ОТЗЫВ

на автореферат диссертации Смирнова Сергея Владимировича на тему «Технология и система автоматической корректировки результатов при распознавании архивных документов», представленной на соискание ученой степени кандидата технических наук по специальности 05.13.11 - «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Актуальность диссертационной работы Смирнова С.В., в которой разрабатываются методы и технология автоматической корректировки результатов распознавания изображений, не вызывает сомнения, так как до сих пор не существует системы оптического распознавания, которая производила бы безошибочное извлечение текстового содержимого из электронных образов документов. Особенно это актуально при обработке архивных документов, созданных десятилетиями назад, отличающихся обилием узкоспециализированной терминологии и потерявшими свое исходное качество, зачастую изначально не высокое, в процессе длительного хранения. В связи с этим исследования в данном направлении весьма перспективны, важны и своевременны.

Работы, направленные на улучшение качества распознавания документов, актуальны для применения практически во всех сферах жизнедеятельности человека, которые связаны с необходимостью ведения истории и архива. В качестве примера можно привести медицинское обслуживание, правовое и законодательное дело, учет актов гражданского состояния и многие другие.

В диссертации решается проблема построения единой системы для обработки документов различных тематик и предметных областей, с учетом их специфики, лексикона и даты происхождения.

Научная новизна диссертационных исследований состоит в разработанной автором теории построения систем массового оптического распознавания и созданных на ее основе алгоритмов и методов автоматической корректировки результатов. Основные научные результаты следующие:

- разработан метод автоматической корректировки результатов распознавания на основе рейтинго-ранговой модели текста, сформированной путем лингвостатистического анализа всего корпуса распознанных документов;
- предложены правила ранжирования и выбора наилучших корректировок для замены ошибочных слов, основанные на вероятности их нахождения на определенной позиции в тексте и частотных характеристиках встречаемости в тексте;
- разработан способ настройки системы для распознавания различных групп документов при помощи специализированного инструментария, позволяющего произвести вычисление широкого набора критериев оценки качества распознавания.

Практический интерес представляет разработанная автором информационная система потокового распознавания архивных документов, реализующая предложенные методы автоматической корректировки и позволяющая экспертам проводить обработку в соответствии с описанной технологией.

Достоверность научных положений и выводов диссертации подтверждается тем, что в работе достаточно четко определено целевое назначение результатов распознавания и разработаны критерии оценки качества, соответствующие заданным требованиям. Опытная

эксплуатация разработанной системы проводилась на реальных документах центральных государственных архивов Санкт-Петербурга. Результаты представлялись на всероссийских и международных конференциях 2011-2014 годов.

Содержание автореферата соответствует специальности, по которой диссертация представляется к защите.

Автореферат не лишен недостатков. Описание корпуса документов, на котором производилась практическая апробация работы, содержит только сводные количественные характеристики. Информация о разбиении на тематические группы отсутствует, хотя это должно было производиться, судя по представленной технологии.

Отсутствует обоснование выбора OCR систем «Abbyy Finereader» и «Tesseract» для участия в оценке разработанного метода автоматической корректировки.

Нет информации о том, какой системой оптического распознавания производилась обработка всего корпуса документов центральных государственных архивов СПб.

Заключение

Несмотря на отмеченные недостатки, диссертационная работа, судя по автореферату и публикациям, представляет завершенный и интересный научный труд, имеющий важное народнохозяйственное значение. По новизне, научной значимости и обоснованности, практической ценности выполненных исследований и разработок диссертация отвечает требованиям Положения ВАК России о порядке присуждения ученых степеней, предъявляемым к кандидатским диссертациям. Смирнов Сергей Владимирович заслуживает присуждения ему ученой степени кандидата технических наук по специальности 05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Доцент кафедры Систем и технологий управления,
Федеральное государственное автономное
образовательное учреждение высшего образования
«Санкт-Петербургский
государственный Политехнический
университет Петра Великого»,
к.т.н., доцент Вячеслав Витальевич Потехин

195220, Санкт-Петербург,
Гражданский пр., д. 28
Телефон: (812) 329-4745
Электронная почта: slava.potekhin@gmail.com

25 мая 2015 г.

Управление персонала

УДОСТОВЕРЯЮ
Ведущий специалист
по кадрам
25 05 2015 г.