «Рязанский государственный радиотехнический университет», профессор кафедры «Вычислительная и прикладная математика»

Умаров Махмуд Абубакирович, кандидат технических наук, Российско-Таджикский (Славянский) университет, доцент кафедры Информатики и информационных систем

Оппонирующая организация: Технологический университет Таджикистана

Защита состоится 14 декабря 2018 г. в 10:00 часов на заседании диссертационного совета 6D.KOA-032 при Таджикском техническом университете имени академика М.С. Осими, г.Душанбе, проспект академиков Раджабовых, 10.

С диссертацией можно ознакомиться в библиотеке Таджикского технического университета имени академика М.С. Осими и на официальном сайте университета: http://ttu.tj/2018/07/09/qosimov_a_a/

Автореферат разослан «___» «_____» 2018 года

Отзывы на автореферат в двух экземплярах, заверенные печатью учереждения, просим направлять по адресу: 734042, г.Душанбе, пр. акад. Раджабовых, 10, тел.: (+992 37) 227-37-81, e-mail: saidaliev.ss@mail.ru

Ученый секретарь диссертационного совета Ш.С. Саъдуллозода

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Проблема распознавания текста возникла, по-существу, одновременно с зарождением письменности. В течение длительного времени она экспонировалась лишь одной своей гранью – необходимостью определения исполнителя письменного произведения. В дальнейшем, с момента изобретения книгопечатания, в проблеме актуализировалась новая грань – потребность опознания автора печатной продукции, что в настоящее время и составило основное содержание всей проблемы.

Распознавание (или идентификация) текста определяется как процедура установления его автора по совокупности признаков, характеризующих особенности текста.

Теоретическая сторона проблемы связана с определением таких характеристик текстовой информации, которые, с одной стороны, не подконтрольны своим создателям, а, с другой стороны, содержат в себе косвенные дескрипторы об особенностях их стилей и, возможно, сведения об их индивидуальных качествах вроде национальности, профессии, образованности, предпочтениях и т.п.

Практическая сторона проблемы имеет отношение *к государственной административной деятельности,* все более охватываемой процессом автоматизации, в которой не последнее место занимает автоматическая обработка текстовой информации; *к криминалистике*, заинтересованной в установлении авторства анонимных текстов; *к сфере образования*, в которой студенческая молодежь не прочь воспользоваться плагиатом при выполнении курсовых и дипломных проектов.

Сказанное говорит в пользу актуальности избранной темы диссертации, в частности, по той причине, что исследования в столь важном направлении в Таджикистане разворачиваются впервые и в ближайшем будущем напрямую будут связываться с разработкой государственной системы информационной безопасности.

Между тем, в дальнем зарубежье работы в этой области знания заметно интенсифицировались в связи с развитием информационных технологий. В подтверждение этого факта достаточно обратиться к трудам A. Abbasi, M.F. Amasyah, C. Apte, S. Argamon, R.H. Baayen, J. Burrows, C.E. Chaski, M. Corney, O. De Vel, J.J. Diederich, B. Efron, J.M. Farringdon, T. Joachims, P. Juola, B. Kjell, M. Koppel, D. Lowe, K. Luyckx, R. Matthews, T.C. Mendenhall, A.Q. Morton, F. Peng, R.D. Peng, J. Rudman, E. Stamatatos, W.J. Teahan, F.J. Tweedie, S. Waugh, R. Zheng.

В России подобным вопросам посвящены исследования Н.А. Морозова, А.А. Маркова, В.П. Фоменко, Т.Г. Фоменко, Д.В. Хмелева, Г. Хетсо, А.А. Рогова, Ю.В.

Сидорова, А.Ю. Комиссарова, О.Г. Шевелева, В.В. Поддубного, М.А. Марусенко, А.А. Шелупанова, А.С. Романова, Р.В. Мещерякова, Ю.Н. Павлова, Е.А. Тихомировой, В.В. Дягилева, А.А. Цхая, С.В. Бутакова, А.О. Шумской, А.В. Седова, З.И. Резановой, Е.В. Шараповой, Р.В. Шарапова.

Настоящая диссертация посвящена изучению проблемы распознавания авторства текстовых фрагментов, написанных на таджикском языке.

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

Задачи исследования. Для достижения цели решаются следующие задачи:

- исследовать информативность признаков, предназначаемых для распознавания автора текста;
- проверить на эффективность математические методы распознавания текстов;
- определить минимальный размер незнакомого 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 незнакомого текста не меньше допустимого размера с привлечением лишь высокоточных элементов алфавита символьных *N*-грамм.

Объект исследования – модельная коллекция печатных текстов на таджикском языке.

