Особенности построения системы массового оптического распознавания архивных документов

С.В. Смирнов

Санкт-Петербургский информационно-аналитический центр
serge.smir@gmail.com

Аннотация

В статье рассматривается проблематика построения систем массового оптического распознавания. Описывается алгоритм корректировки ошибок распознавания, архитектура и компонентная модель разработанной системы.
Также приводятся сведения об эксплуатации в центральных государственных архивах Санкт-Петербурга.

1. Введение

Сфера деятельности архивов и библиотек включает в себя широкий спектр задач, связанных с комплектованием, учетом, использованием и обеспечением сохранности документов.

Эффективность выполнения каждой задачи имеет сильную зависимость от скорости нахождения и получения доступа к нужным документам. Поиск документов является своего рода «узким» местом во всех рабочих процессах и накладывает серьезные ограничения на время выполнения ежедневных задач архива.

Зачастую поиск производится лишь по документам, обладающим текстовым описанием. Текстовое описание вручную заносится в систему операторами и сотрудниками организаций в процессе составления научно-справочного аппарата и оцифровки бумажных документов.

Данный подход к пополнению поисковой базы обладает большой трудоемкостью и как следствие низкой скоростью и малым покрытием.

Выходом из сложившейся ситуации может стать разработка решения, позволяющего пользователям производить поиск по текстовому содержимому изображений документов без необходимости предварительного ручного ввода поисковых метаданных.

В рамках работ по развитию государственной информационной системы «Государственные архивы Санкт-Петербурга» был разработан программный комплекс, решающий поставленную задачу.

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

Программный комплекс состоит из трех подсистем: подсистемы оптического распознавания, подсистемы полнотекстовой индексации результатов распознавания, подсистемы поиска по распознанным изображениям документов.

Ключевым элементом является система оптического распознавания, описанию которой и посвящена данная работа.

2. Проблематика проектирования системы

При разработке системы учитывался ряд особенностей:

1. Отсутствие времени и ресурсов на ручное распознавание и ручную корректировку результатов распознавания
2. Отсутствие времени и ресурсов на ручной отбор и поиск документов, пригодных для распознавания.
3. Отсутствие времени и ресурсов на постановку в очередь на обработку документов, пригодных к распознаванию.
4. Отсутствие времени и ресурсов на ручной контроль качества распознавания каждого документа.
5. Отсутствие времени и ресурсов на ручное обучение.

Также особое внимание на этапах проектирования и разработки системы массового распознавания уделялось следующим проблемным областям [1]:

- оценка качества и характеристик обрабатываемых документов;
- назначение результатов распознавания;
- выбор OCR систем;
- корректировка ошибок распознавания;
- автоматический контроль качества распознавания.

2.1. Качество и характеристики обрабатываемых документов

Документы центральных государственных архивов Санкт-Петербурга, подлежащие распознаванию, подразделяются на следующие тематики: историко-политические документы, документы по литературе и искусству, документы по личному составу ликвидированных

государственных предприятий, научно-техническая документация.

Общий объем корпуса материалов составляет чуть менее 50 тысяч документов, 1 миллиона изображений и 200 миллионов слов.

Примеры изображений документов приведены на рисунке 1. Формат изображений – JPEG, разрешение 300dpi.

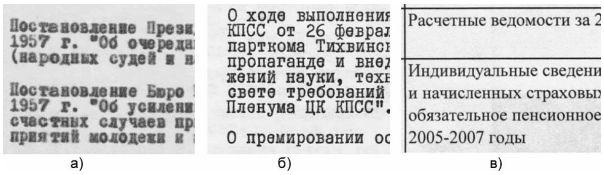

Рис. 1. Примеры изображений: а) печатная машинка, среднее качество; б) печатная машинка, высокое качество; в) принтер, очень высокое качество

2.2. Назначение результатов распознавания

В начале этапа проектирования необходимо было определиться с назначением результатов распознавания. Назначение выбиралось из следующих вариантов [2]:

1. Полнотекстовое индексирование — результат распознавания рассматривается как простой текст и в дальнейшем подается на вход поисковой системы. Текст используется как основа для полнотекстового поиска. Причем, конечному пользователю в результате поиска отображается найденный образ документа без обозначения вхождения поисковой фразы. Данный вид не требователен к точности распознавания и одновременно предоставляет хорошие поисковые возможности.

