УТВЕРЖДАЮ
Ректор Университета ИТМО
член-корреспондент РАН
_____ В.Н. Васильев
«12» 02 2015

ОТЗЫВ

ведущей организации на диссертационную работу Смирнова Сергея Владимировича по теме «Технология и система автоматической корректировки результатов при распознавании архивных документов», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.11 — Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Актуальность темы диссертации

Сохранение исторического наследия является актуальной задачей во всем мире, повсеместно реализуются проекты по массовой оцифровке книг, газет и других бумажных документов. В России для многих библиотек, архивов, музеев работы по сканированию своих фондов уже давно стали плановыми. В результате сформированы большие массивы электронных образов и на первый план выходят задачи по их распознаванию и построению систем эффективного поиска по ним.

В диссертационной работе Смирнова С.В. проводится сравнительный анализ систем оптического распознавания при обработке архивных документов, который свидетельствует о том, что каждая из систем допускает ошибки независимо от качества и временного периода документов. Это обусловливается тем, что архивные документы охватывают широкий набор тематических областей, каждая из которых отличается своей узкоспециализированной терминологией, отсутствующей в базовых словарях систем оптического распознавания. Актуальной задачей является разработка и программная реализация методов корректировки результатов распознавания.

Анализ проблематики, проведенный соискателем, выявил необходимость в разработке системного подхода к распознаванию архивных документов различных предметных областей с последующей корректировкой результатов, который позволил бы выполнять автоматическую обработку электронных образов и расширение поисковой базы архивов в плановом режиме без задействования ручного труда сотрудников.

Тема диссертационной работы Смирнова С.В., посвященная разработке технологии, системы и методов автоматической корректировки результатов оптического распознавания архивных документов в целях повышения качества их электронного представления и точности поиска по ним, является актуальной научно-технической проблемой, решение которой имеет существенное значение для науки и практики.

Анализ содержания диссертационной работы

Текст диссертации состоит из введения, четырех глав, заключения, списка литературы и трех приложений. Материал диссертационного исследования изложен на 130 страницах машинописного текста. Список литературы включает 122 наименования.

Во **введении** обоснована актуальность темы, сформулированы цель и задачи исследования, а также результаты, выносимые на защиту.

В **первой главе** рассматривается концепция построения системы распознавания архивных документов с автоматической корректировкой результатов, определяется назначение системы и основные требования к ней, описывается процесс обработки документов и последующего поиска ним с подсветкой вхождения ключевой фразы при отображении поисковых результатов. Система должна обеспечивать автоматическое распознавание русскоязычных архивных документов различных тематических областей и различных категорий: от высококачественных документов, напечатанных на современных принтерах до документов среднего качества, напечатанных на печатной машинке в середине 20 века.

На основе обзора существующих проектов и систем корректировки ошибок распознавания делается вывод о том, что во многих случаях существующие методы предназначены для обработки современных текстов и зачастую не подходят для обработки исторических текстов, содержащих большое количество

специализированных терминов. Подчеркивается, что в большинстве работ корректировка основана на предварительном ручном обучении системы или участии человека на этапе финального выбора корректировки.

Во **второй главе** предложен метод автоматической корректировки ошибок распознавания на основе рейтинго-ранговой модели текста. Весь процесс корректировки разделяется автором на четыре основных этапа: подготовка структур данных, генерация корректировок, ранжирование корректировок и формирование результата.

В ходе предварительного этапа подготовки структур данных производится сбор статистической информации по всему корпусу распознанных документов, формируется целый ряд словарей, тезаурусов и хэш-таблиц, содержащих необходимые данные для этапа генерации корректировок.

Далее для каждой выявленной ошибки распознавания осуществляется отбор корректировок методом анаграмм по предварительно построенным тезаурусам, учитывающим частотные характеристики повторений слов и словосочетаний во всем корпусе распознанных материалов. Использование таких тезаурусов, позволяет производить корректировку текстов различных предметных областей, содержащих узкоспециализированную терминологию, имена собственные, географические наименования и т.п.

При ранжировании полученных списков корректировок предпочтение отдается словам, которые наиболее часто встречаются во всем корпусе результатов распознавания, обладают наименьшим расстоянием Левенштейна от ошибочного слова, были наиболее часто возвращены в ходе отбора методом анаграмм или содержатся в словарях. Дополнительно учитывается статистическая вероятность нахождения корректировки на месте ошибочного слова в тексте по известным предыдущим словам.

Выбор наилучших корректировок осуществляется на основе ряда эвристик, которые помимо финального ранга учитывают семантику ошибочного слова и корректировок.

В **третьей главе** дается описание технологии распознавания архивных документов с корректировкой результатов и ее применение в процессе массовой

обработки документов электронного архива, определяются группы пользователей системы и последовательность работы с ней.