Предмет исследования – распознавание авторства незнакомого текста на основе частотности его *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 модификатора З.Д.Усманова, способного распознавать с точностью до 96% автора текстового фрагмента размером вплоть до 625 слов (2800 символов) и с точностью не менее 84% автора текстового 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 незнакомого текста среди сколь угодно большого числа предполагаемых авторов.

Теоретическая значимость работы состоит в том, что в ней опробован экспериментально новый метод классификации дискретных случайных величин и установлена эффективность его применения для целей распознавания авторства незнакомых печатных текстов для любых естественных языков с буквенным алфавитом.

Практическая ценность работы состоит в том, что она нацелена на применения 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* для обнаружения плагиата в курсовых и дипломных проектах.

Положения, выносимые на защиту:

- количественный образ текстового фрагмента, характеризующий авторский стиль, в виде распределения частотностей буквенных *N*-грамм (*N*=1,2,3) с учётом пробела;

- целесообразность использования не всех, а только высокоточных элементов алфавита буквенных *N*-грамм;

- эффективность применения модификатора З.Д.Усманова, как метода, лучше других приспособленного для распознавания авторства самых коротких по размерам текстовых фрагментов.

Достоверность и обоснованность полученных результатов подтверждаются серией вычислительных экспериментов по идентификации с достаточно высокой точностью авторов в различных текстовых коллекциях.

Апробация результатов работы. Материалы работы обсуждались на:

- научно-исследовательских семинарах Института математики АН РТ, Худжандского политехнического института Таджикского технического университета имени академика М.С. Осими и Российско-таджикского (славянского) университета 2014-2017 гг.;

- научно-практический семинаре "Новые информационные технологии в автоматизированных системах". 2014 г., 2016 г., Москва;

- международной научно-практической конференции "Перспективы развития науки и образования", 2016 г., Душанбе;

- международной конференции «Kamal Khujandi: Development of literary study and literary relations», 28-29 октября 2016 г., Худжанд;

- международной научно-практической конференции «Роль ИКТ в инновационном развитии экономики Республики Таджикистан», 2017 г., Душанбе;

- международной научной конференции «Современные проблемы математики и их приложения», 14-15 июня 2017 г., Душанбе, Куляб.

Основные публикации. По теме диссертации опубликовано 17 статей, [1-17], из них - 14 наименований в изданиях, рекомендованных ВАК Республики Таджикистан.

Личный вклад автора. Постановка задачи осуществлялась совместно с научным руководителем. Основные результаты диссертационной работы получены автором самостоятельно.

Структура и объем диссертации. Диссертация состоит из введения, пяти глав, заключения и списка литературы из 141 наименования. Основная часть диссертации изложена на 104 страницах. Диссертация содержит 38 таблиц и 20 рисунков.

СОДЕРЖАНИЕ ДИССЕРТАЦИИ[1]

Введение к диссертации включает в себя все основные структурные элементы в соответствии с ГОСТ Р 7.0.11-2011.

В **главе 1** тестируется ряд нетрадиционных количественных характеристик на предмет возможности их использования в качестве информативных признаков для распознавания автора искомого текста. Характеристики нетрадиционные в том смысле, что они не содержатся среди порядка тысячи количественных признаков, которые по подсчётам Рудмана использованы различными исследователями для описания печатных текстов.

В **§1.1** изучается возможность применения обобщенной формулы золотой пропорции для описания положения точек кульминаций художественных произведений. На примере трёх поэм из «Шахнаме» А.Фирдоуси обнаруживаются три параметра, один из которых характеризует само произведение, а два других связываются с творчеством его автора. Полученный результат может успешно применяться для исследования авторства, прежде всего, тех произведений, в которых удается точно определить позицию точки кульминации. Надо полагать, что для текстов иных жанров опробованный метод может оказаться неприемлемым.

В **§ 1.2** для описания одиннадцати поэм из произведения А.Фирдоуси “Шахнаме” на таджикско–персидском языке в кириллической графике и их переводов на русский язык используются пять натуральных единиц измерения текста (число байтов в предложении, число слов в предложении, число букв в слове, число знаков без пробела в предложении и число знаков с пробелами в предложении). Путём проверки статистических гипотез установлено, что выборочные средние пяти показателей для всех поэм в оригинале и в переводе статистически неразличимы. Установленный факт подсказывает, что такой набор из 5 чисел можно интерпретировать в качестве цифрового кода для распознавания произведений А.Фирдоуси, причём его оригинальному творчеству и переводу на русский язык естественно сопоставить арифметические средние взвешенные показатели 11 поэм.