2. Отображение с подсветкой результатов на образе — в данном режиме распознанный текст обрабатывается также как и в предыдущем случае, а отличие заключается в подсистеме отображения поисковых результатов. В результатах поиска пользователю отображается изображение с выделенными фрагментами вхождений поисковой фразы. Очевидно, что в данном случае требования к качеству распознавания возрастают, но одновременно с этим увеличивается и эффективность поисковой системы в отличие от предшествующего способа отображения результатов.

3. Выдача результатов в виде неразмеченного текста — поисковым результатом является непосредственно текст, полученный в результате распознавания, а оригинальное изображение документа не отображается. Если распознанные слова будут сильно искажены, то пользователь не сможет получить искомой информации, и потеряет доверие к системе. Таким образом, точность должна быть очень высокой, что практически не может быть достигнуто без привлечения человеческого ресурса, и, как следствие, ведет к значительным временным и финансовым затратам.

4. Воссоздание оригинального документа — отображение результатов распознавания редко производится без форматирования и разметки текста, с целью сохранения исходной структуры и деталей расположения элементов. В дополнение, размеченный xml документ может содержать дополнительные атрибуты, тэги или ссылки на родственные документы.

В разработанной системе результаты распознавания используются лишь на промежуточном этапе полнотекстового индексирования. Пользователю поисковый результат отображается в виде подсвеченных областей на изображении.

Перечисленные обстоятельства снижают требования к способностям OCR системы проводить структурный анализ документа, что существенно увеличивает круг систем подходящих под задачи исследования. Обязательными требованиями являются лишь умение обрабатывать русскоязычные тексты и наличие в результатах распознавания x,y координат найденных слов.

2.3. Выбор OCR систем

Самостоятельная разработка OCR систем представляет собой довольно сложную научную и техническую задачу и не может являться обоснованной для большинства проектов по оцифровке. Особенно при условии того, что на рынке присутствует порядка десятка различных OCR систем, отличающихся условиями распространения, стоимостью, предоставляемыми функциями и, разумеется, качеством генерируемых результатов.

Поэтому, наиболее актуальной задачей становится выбор подходящей для конкретного проекта OCR системы. Самым надежным подтверждением правильности выбора является проведение сравнительного анализа результатов распознавания. При проведении сравнения необходимо опираться на показатели, которые наиболее полно отвечают будущим целям использования полученных результатов распознавания.

В проведенном сравнительном анализе участвовали наиболее популярные OCR системы: “Abbyy Finereader”, “Cuneiform Linux”, “Cuneiform Windows”, “IRIS Readiris”, “Nuance OmniPage”, “Tesseract”.

Сравнительный анализ выявил наличие ошибок в результатах распознавания архивных документов различного качества среди всех систем оптического распознавания. Минимальные и максимальные показатели точности распознавания на уровне слов отображены в таблице 1.

Таблица 1. Точность результатов распознавания

Качество документов	Максимум	Минимум
среднее	76,80%	22,01%
высокое	90,55%	30,28%
очень высокое	99,25%	90,20%

Анализ точности указал на необходимость проведения корректировки ошибок распознавания.

2.4. Корректировка ошибок распознавания

Качество процесса корректировки во многом зависит от точности нахождения ошибок и их верной классификации. Общая схема классификации ошибок в текстах и уточненная классификация ошибок в результатах оптического распознавания приводятся в работах [3, 4] соответственно.

В работе [5] приводятся результаты проведенного обзора методов и подходов, применимых к задаче контекстной обработке результатов распознавания. Среди которых выделен ряд способов предобработки результатов распознавания путем шаблонной замены, сравнения выходных значение нескольких OCR систем, описаны методы обнаружения ошибок, основанные на оценке вероятности, n-грамм анализе, словарной проверке.

Обзор показал, что существующие методы в общем случае неплохо решают ряд задач по обработке результатов распознавания с использованием словарей, статистических моделей языка, хорошо развита тематика обнаружения и коррекции ошибок в тексте. Тем не менее, во многих случаях указанные методы предназначены для обработки современных текстов и не подходят в чистом виде для обработки исторических текстов, содержащих большое количество специализированных терминов, имен собственных, географических наименований и т.п. В большинстве работ корректировка основана на предварительном ручном обучении системы или участии человека на этапе финального выбора слова-заместителя. Также стоит отметить, очень малое количество работ нацеленных на корректировку именно русскоязычных текстов.