Далее в главе приводится архитектура и компонентная модель разработанной системы. Особое внимание уделяется описанию программных компонентов настройки профилей и сравнительного анализа, представляющих в своей совокупности инструментарий, предназначенный для настройки параметров процессов распознавания и корректировки.

Преимуществом архитектуры разработанной системы является возможность расширения ее функциональных возможностей на каждом шаге обработки за счет подключения дополнительных библиотек. Стоит отметить также модульную организацию разработанного программного обеспечения, благодаря которой обеспечивается возможность эффективного масштабирования и сопровождения.

В **четвертой главе** представлены сведения об опытной эксплуатации разработанных технологии и системы на документах центральных государственных архивов Санкт-Петербурга в составе государственной информационной системы «Государственные архивы Санкт-Петербурга».

Описание технических характеристик инфраструктуры развертывания разработанной системы распознавания, схемы ее интеграции с другими подсистемами и особенностей корпуса документов позволяет сформировать достаточно полное и ясное представление о среде проведения испытаний и эксплуатации. С электронными образами архивных документов можно свободно ознакомиться на сайте архивов Санкт-Петербурга в сети интернет.

Оценка метода автоматической корректировки производилась по разработанным соискателем критериям, учитывающим особенности применения результатов распознавания в поисковой системе электронного архива и отвечающим требованиям, описанным в первой главе, что подтверждает его практическую пригодность и значимость.

Представленная в конце главы положительная оценка результатов распознавания и корректировки корпуса документов размером около 700 тысяч изображений позволяет сделать заключение о корректности выдвинутых теоретических положений и эффективности их практической применимости.

В **заключении** перечислены основные результаты выполненного диссертационного исследования, приведена информация об их внедрении, сделаны предложения по дальнейшему развитию.

Список литературы включает в себя работы как отечественных, так и зарубежных авторов. Достаточно полно представлены основные работы по теме исследования.

Новизна исследований и полученных результатов

Научная новизна диссертации заключается в разработке методологических основ подготовки, настройки и проведения потокового распознавания архивных документов различных тематических областей с автоматической корректировкой получившихся результатов, в создании прикладных основ решения проблемы в виде методов и правил корректировки ошибок оптического распознавания и определения наилучших корректировок.

В целом научную новизну диссертации составляют следующие положения.

Разработанный метод автоматической корректировки ошибок распознавания, отличительной особенностью которого является способность производить корректировку текстов, содержащих большое количество узкоспециализированной терминологии, отсутствующей в современных словарях, а также самодостаточность и автономность. Для обнаружения ошибочных слов, поиска и выбора наилучших корректировок используются тезаурусы и рейтинго-ранговые распределения, построенные в автоматическом режиме на основе статистического анализа корпуса распознанных документов и дополнительных электронных материалов архива. От оператора не требуется проведения ручного подбора, составления словарей и обучения системы.

Реализован подход к выбору наилучших корректировок на основе двухэтапной модели ранжирования, учитывающей инвариантную и контекстно-зависимую статистическую вероятность корректировки, и последующего применения ряда эвристических правил определения наиболее подходящих корректировок для замены ошибочного слова.

Разработана технология массового распознавания архивных документов, регламентирующая последовательность операций по кластеризации документов по тематическим группам, настройки конфигурационных профилей обработки

сформированных групп и их сравнительного анализа, автоматическому распознаванию и корректировки получившихся результатов. Автоматизация всех стадий обработки обеспечивается разработанной системой с включенным в ее состав инструментарием настройки конфигурации.

Практическая значимость результатов исследования

Методологические и прикладные основы решаемой в диссертации проблемы позволяют перейти на новый, отвечающий современным потребностям уровень формирования электронных баз данных оцифрованных документов и организации систем поиска по ним.

Практическая ценность диссертационного исследования состоит в том, что теоретические результаты доведены до уровня методического, алгоритмического и программного обеспечения, получившего реализацию при обработке документов различных предметных областей (литература и искусство, личный состав ликвидированных предприятий, научно-техническая документация, историко-политические документы) в виде технологии и системы автоматического распознавания архивных материалов с корректировкой результатов.

Применение результатов диссертационной работы в Архивном комитете и подведомственных ему архивах Санкт-Петербурга позволило произвести массовое распознавание накопленного массива электронных образов научно-справочного аппарата и организовать потоковое распознавания вновь загружаемых изображений документов в автоматическом режиме с целью расширения поискового пространства и пополнения мета базы данных

В центральном государственном архиве документов по личному составу ликвидированных государственных предприятий Санкт-Петербурга, документы которого содержат большое количество наименований предприятий, специальностей, должностей, имен собственных и других узкоспециализированных терминов, предложенные в диссертационном исследовании методы и алгоритмы автоматической корректировки ошибок распознавания оказались особенно актуальными и эффективными.