В **§ 1.3** в качестве количественной характеристики тестируется частотность встречаемости букв в печатных текстах. Для изучения этого вопроса используются те же самые 11 поэм из произведения А.Фирдоуси “Шахнаме”, дополненные поэмами Н.Хисрава, М.Турсунзода, М.Каноат, Л.Шерали и прозой С.Айни и Дж.Икроми. Обработка статистических данных показала высокую степень коррелируемости частотностей букв всевозможных пар произведений. Из этого последовал вывод о том, что частотности знаков таджикского алфавита (букв с пробелами и без них) в произведениях поэтов классической таджикско-персидской литературы, а также различных авторов современной таджикской поэзии и прозы, с точки зрения степени их коррелируемости во всевозможных

[1] В автореферате используются нумерация параграфов, формул, таблиц, рисунков и т.п. в соответствии с обозначениями, принятыми в диссертации.

парах произведений оказались статистически неразличимыми и потому не является показателем, пригодным для распознавания авторства текста.

Еще один набор признаков для идентификации авторов художественных произведений связывался с соотношением чисел словоформ $N_{сф}$ и словоупотреблений $N_{су}$. В пяти пунктах **§ 1.4** связь между ними описывалась с помощью двух формул

$$N_{сф} : N_{су} = a\,n + b$$

и

$$N_{сф} = \frac{a\,N_{су}}{1 + b\,N_{су}}$$

где a и b - константы, значения которых вычислялись по данным соответствующих произведений. С одной стороны, п.п. 1.4.1, 1.4.2 и 1.4.5 указывали на “сходство” значений констант для оригинальных (в кириллической и персидской графиках) и переводных (на русский язык) для поэм А.Фирдоуси. С другой стороны, п.п. 1.4.3 и 1.4.4 показывали существенные различия их значений как при сравнениях творчества А.Фирдоуси и К.Худжанди, не говоря уже о творчестве А.С.Пушкина.

В **§ 1.5** продолжается тестирование количественных характеристик на предмет их пригодности для идентификации авторов текстов. В п. 1.5.1 устанавливается, что *распределения частотности биграмм* в произведениях классической и современной поэзиях, а также в современной прозе таджикского языка *с позиции их корреляции статистически неразличимы*. В то же время с помощью распределений частотностей буквенных триграмм удаётся установить статистически достоверные различия как между произведениями, так и между творчествами различных авторов, см. п. 1.5.2.

Глава 2 посвящена предварительным исследованиям проблемы распознавания автора незнакомого текста. В основу разработки математических методов рассматриваемой предметной области закладывается вполне естественная гипотеза о том, что “почти все” произведения одного автора однородны, а двух разных авторов “почти всегда” неоднородны. Иными словами, каждому автору присущ свой стиль. И, в общем случае, два автора различаются по стилям.

Несмотря на то, что опытные литературные критики по неуловимым признакам распознают творения различных авторов, для математика авторский стиль остаётся понятием качественным, подлежащем шкалированию. Приступая к разработке этой проблемы для текстов на таджикском языке, мы должны воспользоваться некоторым количественным образом в качестве математической модели восприятия печатного текста, с одной стороны, и какими-то известными или же новыми математическими методами обработки текстовой информации, с другой стороны.

В **§ 2.1** представляются два метода, используемые для идентификации автора текста – статистический критерий однородности Н.В.Смирнова и сопутствующий ему модификатор, предложенный З.Д.Усмановым. В качестве количественных образов текста рассматриваются извлекаемые из него распределения символьных *N*-грамм (*N* = 1, 2, 3).

В **§§ 2.2 – 2.4** указанные методы тестируются на модельной коллекции из 10 таджикских текстов, принадлежащих 5 авторам: **С.Айни** “Аҳмади Девбанд”, “Одина”; **Дж.Руми “**Дафтари Аввал”, “Дафтари Дуввум”; **М.Турсунзода “**Садои Осиё”, “Ҳасани Аробакаш”; **А.Фирдовси** “Бежан ва Манижа”, “Рустам ва Сӯҳроб”; **Л.Шерали** “Катиба”, “Суханреза”. Количественными образами этих произведений служат поначалу распределения униграмм, затем биграмм и, наконец, триграмм. Итоговые результаты относительно эффективности тестируемых методов представлены в последующей таблице.

Таблица 2.7. - Эффективность идентификации авторства текста (в %) на основе символьных *N*-грамм (с учётом и без учёта пробелов)

	Униграмма		Биграмма		Триграмма	
	без проб.	с проб.	без проб.	с проб.	без проб.	с проб.
по критерию Н.В.Смирнова	92	92	84	92	93	100
по модификатору	96	96	92	96	93	100

Из полученных результатов следует, что для целей распознавания авторства описание текста

- с помощью символьных *N*-грамм с учётом пробелов более эффективно, чем без учёта пробелов;
- с помощью триграмм более эффективно, чем посредством униграмм и биграмм.

И в дополнении к сказанному, модификатор выдает более точные результаты, нежели критерий Н.В.Смирнова, причём оба метода оказываются более приспособленными к идентификации авторов, чем метод сравнения частотностей *N*-грамм пар текстов средствами корреляционного анализа.