Это вызвало потребность разработки алгоритма корректировки, учитывающего особенности русского языка и позволяющего обрабатывать корпуса текстов больших объемов в полностью автоматическом режиме.

3. Алгоритм корректировки ошибок оптического распознавания

Для корректировки результатов распознавания был разработан метод, основанный на рейтинго-ранговой модели текста. Подробное описание метода приводится в работе [6], здесь же рассмотрим только ключевые особенности.

Разделим весь процесс корректировки результатов распознавания на четыре основных этапа (рисунок 2):

1. Подготовка структур данных.
2. Генерация корректировок.
3. Ранжирование корректировок.
4. Формирование результата.

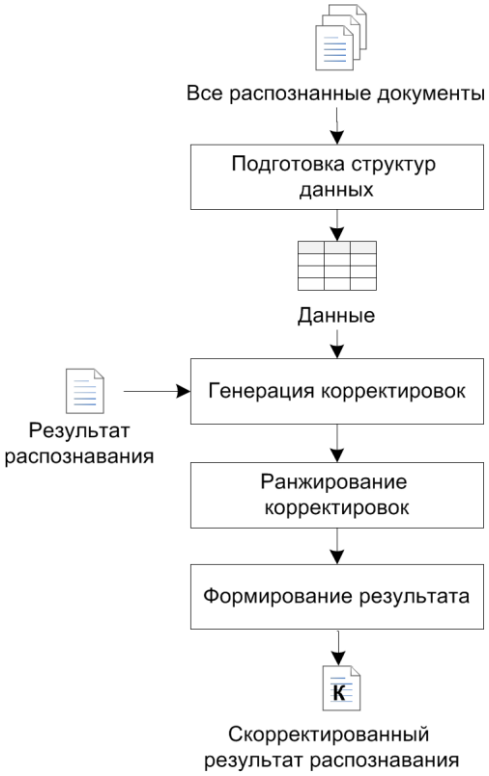

Рис. 2. Основные этапы корректировки

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

2. На этапе генерации корректировок для каждого ошибочно распознанного слова формируются списки слов-кандидатов на замену.

Отбор слов-кандидатов осуществляется по специальным глоссариям, построенным на основе частотных характеристик повторений слов и словосочетаний со всего корпуса распознанных материалов. Использование таких глоссариев, позволяет производить корректировку текстов различных предметных областей, содержащих узкоспециализированную терминологию, имена собственные, географические наименования и т.п. Все множество ошибок распознавания можно разделить на множество ошибок 1-го рода (пропущенные слова) и множество ошибок 2-го рода (ошибочно распознанные слова). На этапе генерации корректировок обработке подвергаются только ошибки 2-го рода.

3. Далее каждой корректировке присваивается ранг, на основе достоверности ее появления в тексте на месте ошибочного слова. После производится финальное упорядочивание корректировок по убыванию их ранга.

Финальное ранжирование отобранных слов-кандидатов производится с учетом контекста и основывается на результатах статистического n-

грамм анализа всего корпуса текста в нормальной форме. Реализация функции перевода слова в нормальную форму на основе морфологических предсказаний позволяет генерировать нормальные (базовые) формы даже для слов русского языка, отсутствующих в словаре.

4. На последнем этапе производится выборка наиболее вероятных слов-заместителей, их подстановка и сохранение финального результата распознавания в формате XML.

Разработанный метод использует расстояние Левенштейна [7] в качестве критерия оценки близости корректировки и ошибочного слова, и алгоритм поиска схожих слов методом анаграмм [8].

4. Архитектура и компонентная модель системы

4.1. Архитектура

Архитектура разработанной системы, изображенная на рисунке 3, построена по классической трёхуровневую модели.

Работа с системой осуществляется через веб приложение, разработанное на языке программирования Java [9]. Веб-приложение может функционировать в любом контейнере сервлетов [10].

Программная реализация системы состоит из веб-приложения, набора прикладных программ и утилит, базы данных.

БД системы может быть развернута на любой СУБД, поддерживающей реляционную модель хранения данных.

Вызов прикладных программ осуществляется через программную оболочку веб-приложения, что позволяет производить увеличение вычислительной мощности за счет горизонтального масштабирования серверов приложений.

