Сервис-ориентированная архитектура для системы распознавания печатных исторических текстов

© А.В. Рудалёв[1], А.Г. Варфоломеев[2]

[1]Северный (арктический) федеральный университет имени М.В. Ломоносова,
[2]Петрозаводский государственный университет
alexv@pomorsu.ru, e-humanities@yandex.ru

Аннотация

Данная статья посвящена описанию проекта программной системы распознавания печатных исторических текстов. Предлагается создавать такую систему в среде Web на основе архитектуры Web-сервисов, что позволит системе накапливать опыт и адаптироваться к различным авторам, жанрам и шрифтам. Особенностью системы является использование REST-интерфейса для обмена информацией между сервисами.

1. Введение

Развитие технологий Web-сервисов и распределенных вычислений привело в последние годы к тому, что все больший круг задач, традиционно решавшихся с помощью настольных приложений, решается теперь в среде Web. К таким задачам относится оптическое распознавание текста (OCR).

Преимущество Web-сервисной архитектуры при распознавании текстов заключается не столько в возможности использования современных OCR-систем с различных устройств без установки на них этих систем, или автоматизации процесса пакетного распознавания большого числа страниц без нагрузки на свой компьютер. Их главное преимущество, на наш взгляд, в другом. Вынесение процесса распознавания в среду Web позволяет создавать гибкие, самообучающиеся среды распознавания, накапливающие опыт и способные адаптироваться к особенностям различных текстов. Особенно интересной выглядит перспектива применения таких сервисов для распознавания исторических текстов, в частности, книг XVIII – начала XX вв.

Оставляя в стороне серьезные проблемы улучшения качества изображений и восстановления поврежденных фрагментов текстов, можно выделить две проблемы, с которыми неизбежно сталкиваются исследователи при распознавании печатных исторических текстов – устаревшие шрифты и отсутствие достаточно полных электронных словарей для различных эпох. Так, для русского языка XIX века существует частотный словарь системы «Смалт» [1], содержащий более 50000 словоформ, но он был создан на основе литературно-публицистических текстов, и степень его соответствия лексике газетных статей или мемуаров неизвестна. Поэтому актуальной является задача разработки адаптивных систем распознавания исторических текстов, способных к самонастройке и самообучению – в частности, к пополнению образцов шрифтов и словарей. Подобные проекты существуют (см, например, [2]), однако в них не применяется сервис-ориентированный подход. Цель нашей статьи – представить сервис-ориентированную архитектуру для адаптивной системы распознавания печатных исторических текстов.

2. Архитектура системы

Рассмотрим представленную на рис.1 схему функционирования системы на основе Web-сервисной архитектуры [3]. В ней явно выделяется основной компонент — сервисы хранения исторических документов, задачей которых является постепенное накопление общей информации. Другая часть системы — это некоторые приложения, использующие полученную информацию, которые по своей сути являются вычислительными GRID-службами, созданными для обработки конкретной информации в определённое время. Эти вычислительные сервисы могут появляться и исчезать, временно используя не только вычислительные мощности среды, но и место для хранения промежуточных результатов (частотных словарей и образцов изображений символов различных шрифтов).

Исторический документ в системе может быть представлен в виде метаинформации, изображения документа, распознанного текста, а также дополнительных файлов (например, онтологий). Всё это связанные логически, но разные по формату и назначению электронные документы, поэтому они хранятся в разных подсистемах:

• *Сервис хранения метаинформации* — центральный сервис системы, позволяющий

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

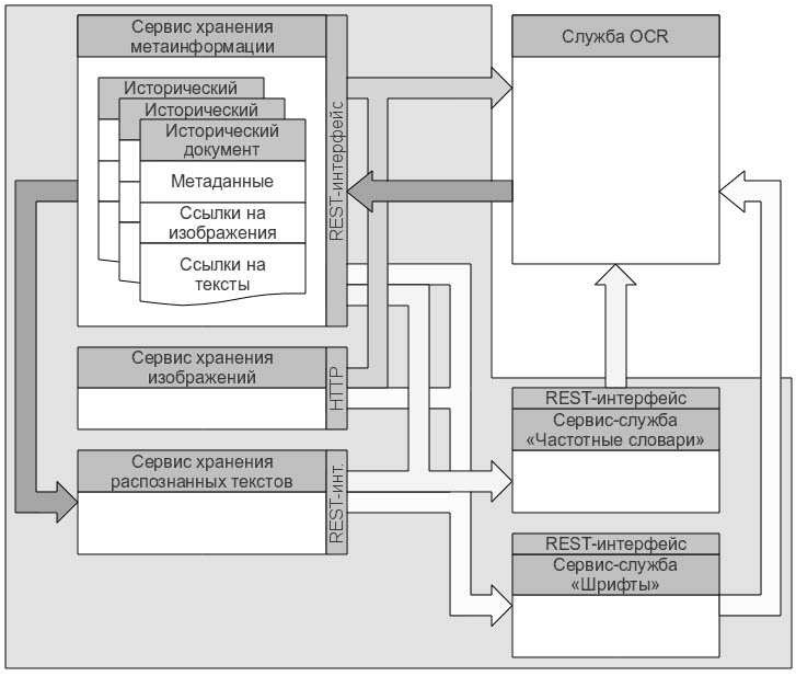

Рис.1. Схема функционирования системы

управлять всей информацией, включая изображения и тексты;

• *Сервис хранения изображений*;

• *Сервис хранения распознанного текста* — эта отдельная специализированная система, в задачи которой входит: хранение распознанного текста с привязкой к исходному изображению, текстовый поиск, построение и хранение частотных словарей.

Кроме сервисов хранения, в общую систему входят дополнительные службы и сервис-службы (вычислительные сервисы в терминологии GRID), среди них выделим:

• *Служба оптического распознавания* — полуавтоматическая (с участием человека как верификатора) или автоматическая система, на вход которой, кроме исходного изображения документа, может подаваться и дополнительная информация, обеспечивающая более качественный перевод (частотные словари и наборы векторных и растровых шрифтов, характерные для данного периода, автора и места публикации).

• *Сервис-служба «Частотные словари»* — в её задачу входит генерация частотных словарей, соответствующих данной группе документов, и их временное хранение.

• *Сервис-служба «Шрифты»* — служба по генерации и хранению векторных или растровых шрифтов, соответствующих данной группе документов

Следует отметить, что построенная архитектура приложения имеет высокий уровень масштабируемости и дальнейшей расширяемости. Особенностью системы является использование REST-интерфейса для обмена информацией между сервисами [4], что позволяет реализовать систему как обычное Web-приложение, без больших временных и финансовых затрат на разработку.

Литература

[1] Сайт проекта «СМАЛТ». http://smalt.karelia.ru

[2] Kluzner V. et al. Word-Based Adaptive OCR for Historical Books // Proceedings of 10th Int. Conf. on Document Analysis and Recognition. Barcelona, 2009. p.501-505.

[3] Alonso G. et al. Web Services: Concepts, Architecture, Applications. Springer, 2004.

[4] Pautasso C.; Zimmermann O.; Leymann F. RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision // Proceedings of 17th International World Wide Web Conference (WWW2008). Beijing, 2008. p.805-814.

Service-Oriented Architecture for Published Historical Texts Recognition

© A. Rudalev, A. Varfolomeyev

This paper describes the project of service-oriented Web information system for published historical texts recognition. That system should be learnable, able to gain experience, adaptable for different authors, genres, publishers, fonts etc. The particular feature of that system is REST-interface for information interchange between components. This architectural decision allows to create the system as usual Web application without substantial costs. The architecture described in this paper is highly scalable and expandable.