Федеральное государственное бюджетное учреждение науки
Санкт-Петербургский институт информатики и автоматизации Российской академии наук
(СПИИРАН)

199178, Санкт-Петербург, 14 линия, 39
Телефон: (812)328-33-11
Факс: (812)328-44-50
E-mail: spiiran@iias.spb.su
http://www.spiiras.nw.ru
ОКПО 04683303, ОГРН 1027800514411
ИНН/КПП 7801003920/780101001

«27» 01 2015 № 11607-02/62/35
На N .....

УТВЕРЖДАЮ
Директор СПИИРАН
член-корреспондент РАН

_____ Р.М. Юсупов

«27» января 2015 г.

ЗАКЛЮЧЕНИЕ
Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН)

Диссертация «Технология и система автоматической корректировки результатов при распознавании архивных документов» выполнена в лаборатории автоматизации научных исследований. В период подготовки диссертации соискатель Смирнов Сергей Владимирович работал в Санкт-Петербургском государственном унитарном предприятии «Санкт-Петербургский информационно-аналитический центр» в отделе проектирования и разработки электронных архивов в должности начальника сектора.

В 2008 году окончил «Санкт-Петербургский государственный университет информационных технологий, механики и оптики», факультет компьютерных технологий и управления по специальности «вычислительные машины, комплексы, системы и сети».

С 2010 года является соискателем СПИИРАН. Удостоверение о сдаче кандидатских экзаменов выдано в 2014 году Федеральным государственным бюджетным учреждением науки «Санкт-Петербургский институт информатики и автоматизации Российской академии наук». Закончил диссертацию на соискание ученой степени кандидата технических наук.

Научный руководитель — Кулешов Сергей Викторович, доктор технических наук, ведущий научный сотрудник лаборатории автоматизации научных исследований СПИИРАН.

По результатам рассмотрения диссертации «Технология и система автоматической корректировки результатов при распознавании архивных документов» принято следующее заключение:

Оценка выполненной соискателем работы

В диссертационной работе Смирнова Сергей Владимировича приведен сравнительный анализ качества результатов распознавания архивных документов современными системами оптического распознавания текста, выявлена необходимость в корректировке получаемых результатов распознавания и проведен анализ существующих методов автоматической и полуавтоматической корректировки ошибок в текстах. Разработан специализированный подход к распознаванию архивных документов различной тематики и предложен метод корректировки допускаемых ошибок на основе рейтинго-ранговой модели текста. Метод основывается на поиске корректировок по заранее подготовленным корпусным и тематическим тезаурусам и позволяет производить корректировку результатов распознавания, содержащих узкоспециализированную терминологию, специфичную для определенной предметной области. Предложенные в работе алгоритмы и методы были реализованы в виде системы потокового распознавания архивных документов, включающую подсистему автоматической корректировки ошибок распознавания, и подсистему поиска по изображениям документов архивного фонда.

Актуальность и востребованность данной тематики также подтверждается большим вниманием, уделяемым задачам и проектам по оцифровке документов культурного наследия, хранящихся в архивах, библиотеках, музеях и других учреждениях, для расширения электронного информационного пространства страны и увеличения его доступности для граждан.

Личное участие соискателя в получении результатов, изложенных в диссертации.

Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованных работах. Подготовка к публикации полученных результатов проводилась автором самостоятельно с незначительным участием соавторов. Представленные к защите результаты получены лично автором.

Достоверность результатов проведенных исследований.

Достоверность подтверждена аналитическим обзором исследований и разработок в области построения систем распознавания изображений и корректировки ошибок в тексте, положительными итогами практического использования результатов диссертационной работы в прикладных системах электронных архивов, а также апробацией основных научно-практических положений в печатных трудах и докладах на всероссийских и международных конференциях.

Научная новизна полученных результатов.

Научная новизна состоит в разработанном методе автоматической корректировки ошибок распознавания текста на основе рейтинго-ранговой модели текста, разработанных правилах ранжирования корректировок, основанных на предварительно проведенном статистическом n-грамм анализе корпуса результатов распознавания и тематических текстов, разработанной системе распознавания архивных документов с автоматической корректировкой результатов и инструментарии, позволяющем эксперту ограничивать пространство конфигураций для наиболее эффективного решения задачи распознавания.

Практическая значимость полученных результатов.