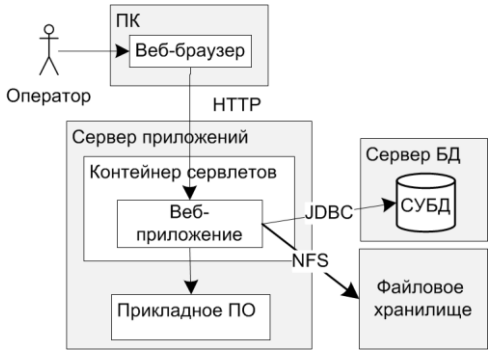

Рис. 3. Архитектура системы

В основе программной реализации системы лежит свободно распространяемое ПО, что делает систему потенциально более доступной для применения в других проектах.

4.2. Компонентная модель

Компонентная модель разработанной системы представлена на рисунке 4.

Разработанная система состоит из трех программных комплексов (ПК) связанных между собой единой БД и программным интерфейсом для взаимодействия с внешними подсистемами:

1. ПК подготовки к работе предназначен для ручного тестирования процесса распознавания на единичных изображениях. В результате тестирования для разных типов изображений создаются профайлы, содержащие в себе настройки каждого этапа распознавания и общую технологическую схему обработки. После этого необходимо произвести сравнительный анализ созданных профайлах на различных наборах данных и выявить наиболее подходящий.

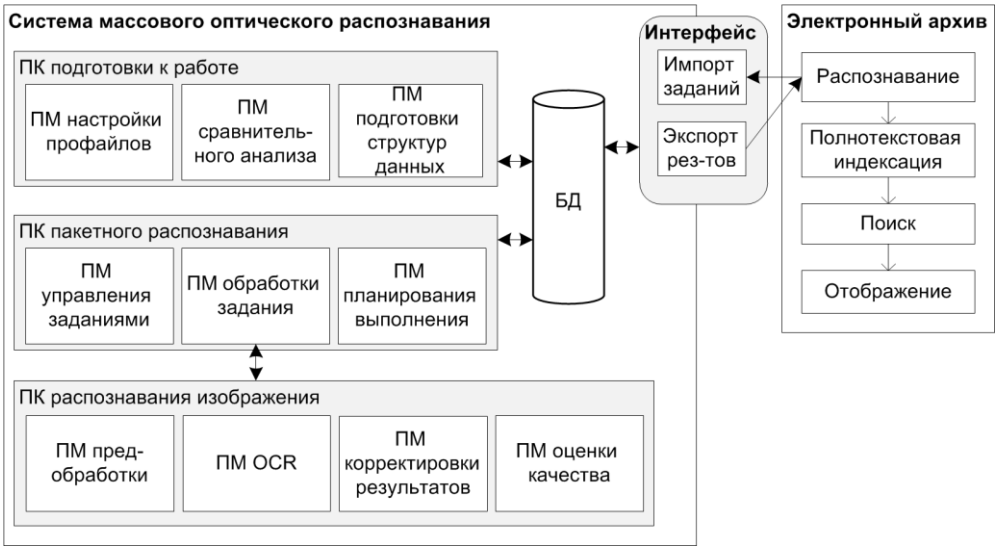

Рис. 4. Компонентная модель системы

Также в задачи подготовки к работе входит предварительный разбор всего корпуса распознанных текстов и построение структур данных, необходимых для автоматической корректировки ошибок

2. ПК пакетного распознавания предназначен для управления ходом выполнения заданий на обработку изображений документов. Его основными задачами являются: предоставление возможности просмотра журнала заданий, управление приоритетами заданий, вызов процедур распознавания отдельных изображений, сбор результатов и запись их в БД.

3. ПК распознавания изображения отвечает за процесс обработки отдельного изображения в соответствии с заданной в профайле технологической схемой.

Программный модуль (ПМ) предобработки занимается подготовкой изображений к процессу извлечения текста. В основу программной реализации положена библиотека ImageMagick [11]. ПМ выполняет следующие задачи: бинаризацию, устранение «шума», выравнивание угла наклона и т.п.

ПМ OCR отвечает за перевод изображений текста в машиночитаемую и редактируемую форму. Программная реализация основана на ряде свободно распространяемых систем оптического распознавания: Tesseract [12], Cuneiform [13].

ПМ корректировки реализует алгоритмы автоматической корректировки результатов распознавания, описанные в главе 3.