Реализация предложенных в исследовании систем распознавания, корректировки и поиска архивных документов в составе государственной информационной системы «Государственные архивы Санкт-Петербурга» значительно облегчило

процессы обнаружения необходимой информации, что положительно отразилось на эффективности исполнения запросов и скорости оказания государственных услуг.

Достоверность и обоснованность результатов исследований

Достоверность результатов обусловлена корректным выбором методики исследования и подтверждена данными, полученными в ходе экспериментальной апробации разработанных технологии и системы автоматической корректировки результатов при распознавании документов центральных государственных архивов Санкт-Петербурга в составе государственной информационной системы «Государственные архивы Санкт-Петербурга». Основные выводы по результатам исследований и рекомендации по их использованию обоснованы.

Рекомендации по использованию результатов и выводов диссертации

Разработанные в ходе диссертационного исследования алгоритмы и программы могут найти практическое применение в достаточно широком круге предметных областей, таких как системы электронного документооборота, электронно-библиотечные системы, системы электронных архивов. Это подчеркивает важность расширения сферы практических приложений проводимых исследований и позволяет рекомендовать их продолжение и развитие в государственных и муниципальных архивах РФ, центральных государственных архивах и Архивном комитете Санкт-Петербурга, Санкт-Петербургском государственном унитарном предприятии «Санкт-Петербургский информационно-аналитический центр», федеральных и региональных библиотеках и музеях, а также в других учреждениях РФ.

На основе результатов диссертации может быть создан электронный сервис, предоставляющий услуги по оптическому распознаванию документов с последующей автоматической и ручной корректировкой результатов для различных бюджетных и коммерческих организаций или физических лиц посредством открытых или защищенных каналов передачи данных.

Замечания по диссертационной работе

1. Из работы не ясно, в какой степени разработанные технология, система и метод автоматической корректировки результатов оптического распознавания архивных материалов подходят для обработки документов на иностранных языках.

2. В работе отсутствуют практические примеры и количественное описание структур данных для обнаружения ошибок, отбора и ранжирования корректировок, сформированных во время обработки документов центральных государственных архивов, что несколько затрудняет понимание реальной работы разработанного метода.

3. В четвертой главе при описании экспериментальной апробации основных результатов диссертационной работы не указываются оценки временных затрат на проведение ручной настройки конфигурационных профилей системы и на выполнение процессов распознавания и автоматической корректировки всего корпуса документов.

4. В тексте диссертации присутствует смешение русских и английских терминов, например одновременное использование «точность», «полнота», «Precision» и «Recall» на странице 101.

Перечисленные замечания не снижают высокий научный уровень проведенных исследований и не влияют на общий положительный вывод о качестве представленной к защите диссертации. Замечания носят рекомендательный характер и могут быть учтены автором в дальнейших публикациях по теме исследования.

Вывод

Учитывая вышеизложенное, можно сформулировать вывод о том, что диссертационная работа Смирнова Сергея Владимировича является законченной научно-исследовательской работой, выполненной на актуальную тему, отличается научной новизной и практической значимостью полученных результатов.

Автором в диссертации сформулирована и решена важная научно-техническая проблема разработки технологии и системы автоматической корректировки результатов оптического распознавания архивных документов.

Основные этапы работы, выводы и результаты представлены в автореферате, который достаточно полно отражает содержание диссертации. По материалам диссертационной работы опубликовано 13 научных работ, в том числе 6 в периодических журналах, рекомендованных ВАК, получено 2 свидетельства о государственной регистрации программы для ЭВМ.

Тематика диссертации, формулировка ее целей, научной новизны и областей применения полученных результатов подтверждают соответствие диссертации специальности 05.13.11.

Диссертационная работа отвечает критериям «Положения о порядке присуждения ученых степеней» и соответствует требованиям ВАК РФ, предъявляемым к кандидатским диссертациям, а ее автор, Смирнов Сергей Владимирович, заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Диссертационная работа и отзыв обсуждены и одобрены на заседании кафедры вычислительной техники Университета ИТМО, присутствовало 19 из 23 человек, протокол № 11 от «28» апреля 2015г.

Заведующий кафедрой вычислительной техники Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», заслуженный работник высшей школы Российской Федерации, профессор, доктор технических наук
Почтовый адрес: 197101, г. Санкт-Петербург, Кронверкский проспект, д.49
Телефон: +7 (812) 233-22-54
Электронная почта: aliev@cs.ifmo.ru

Алиев Тауфик Измайлович
12.05.2015

Профессор кафедры вычислительной техники Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», профессор, доктор технических наук
Почтовый адрес: 197101, г. Санкт-Петербург, Кронверкский проспект, д.49
Телефон: +7 (812) 232-52-78
Электронная почта: tau@d1.ifmo.ru

Тропченко Александр Ювенальевич