В **главе 3** изучается следующий вполне естественный вопрос: возможно ли идентифицировать автора по *фрагменту текста* столь же успешно, что и по полному произведению, как это делалось в предыдущей главе. В диссертации решение проблемы даётся на примере модельной базы данных – той же самой коллекции текстов, что и в главе 2, составленной из 10 художественных произведений 5 авторов и расширенной до 12 произведений за счёт присоединения к ней рассказов С.Турсуна “Нисфирӯзӣ”, разделенных на две части, приблизительно одинаковых размеров по числу слов.

В **§ 3.1** уточняется постановка задачи. В частности, для исследовательских целей из каждого произведения извлечены упорядоченные по убыванию

последовательности текстовых фрагментов, начиная с достаточно больших размеров в 10000 слов и кончая относительно малыми размерами в 75 слов.

В качестве количественной характеристики текстов и их фрагментов рассматриваются только лишь распределения *N*-грамм (*N*=1,2,3). Для идентификации авторства применяется один метод – классификатор текстов З.Д.Усманова, который в сравнении с критерием Н.В.Смирнова показал на модельной коллекции текстов большую эффективность в идентификации авторов произведений.

В **§ 3.2** дается описание существа его метода в применении к задачам лингвистики. Пусть T_1 и T_2 – два каких-либо текста, законы распределения символьных *N*-грамм которых задаются в табличном виде

$$\begin{array}{ll} T_i: & 1 \ \cdots \ k \ \cdots \ m \\ P^{(i)}: & p_1^{(i)} \cdots p_k^{(i)} \cdots p_m^{(i)}, \end{array} \tag{3.1}$$

причём

$$\sum_{k=1}^{m} p_k^{(i)} = 1.$$

В этих выражениях k $(k = \overline{1,m})$ – порядковый номер k –й N-граммы в алфавите N-грамм, $p_k^{(i)}$ - относительная частота встречаемости k–й N-граммы в тексте T_i, $i = 1,2$. Тогда расстояние между T_1 и T_2 определяется по формуле

$$\rho(T_1, T_2) = \sqrt{\frac{m}{2}} \max_s \left| \sum_{k=1}^{s} (p_k^{(1)} - p_k^{(2)}) \right|, \tag{3.2}$$

где $s = \overline{1,m}$.

Пусть γ - некоторое положительное число. Тексты T_1 и T_2 называются γ-*однородными*, если

$$\rho(T_1, T_2) \leq \gamma. \tag{3.3}$$

и γ-*неоднородными*, если

$$\rho(T_1, T_2) > \gamma. \tag{3.4}$$

Предположим, что коллекция текстов T разделена на подмножества $T^{(j)}$, $j = \overline{1,n}$. Для фиксированного значения γ подсчитывается число $\aleph^0$ - сумма однородных пар текстов, принадлежащих подмножествам $T^{(j)}$, $j = \overline{1,n}$, и число $\aleph^н$ -сумма γ -неоднородных пар текстов, принадлежащих различным подмножествам. Отношение

$$\eta = \frac{\aleph^0 + \aleph^н}{\mathbb{N}}, \tag{3.5}$$

в котором $\mathbb{N}$ - общее число пар текстов в коллекции T, характеризует для заданного γ эффективность применения математической модели (3.1) – (3.4) к автоматическому разбиению коллекции T на подмножества $T^{(j)}$. З.Д.Усмановым предложен алгоритм для вычисления оптимального значения $\gamma^{\text{опт}}$, при котором достигается максимальная эффективность η для коллекции $T = \{T^{(j)}\}$.

В **§§ 3.3 - 3.5** определяется минимальный размер текста, необходимый для распознавания авторства с помощью символьных униграмм, биграмм и триграмм. Изучение вопроса разделяется на 5 этапов**.**

Этап 1. Из всех 10 произведений 5 авторов и двух частей рассказов С.Турсуна извлекаются по 10 текстовых фрагментов размерами в 10000, 5000, 4000, 2500, 1250, 800, 625, 300, 150 и 75 слов.

Этап 2. Для всех произведений и текстовых фрагментов вычисляются их количественные характеристики – распределения частот встречаемости *N*-грамм (*N* = 1, 2, 3).

Этап 3. Для текстовых фрагментов одинаковых размеров по формуле (3.2) из **§** 3.2 вычисляются их расстояния до 12 произведений 6 авторов.

Этап 4. Путём подбора оптимального значения $\gamma^{опт}$ по формуле (3.5) определяется коэффициент максимальной эффективности η разделения текстовых фрагментов фиксированного размера на подмножества однородных и неоднородных текстов.