Практическая значимость диссертационной работы заключается в том, что полученные в ней результаты являются в достаточной степени универсальными и фундаментальными, что подтверждается эффективностью их применения при распознавании разнообразных по качеству и тематике документов центральных государственных архивов Санкт-Петербурга, а также успешной реализацией в составе государственной информационной системы «Государственные архивы Санкт-Петербурга».

Специальность, которой соответствует диссертация.

Работа соответствует требованиям, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук по специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Полнота изложения материалов диссертации в работах, опубликованных соискателем.

Основные положения и результаты диссертации получили полное отражение в докладах на 7 всероссийских и международных конференциях, в 13 печатных работах, среди которых 6 работ в рецензируемых изданиях из перечня ВАК, получено 2 свидетельства о государственной регистрации программы для ЭВМ.

Основные результаты диссертации изложены в следующих работах в необходимой полноте:

1. Смирнов С.В., Белозерова М.В. Оцифровка, каталогизация, хранение и поиск архивной документации // Информационно-измерительные и управляющие системы. 2010. Т.8. №7. С. 97-101.
2. Кулешов С.В., Смирнов С.В. Методы сегментации OCR систем в задачах автоматической обработки архивных документов // Труды СПИИРАН. 2011. Выпуск 1 (16). С. 110–122.

3. Смирнов С.В. Подсистема массового распознавания изображений архивных документов // Труды СПИИРАН. 2012. Выпуск 3 (22). С. 234-248.
4. Смирнов С.В. Методы автоматической постобработки результатов распознавания в задачах оцифровки архивных документов // Информационно-измерительные и управляющие системы. 2013. Т.11. №9. С. 22-32.
5. Смирнов С.В. Сравнительный анализ OCR систем в контексте построения системы поиска по изображениям архивных документов // Информационно-измерительные и управляющие системы. 2014. Т.12. №12. С. 62–69.
6. Смирнов С.В. Корректировка ошибок оптического распознавания на основе рейтинго-ранговой модели текста // Труды СПИИРАН. 2014. Выпуск 4 (35). С. 64-82.
7. Смирнов С.В. Таксономия информационных объектов электронного архива // Сборник научных трудов Всероссийской научно-практической конференции-форума молодых ученых и специалистов «Современная российская наука глазами молодых исследователей». Красноярск. 2011. С. 192-194.
8. Смирнов С.В. Логическая модель представления информации в электронном архиве // Сборник научных трудов IV Всероссийской научно-практической конференции с международным участием «Научное творчество XXI века». Красноярск. 2011. Выпуск 2. С. 93-94.
9. Смирнов С.В. Критерии оценки качества результатов оптического распознавания // Сборник материалов XVI Международной научно-практической конференции «Перспективы развития информационных технологий». Новосибирск. 2013. С. 33–38.
10. Смирнов С.В. Система массового оптического распознавания архивных документов с автоматической корректировкой результатов // Материалы XIV Санкт-Петербургской международной конференция «Региональная информатика (РИ-2014)». Санкт-Петербург. 2014. С. 302.
11. Смирнов С.В. Особенности построения системы массового оптического распознавания архивных документов // Труды XVII Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург. 2014. С. 37-42.
12. Смирнов С.В. Система полнотекстового поиска по изображениям архивных документов // Сборник материалов XXI Международной научно-практической конференции «Перспективы развития информационных технологий». Новосибирск. 2014. С. 16–21.
13. Воронцов А.В., Кожин А.В., Смирнов С.В. Настоящее и будущее государственных электронных архивов Санкт-Петербурга // Материалы X всероссийской научно-практической конференции «Электронные ресурсы библиотек, музеев, архивов». Санкт-Петербург. 2014. С. 106-114.

Диссертация «Технология и система автоматической корректировки результатов при распознавании архивных документов» Смирнова Сергея Владимировича рекомендуется к защите на соискание ученой степени кандидата технических наук по специальности 05.13.11 — математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей. Заключение принято на расширенном семинаре лабораторий автоматизации научных исследований, речевых и многомодальных интерфейсов, информационных технологий в системном анализе и моделировании. Присутствовало на семинаре 7 чел. Результаты голосования: «за» — 7 чел., «против» — 0 чел., «воздержалось» — 0 чел., протокол №5 от 15.12.2014 г.

Заведующий лабораторией автоматизации
научных исследований СПИИРАН,
доктор технических наук, профессор В.В.Александров

199178, Санкт-Петербург, 14-линия В.О., д.39,
тел. (812) 323-51-39, alexandr@iias.spb.su