ОТЗЫВ

официального оппонента доктора технических наук, профессора Городецкого Андрея Емельяновича на диссертационную работу Смирнова Сергея Владимировича «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 диссертационной работы

В последние годы активно развивается тенденция перевода бумажных документов в электронный вид. Основной целью ставится обеспечение их сохранности и увеличение доступности посредствам цифровых систем и каналов передачи данных. Особенную актуальность данная тема приобретает для материалов, представляющих историческую культурную ценность для страны и общества. В современную цифровую эпоху информация и знания, содержащиеся в документах на бумажной основе, являются намного более значимыми и доступными, если они переведены в цифровой вид. В противном случае существует высокая вероятность того, что они будут утеряны, забыты или вовсе уничтожены вследствие пожаров и других чрезвычайных ситуаций.

Первым шагом к переводу бумажного архива в электронную форму является сканирование. На следующем этапе обычно производится оптическое распознавание каждого изображения машинописного документа с целью извлечения текста для дальнейшей обработки (классификации, извлечения метаданных, автореферирования, полнотекстовой индексации и т.п.).

Очевидно, что обработка данных, содержащих ошибки, трудна, а ее результаты влекут за собой искажение информации и последующую

некорректную интерпретацию. Появление ошибок оптического распознавания, как убедительно показано в диссертационной работе Смирнова С.В., происходит в наше время даже при обработке современных документов, обладающих хорошим качеством печати. Когда же дело доходит до распознавания исторических архивных документов, на которых текст имеет меньшую контрастность, возможны повреждения бумаги, начертание букв не позволяет провести четкое сопоставление с шаблоном, а лексикон не соответствует типовым встроенным словарям систем распознавания, то отработанные технические решения, успешно применяемые для офисного документооборота, становятся неработоспособными в связи со значительным возрастанием количества ошибок.

Предложенные в диссертации Смирнова С.В. методологические основы и прикладные методы, направленные на решение вышеописанной проблемы, в виде разработанных технологии и системы автоматической корректировки результатов оптического распознавания, имеют важное теоретико-практическое значение и являются своевременными и актуальными.

Научная новизна результатов и выводов работы

Новизна диссертационной работы состоит в разработке технологии распознавания больших массивов архивных документов с последующей постобработкой результатов посредством разработанных системы и метода автоматической корректировки ошибок распознавания.

Следует отметить следующие результаты диссертационной работы:

1. Разработан метод автоматической корректировки результатов оптического распознавания на основе рейтинго-ранговой модели текста, построенной на основе лингвостатистического анализа всего корпуса распознанных архивных материалов. Новизна метода заключается в полностью автоматическом процессе формирования необходимых тезаурусов, рейтинговых распределений и других

структур данных для обнаружения ошибок, отбора и ранжирования корректировок, а также применяемых правилах выбора наилучших корректировок, основанных на предварительно проведенном n-грамм анализе, и учитывающих статистическую вероятность сочетаемости с предшествующими словами.

2. Разработаны технология и система потокового распознавания архивных документов, адаптированные под специфику предметной области и предназначенные для корректировки текстов, содержащих большое количество узкоспециализированной терминологии.
3. Разработан инструментарий настройки конфигурации процессов распознавания и корректировки, основанный на формировании профилей, определяющих множество значений параметров для наиболее эффективного распознавания отдельных типов документов. Особенность данного инструментария заключается в возможности проведения оценки и сравнительного анализа качества распознавания различных групп документов с разнообразными конфигурационными профилями по множеству рассчитываемых системой критериев.
4.

Обоснованность, достоверность и практическая значимость результатов

Обоснованность и достоверность научных положений, основных выводов и результатов диссертации обеспечиваются тщательным анализом видов ошибок оптического распознавания и существующих на текущий момент способов их корректировки. Работа хорошо структурирована, в конце каждой главы имеются четкие обоснованные выводы, отражающие суть производимых исследований и разработок.

Корректность предложенного метода и алгоритма его реализации подтверждается экспериментальной апробацией в составе разработанной системы при распознавании документов центральных государственных архивов

Санкт-Петербурга.

Практическая значимость работы состоит в разработке системы и технологии распознавания документов различных тематических областей, которая может быть использована в проектах по оцифровке фондов библиотек, музеев, архивов государственных и коммерческих предприятий и других учреждений. Применение результатов диссертационной работы позволит значительно сократить временные затраты на проведение дорогостоящей ручной корректировки ошибок распознавания, а в некоторых случаях и вовсе позволит отказаться от нее.

Результаты, полученные в диссертации, нашли практическое применение и реализацию в государственной информационной системе «Государственные архивы Санкт-Петербурга» и были внедрены в центральных государственных архивах и Архивном комитете Санкт-Петербурга, а также в СПб ГУП «Санкт-Петербургский информационно-аналитический центр». Реализация результатов подтверждена актами внедрения, представленными в приложении к диссертации.

Апробация работы и публикации

Основные полученные результаты полностью соответствуют паспорту специальности 05.13.11 и неоднократно докладывались и были одобрены на различных всероссийских и международных конференциях. Содержание работы, суть выполненных исследований и личный вклад соискателя с необходимой степенью полноты отражены в 13 публикациях, из них 6 из перечня ВАК.

Замечания по работе

1. В главе 1.4 указывается то, что в работе будут применяться методы корректировки, использующие результаты распознавания одной OCR

системы, но не объясняется, почему не планируется использовать методы, основанные на сравнении результатов нескольких OCR систем.

2. Описание метода вычисления расстояния Левенштейна в главе 2 можно было бы опустить, сославшись на соответствующую литературу.

3. Отсутствует информация об элементах списка стоп слов D^{stop}, используемого для фильтрации слов при построении структур данных для ранжирования корректировок. Состав списка стоп слов можно было бы привести в приложении к диссертации.

4. Одним из результатов диссертации является разработанный инструментарий, позволяющий ограничивать пространство конфигураций для наиболее эффективного решения задачи, однако в работе приводится лишь общее описание множества допустимых параметров, конкретные примеры и возможные значения параметров отсутствуют, их размещение в приложении внесло бы дополнительную наглядность и ясность.

5. На странице 61 указывается, что одной из групп пользователей системы являются «Эксперты, осуществляющие обработку всего корпуса документов» без уточнения, что это за эксперты и в каких областях знаний.

Отмеченные недостатки не влияют на общую положительную оценку теоретических и практических результатов выполненной работы.

Заключение

Диссертационная работа Смирнова С.В. по своему содержанию, объему выполненных исследований, новизне, научной и практической значимости результатов представляет завершенный и интересный научный труд, является научно-квалификационной работой, в которой разработаны метод, система и

технология, позволяющие существенно повысить эффективность процесса перевода бумажных документов в электронный вид. Автореферат соответствует основному содержанию диссертационной работы.

Считаю, что диссертационная работа Смирнова Сергея Владимировича удовлетворяет требованиям п.9 «Положения о присуждении ученых степеней», предъявляемым ВАК РФ к кандидатским диссертациям, а ее автор заслуживает присуждения ему ученой степени кандидата технических наук по специальности 05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Заведующий лабораторией
интеллектуальных электромеханических систем
Федерального государственного бюджетного
учреждения науки
Института проблем машиноведения
Российской академии наук,
доктор технических наук, профессор Андрей Емельянович Городецкий

«7» мая 2015 г.

Почтовый адрес: В. О. Большой пр., д. 61, Санкт-Петербург, 199178
Телефон: (812) 321 90 07
Электронная почта: g27764@yandex.ru

Подпись Городецкого А.Е. удостоверяю
Отдел кадров ФИО
07.05.2015