Произвольный текстовый фрагмент относится к фиксированному множеству, порождаемому одним из 12 произведений 6 авторов, если расстояние между фрагментом и порождающим произведением меньше или равно $\gamma^{опт}$, и не относится к рассматриваемому множеству, если расстояние между ними строго больше $\gamma^{опт}$.

Этап 5. Процедуры 4-го этапа повторяются для каждого из 9 выбранных размеров.

Описанный алгоритм, программно реализованный, применён к упомянутой коллекции текстов, естественная классификация которой выражается в том, что два произведения одного автора однородны, а разных авторов неоднородны.

Сравнение итоговых результатов §§ 3.3–3.5 представлено в таблице 3.4 **§ 3.6**.

Таблица 3.4. - Зависимость коэффициента η эффективности классификации на основе *N*-грамм от размера текстового фрагмента (в словах) и значения $\gamma^{опт}$

Число слов	η и ($\gamma^{опт}$)					
	Униграммы		Биграммы		Триграммы	
	без проб.	с проб.	без проб.	с проб.	без проб.	с проб.
10000	1.00 **(0.06)**	1.00 **(0.06)**	1.00 **(0.40)**	1.00 **(0.40)**	1.00 **(2.20)**	1.00 **(1.95)**
5000	1.00 **(0.06)**	1.00 **(0.05)**	1.00 **(0.40)**	1.00 **(0.40)**	1.00 **(2.20)**	1.00 **(1.95)**
4000	0.98 **(0.06)**	1.00 **(0.05)**	1.00 **(0.40)**	1.00 **(0.40)**	1.00 **(2.30)**	1.00 **(1.95)**
2500	0.98 **(0.08)**	0.98 **(0.06)**	0.98 **(0.40)**	**1.00 (0.40)**	0.98 **(2.30)**	1.00 **(2.26)**
1250	0.98 **(0.08)**	0.98 **(0.07)**	0.98 **(0.40)**	0.98 **(0.40)**	0.98 **(2.40)**	**1.00 (2.31)**
800	0.93 **(0.08)**	0.93 **(0.07)**	0.93 **(0.50)**	0.93 **(0.50)**	0.93 **(2.50)**	0.96 **(2.71)**

625	0.93 **(0.08)**	0.93 **(0.07)**	0.93 **(0.50)**	0.93 **(0.50)**	0.93 **(2.50)**	0.96 **(2.71)**
300	0.67 **(0.17)**	0.84 **(0.11)**	0.71 **(0.90)**	0.82 **(0.70)**	0.82 **(4.90)**	0.87 **(3.91)**
150	0.62 **(0.17)**	0.73 **(0.12)**	0.71 **(0.90)**	0.82 **(0.70)**	0.76 **(5.20)**	0.84 **(4.07)**
75	0.49 **(0.23)**	0.53 **(0.19)**	0.44 **(1.50)**	0.62 **(1.10)**	0.51 **(8.06)**	0.62 **(6.41)**

В этой таблице в 1-м столбце отмечены длины фрагментов текста в словах. Затем следуют три блока (по два столбца в каждом), указывающие конкретно, какими *N*-граммами характеризовались тексты и их фрагменты. Первый и второй столбцы в блоках отмечают наличие или отсутствие пробела в алфавите *N*-грамм. И, наконец, в каждой ячейке таблицы представлены значения двух чисел – η и γ^{onm} (в скобках).

Из таблицы следует, что

- символьные униграммы, биграммы и триграммы являются вполне приемлемыми количественными характеристиками для решения проблемы идентификации авторов текстов;

- учёт пробелов в *N*-граммах повышает точность классификации;

- классификатор З.Д.Усманова (3.1) – (3.5) показывает достаточно высокий уровень идентификации авторов фрагментов текста размерами вплоть до 625 слов;

- по мере уменьшения размеров текстовых фрагментов эффективность их идентификации понижается, тем не менее представляется возможным использовать работу классификатора для текстов длиной в 300, 150 и даже 75 слов.

Отметим также, что из таблицы 3.4 следует, что с помощью 36 символьных униграмм (35 букв + пробел) удаётся 100% -но распознать автора текста размером не менее 4000 слов; соответствующий порог для биграмм (с учётом пробела их число равно 1296) измеряется 2500 словами, а для триграмм (численностью 46656) – даже 1250 словами.

В **главе 4** проверяется 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 (*N* =1,2,3). Соответствующие вычислительные эксперименты выполняются на коллекции текстов главы 3.

В § 4.1 указываются 4 этапа предобработки экспериментальных данных.

Этап 1. Для коллекции текстов в целом поначалу вычисляются частоты встречаемости униграмм, биграмм и триграмм. Из списка униграмм извлекаются 11 высокочастотных, осуществляющих 75.65% - покрытие коллекционных текстов; из списка биграмм - 55 высокочастотных, покрывающих 56.04% текстов; из списка триграмм - лишь первых 1000 высокочастотных, покрывающих 75.65% текстов.