ПМ оценки точности производит расчет метрик, характеризующих точность распознавания.

4. Программный интерфейс системы предоставляет ряд API вызовов для постановки на распознавание отдельных документов ЭА, опроса состояния и получения результатов.

5. Апробация системы

Испытания системы проводились в составе государственной информационной системы «Государственные архивы Санкт-Петербурга» на базе архивов:

- Центральный государственный архив Санкт-Петербурга (ЦГА),
- Центральный государственный архив историко-политических документов Санкт-Петербурга (ЦГАИПД),
- Центральный государственный архив литературы и искусства Санкт-Петербурга (ЦГАЛИ),
- Центральный государственный архив документов по личному составу ликвидированных государственных предприятий, учреждений, организаций Санкт-Петербурга (ЦГАЛС),
- Центральный государственный архив научно-технической документации Санкт-Петербурга (ЦГАНТД).

За первые месяцы эксплуатации было обработано 32 тысячи документов, состоящих из 700 тысяч изображений. Полученные результаты распознавания содержат около 200 миллионов слов.

Следующей задачей ставится, проведение оценки точности распознавания, выявление результатов ненадлежащего качества и оптимизации методов предобработки, распознавания и корректировки для повышения итогового качества обработки.

6. Заключение

Основной особенностью разработанной системы является гибкая архитектура, позволяющая подключать различные коммерческие и свободно распространяемые OCR системы и библиотеки предобработки изображений.

Система может быть настроена на распознавание документов различных категорий качества. Для одной категории потребуется подключение дорогостоящих движков распознавания, для другой хорошие результаты будет давать бесплатная OCR система.

Также главным преимуществом системы является наличие процедур автоматической корректировки ошибок распознавания, позволяющих выявлять и исправлять ошибки даже в текстах, изобилующих специфическими терминами, именами собственными, узкоспециализированным лексиконом. Это особенно важно для исторических, архивных документов.

Система разработана в виде автономного программного комплекса и может быть интегрирована с другими информационными системами.

Литература

[1] Anderson N. IMPACT Best Practice Guide: Optical Character Recognition – Part 1. 2010 URL: http://www.impact-project.eu/uploads/media/IMPACT-ocr-bpg-pilot-s1.pdf (дата обращения: 06.06.2012)

[2] Tanner S. Deciding Whether Optical Character Recognition is Feasible. 2004. URL: http://www.odl.ox.ac.uk/papers/OCRFeasibility_final.pdf (дата обращения: 06.06.2012)

[3] Kai N. Unsupervised Post-Correction of OCR Errors // Hannover: Leibniz University. 2010.

[4] Kukich K. Techniques for automatically Correcting Words in Text // ACM computing survey Computational Linguistic. 1992. vol. 24. no. 4. pp. 377–439.

[5] Смирнов С.В. Методы автоматической постобработки результатов распознавания в задачах оцифровки архивных документов // Информационно-измерительные и управляющие системы. 2013. №9. С. 22–32.

[6] Смирнов С.В. Корректировка ошибок оптического распознавания на основе рейтинго-

ранговой модели текста // Труды СПИИРАН. 2014. Выпуск 4(35).
[7] Левенштейн В. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. 1965. Т. 163. № 4. С. 845-848.
[8] Reynaert M. Non-interactive OCR Post-correction for Giga-Scale Digitization Projects // Computational Linguistics and Intelligent Text Processing. 2008. pp. 617–630.
[9] Java. URL: http://java.com (дата обращения 18.09.2014).
[10] Mordani R. Java Servlet Specification, Version 3.0. USA. 2009.
[11] ImageMagick: Convert, Edit, Or Compose Bitmap Images. URL: http://www.imagemagick.org/ (дата обращения 18.09.2014).
[12] Tesseract-ocr. URL: http://code.google.com/p/tesseract-ocr/ (дата обращения: 29.04.2014).
[13] Cuneiform Windows. URL: http://cognitiveforms.com/ru/products_and_services/ cuneiform (дата обращения: 29.04.2014).

Design Features of Mass Optical Character Recognition system of archival documents

S.V. Smirnov

The article focuses on the problems of constructing systems of mass OCR. An algorithm for correction of recognition errors, architecture and component model of the system are described.
Also information about the deployment in the central state archives of St. Petersburg is present.