Этап 2. Вычисляются частоты *N*-грамм (*N*=1,2,3) для каждого из 12 произведений модельной коллекции.

Этап 3. Из каждого произведения коллекции случайным образом извлекаются по три фрагмента-выборки размерами в 4000, 2500 и 1250 слов. Для

всех фрагментов размером в 4000 слов определяются частотности униграмм, для размеров в 2500 слов – частотности биграмм и для размеров в 1250 слов – частотности триграмм. Далее из списков *N*-грамм (*N*=1,2,3) выделяются высокочастотные: *11 первых униграмм, 55 первых биграмм и 1000 первых триграмм*.

Этап 4. Извлеченные на 3-ем этапе высокочастотные *N*-граммы, предназначавшиеся для сравнения с соответствующими *N*-граммами всех произведений, оказывались нетождественными по составу элементов. В этой связи на данном этапе производилась процедура “выравнивания” каждой пары сравниваемых распределений *N*-грамм. На примере $N = 1$ она заключалась в том, что с обеих сторон по 11 высокочастотных элементов объединялись в единое множество с тем, чтобы уже по новому списку элементов с частотностями, с которыми они присутствуют в соответствующих множествах, производить их тестирование на предмет однородности сопоставляемых пар.

В § 4.5 итоги исследований §§ 4.2.- 4.4 сведены в единую таблицу.

Таблица 4.7. - Эффективность классификации в зависимости от числа используемых высокочастотных *N*-грамм

Число *N*-грамм и процент покрытия – эффективность η и значение $\boldsymbol{\gamma^{опт}}$		
Униграмма (с проб.)	Биграмма (с проб.)	Триграмма (с проб.)
11 (75.65%) – 1.00 **(0.017)**	55 (56.04%) – 1.00 **(0.08)**	1000 (77.07%) – 1.00 **(0.17)**
8 (65.67%) – 1.00 **(0.014)**	50 (53.33%) – 0.98 **(0.07)**	500 (62.43%) – 0.93 **(0.21)**
7 (61.86%) – 1.00 **(0.013)**	45 (50.45%) – 0.98 **(0.07)**	300 (51.48%) – 0.91 **(0.31)**
6 (56.33%) – 1.00 **(0.013)**		
5 (50.74%) – 0.93 **(0.012)**		
3 (38.74%) – 0.84 **(0.009)**		

В каждой ячейке таблицы представлены четыре числа:

- 1-е число - число высокочастотных *N*-грамм, использованных для идентификации авторов текстов;
- 2-е число (в скобках) - процент покрытия текста высокочастотными *N*-граммами;
- 3-е число (после тире) - это значение η;
- 4-е число (в скобках, выделено жирным шрифтом) - значение $\gamma^{опт}$.

В пояснение таблицы напомним, что число различных униграмм (с учетом пробела) 36, биграмм – не более 1296 и триграмм – не более 46656. Из них в текстах встретилось 1073 биграмм (не встретилось – 223), 9916 триграмм (не встретилось – 36740).

На основе табличных данных заключаем, что

- учёт пробелов в *N*-граммах повышает точность классификации;
- классификатор З.Д.Усманова (3.1) – (3.5) показывает достаточно высокий уровень идентификации авторов ($\eta = 1$) при использовании 6 высокочастотных униграмм, 55 высокочастотных биграмм и 1000 высокочастотных триграмм;

- по мере уменьшения числа высокочастотных N-граммах показатель η эффективности идентификации авторов текстов понижается, тем не менее остается достаточно близким к значению $\eta = 1$; следовательно, классификатор З.Д.Усманова может быть использован для случаев использования 5 высокочастотных униграмм, 50 и 45 высокочастотных биграмм, 500 и даже 300 высокочастотных триграмм.

В §§ 5.1 – 5.6 **главы 5** дано подробное описание программного комплекса **“TTA”** (tajik text author), предназначенного для распознавания автора незнакомого текста.

ЗАКЛЮЧЕНИЕ

Основные результаты диссертации:

1. Проанализированы имеющиеся в зарубежной научной литературе данные о количественных признаках текстов и алгоритмах, применяемых при распознавании авторов. Определены перспективные направления исследований.

2. Вычислительными экспериментами обосновано использование символьных *N*-грамм (*N*=1,2,3) в качестве математической модели описания текстов на таджикском языке.

3. Установлена эффективность модификатора З.Д.Усманова, способного распознавать с точностью до 96% автора текстового фрагмента размером вплоть до 625 слов (2800 символов) и с точностью не менее 84% автора текстового фрагмента размером даже до 150 слов (670 символов).

4. Установлена возможность существенного сокращения объёма вычислительных процедур за счёт использования не всех, а только высокочастотных элементов алфавита буквенных *N*-грамм (*N*=1,2,3).

5. Создан первый в Таджикистане объектно-ориентированный компьютерный программный комплекс для идентификации авторства незнакомого текста для сколь угодно большого множества предполагаемых авторов.

Спроектированный комплекс рекомендуется для применения в автоматизации процесса обработки текстовой информации в государственной административной деятельности, для установления авторства анонимных текстов в сфере криминалистики, для обнаружения плагиата в курсовых и дипломных проектах в области образования, а также для использования в изучении самых разнообразных научных проблем, связанных с вопросами распознавания авторства печатных текстов.

СПИСОК ПУБЛИКАЦИЙ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

В журналах, рекомендованных ВАК:

1. Косимов, А.А. К вопросу о положении точки кульминации в художественных произведениях [Текст] / З.Д. Усманов, А.А. Косимов // Материалы 17 научно-практического семинара "Новые информационные технологии в автоматизированных системах", Москва. - 2014. - С. 392-395.
2. Косимов, А.А. Цифровой образ “Шахнаме” (“Книги царей”) А.Фирдоуси [Текст]. / З.Д. Усманов, А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2014. - Том 57. - № 6. - С. 471-476.
3. Косимов, А.А. Частотность букв таджикской литературы [Текст] / З.Д. Усманов, А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2015. - Том 58. - № 2. - С. 112-115.
4. Косимов, А.А. О соотношении словоформ и словоупотреблений в произведении А.Фирдоуси “Шахнаме” [Текст] / З.Д. Усманов, А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2015. - Том 58. - № 8. - С. 678-683.
5. Косимов, А.А. О соотношении словоформ и словоупотреблений в русском переводе произведения А.Фирдоуси “Шахнаме” [Текст] / Х.А. Худойбердиев, А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2015. - Том 58. - № 9. - С. 786-792.
6. Косимов, А.А. О соотношении словоформ и словоупотреблений в творчестве А.С. Пушкина [Текст] / З.Д. Усманов, А.А. Косимов // Материалы девятнадцатого научно-практического семинара "Новые информационные технологии в автоматизированных системах", Москва. - 2016. - С. 131-134.
7. Косимов, А.А. Частотность биграмм в таджикской литературе [Текст] / З.Д. Усманов, А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2016. - Том 59. - № 1-2. - С. 28-32.
8. Косимов, А.А. О распознавании авторства таджикского текста [Текст] / З.Д. Усманов, А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2016. - Том 59. - № 3-4. - С. 114-119.
9. Косимов, А.А. О множестве анаграмм в поэме А.Фирдауси “Шахнаме” [Текст] / А.А. Косимов // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. - 2016. - № 1 (162). - С. 48-53.
10. Косимов, А.А. Оценка эффективности использования униграмм при идентификации текста [Текст] / А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2017. - Том 60. - № 3-4. - С. 132-137.

11. Косимов, А.А. Оценка эффективности использования биграмм при идентификации текста [Текст] / А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2017. - Том 60. - № 5-6. - С. 224-229.

12. Косимов, А.А. Оценка эффективности использования триграмм при идентификации текста [Текст] / А.А. Косимов // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. - 2017. - № 1 (166). - С. 51-57.

13. Косимов, А.А. О минимальном объёме текста, необходимого для распознавания его автора [Текст] / А.А. Косимов // Доклады Академии наук Республики Таджикистан. - 2017. - Том 60. - № 9. - С. 398-401.

14. Косимов, А.А. О минимальном числе высокоточных n-грамм, необходимых для распознавания автора текста [Текст] / А.А. Косимов // Российско-китайский научный журнал «Содружество», Ежемесячный научный журнал, научно-практической конференции. - 2017. - Часть 1. - № 17. - С. 58-59.

Другие публикации:

15. Косимов, А.А. О соотношении словоформ и словоупотреблений в творчестве К.Худжанди [Текст] / Х.А. Худойбердиев, А.А. Косимов // Материалы VIII международной научно-практической конференции "Перспективы развития науки и образования", Душанбе. - 2016. - Часть 2. - С. 421-425.

16. Қосимов, А.А. Оиди муносибати шаклҳои калима ва калимаҳо дар ҳуруфоти форсии китоби “Шоҳнома”-и А.Фирдавсӣ [Текст] / А.А. Қосимов // Роль ИКТ в инновационном развитии экономики Республики Таджикистан, Материалы международной научно-практической конференции, Душанбе: Баҳманрӯд. - 2017. - С. 321-328.

17. Косимов, А.А. Определение минимального объёма выборки слов для идентификации текста [Текст] / А.А. Косимов // Вестник Таджикского национального университета, Серия естественных наук, Международной научной конференции «Современные проблемы математики и их приложения», Душанбе, Куляб. - 14-15 июня 2017. - №1/5. - С. 178-180.

АННОТАЦИЯ

диссертации Косимова Абдунаби Абдурауфовича на тему «Разработка основ автоматической системы распознавания автора незнакомого текста (на примере художественных произведений на таджикском языке)» на соискание ученой степени кандидата технических наук по специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

В работе алгоритмизируется процесс распознавания авторства таджикских текстов, для чего:

- исследуется информативность нетрадиционных лингвистических признаков на предмет количественного описания таджикских текстов;

- устанавливается эффективность использования в качестве текстовых дескрипторов символьных *N*-грамм и модификатора З.Д.Усманова для распознавания автора тестового фрагмента;

- для сокращения объёма вычислительных процедур устанавливается возможность эффективного использования только высокочастотных элементов алфавита символьных *N*-грамм.

Итогом работы является создание объектно-ориентированного компьютерного программного комплекса для идентификации авторства неизвестного текста.

Спроектированный комплекс рекомендуется для применения в автоматизации процесса обработки текстовой информации в государственной административной деятельности, для установления авторства анонимных текстов в сфере криминалистики, для обнаружения плагиата в курсовых и дипломных проектах в области образования, а также для использования в изучении самых разнообразных научных проблем, связанных с вопросами распознавания авторства печатных текстов.

ШАРҲИ МУХТАСАР

ба рисолаи диссертатсионии Қосимов Абдунаби Абдурауфович дар мавзӯи «Коркарди асосҳои системаи автоматии муайянкунии муаллифи матни номаълум (дар мисоли асарҳои бадеӣ бо забони тоҷикӣ)» барои дарёфти дараҷаи илмии номзади илмҳои техникӣ аз рӯйи ихтисоси 05.13.11 - Таъминоти математикӣ ва барномавии мошинҳои ҳисоббарор, муҷтамаъҳо ва шабакаҳои компютерӣ

Дар рисола раванди муайянкунии муаллифи матни тоҷикӣ алгоритмизатсия шуда, барои он:

- тадқиқи информативнокии аломатҳои ғайрианъанавӣ барои тавсифи миқдории матнҳои тоҷикӣ гузаронида шуд;
- самаранокии истифодаи *N*-граммаҳои ҳуруфӣ (*N*=1,2,3) ва модификатори Усмонов З.Ҷ. ба сифати аломатҳои муайян кардани муаллифи порчаи матн муқаррар карда шуд;
- бо мақсади комилан кам кардани ҳаҷми протседураи ҳисобкуниҳо имконияти самаранокии истифодаи на ҳама, балки танҳо элементҳои алифбои рамзии *N*-граммаҳои (*N*=1,2,3) баландбасомад муқаррар карда шуд.

Натиҷаи охири кор ин сохтани комплекси барномаҳои компютерии ба объект нигаронидашуда барои муайян кардани муаллифи матнҳои номаълум мебошад.

Комплекси тарҳрезишуда барои ҷорикунӣ дар автоматикунонии раванди коркарди маълумоти матнӣ дар фаъолияти маъмурияти давлатӣ барои муқаррар кардани муаллифи матни номаълум дар соҳаи амният, барои муайян кардани асардуздӣ дар корҳои курсӣ ва дипломӣ дар соҳаи маориф, аз он ҷумла барои омӯзиши проблемаҳои илмии гуногун, ки бо масъалаи муайян кардани муаллифи матни чопӣ алоқа дорад, истифода мешавад.

ANNOTATION

on the dissertation of Kosimov Abdunabi Abduraufovich on the theme «Development of the fundamentals of the automatic recognition system for the author of an unfamiliar text (on the example of art works in the Tajik language)» for candidate a degree of technical sciences on a specialty 05.13.11 – Mathematical and software of computers, complexes and computer networks

In the work, the process of recognition of the authorship of Tajik texts is algorithmized, for which:

- the informativeness of non-traditional linguistic characteristics is investigated for quantitative description of Tajik texts;

- the efficiency of using as character descriptors character *N*-gram and modifier Z.D. Usmanov for recognition of the author of the test fragment is established;

- to reduce the amount of computational procedures, setups possibility of efficiently using only high-frequency elements of the alphabet of symbolic *N*-grams.

The result of the work is the creation of an object-oriented computer software package for identifying the authorship of an unknown text.

The developed package is recommended for use in automating the processing of textual information in state administrative activities, for establishing the authorship of anonymous texts in the field of forensic science, for detecting plagiarism in course and diploma projects in the field of education, and for use in studying a wide variety of scientific problems related to issues of recognition of authorship of printed texts.

Подписано к печати 09.11.2018
Формат 6х84/16. Бумага офсетная
Тираж 100 экз. Объём 1,3 п.л. Заказ №181
Отпечатано в типографии «Мехвари дониш»
г. Худжанд, ул. Ленина, 226