Сервис поиска электронных книг

Участники проекта: Жезлов Павел
Мальцев Даниил
Сердюк Дмитрий
Туркин Александр

Руководитель проекта: Тузова Екатерина

Академия современного программирования, 2010-2011

eBookSearch
ebook search service

Проблемы

- Существует множество интернет-библиотек – пользователю надо искать по всем
- Доступность для мобильных устройств
- Проблема поддержки форматов
- Нет удобного инструмента для создания книг в формате ePub

eBookSearch
ebook search service

Решения

- Поиск сайтов с книгами
- Агрегация данных с сайтов
- Аналитический классификатор
- Специальная система жанров
- Конвертор txt, html, doc → ePub
- Предоставление доступа в HTML/OPDS

eBookSearch
ebook search service

Использованные средства

- Python (язык разработки)
- Django (веб-фреймворк)
- Sphinx (быстрый поиск)
- wvWare (конвертация)
- BeautifulSoup (парсер HTML/XML)
- Stemmer (анализ текста)

eBookSearch
ebook search service

Существующие сервисы

Google Книги http://books.google.com

- Ориентированность на продажу
- Отсутствие OPDS
- Неудобный доступ с мобильных устройств

Electronic Books Database http://ebdb.net

- Отсутствие OPDS
- Непрямые ссылки

Bookserver http://bookserver.archive.org

- Есть OPDS
- Отсутствие HTML интерфейса
- Отсутствие книг на русском

eBookSearch
ebook search service

Структура проекта

- Crawler
  поиск сайтов, содержащих электронные книги, и отдельных книг
- Analyser
  агрегация информации о книгах с сайтов-библиотек
- Classifier
  классификация книг по жанрам, подбор рекомендаций
- Epuber
  преобразование книг в формат Epub
- Веб-интерфейс
  предоставление доступа к книгам (HTML/OPDS)

eBookSearch
ebook search service

Crawler

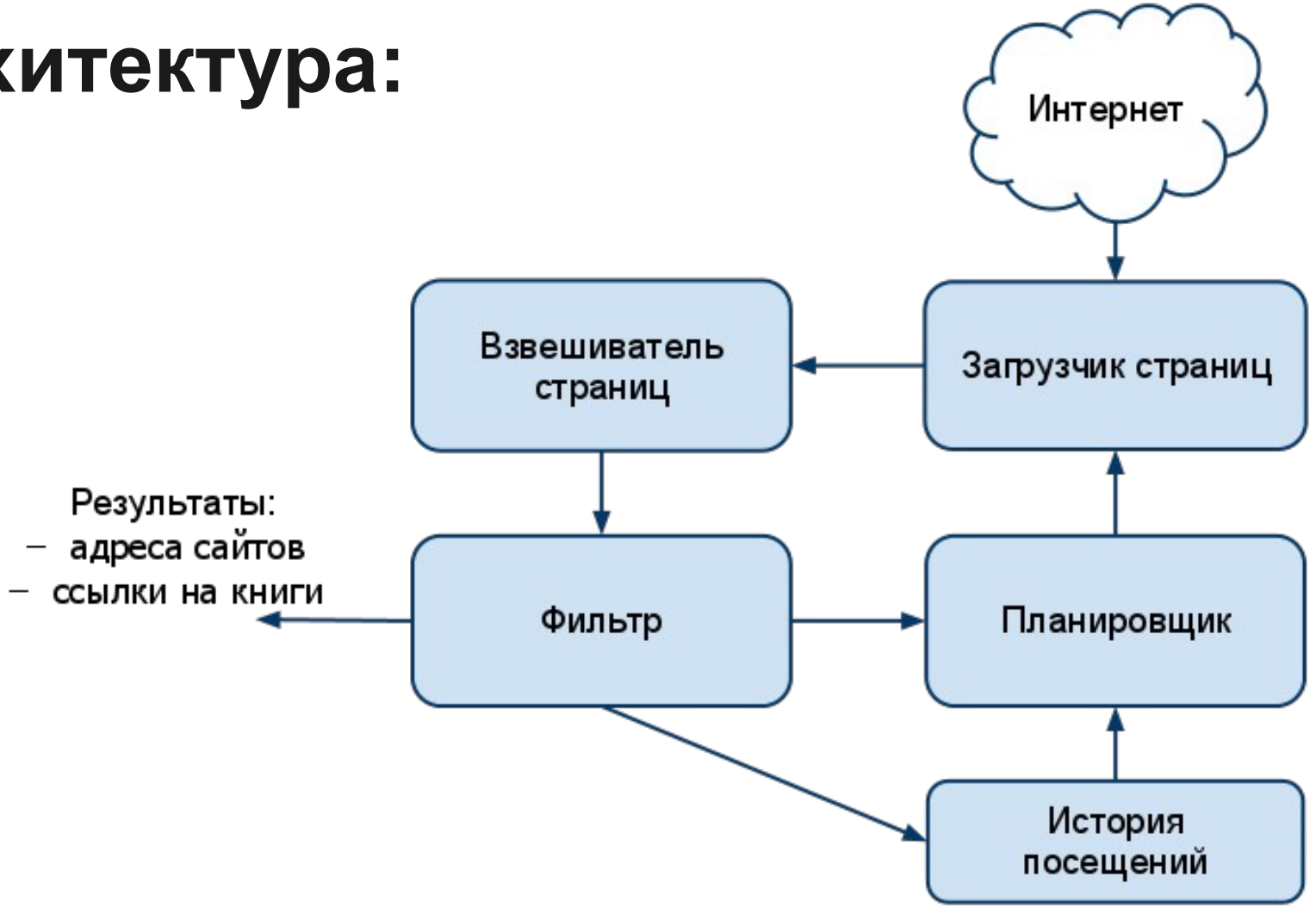
Решаемые задачи:

- поиск сайтов с книгами для Analyser'a
- игнорирование интернет-магазинов
- поиск отдельных книг (ссылки вида *.epub, *.fb2)

eBookSearch
ebook search service

Crawler

Архитектура:



eBookSearch
ebook search service

Crawler

Настройка:

- «избирательность» фильтра страниц
- тщательность поиска
- словари
- использование ресурсов

eBookSearch
ebook search service

Crawler

Развитие:

- иные методы оценки страниц
- смена стратегии планировщика
- самообучение словарей

eBookSearch
ebook search service

Analyser

Решаемые задачи:

- Обход сайтов
- Агрегация данных
- Обработка данных

Собираемая информация:

название, ссылки, авторство, аннотация, язык, жанры, обложка

eBookSearch
ebook search service

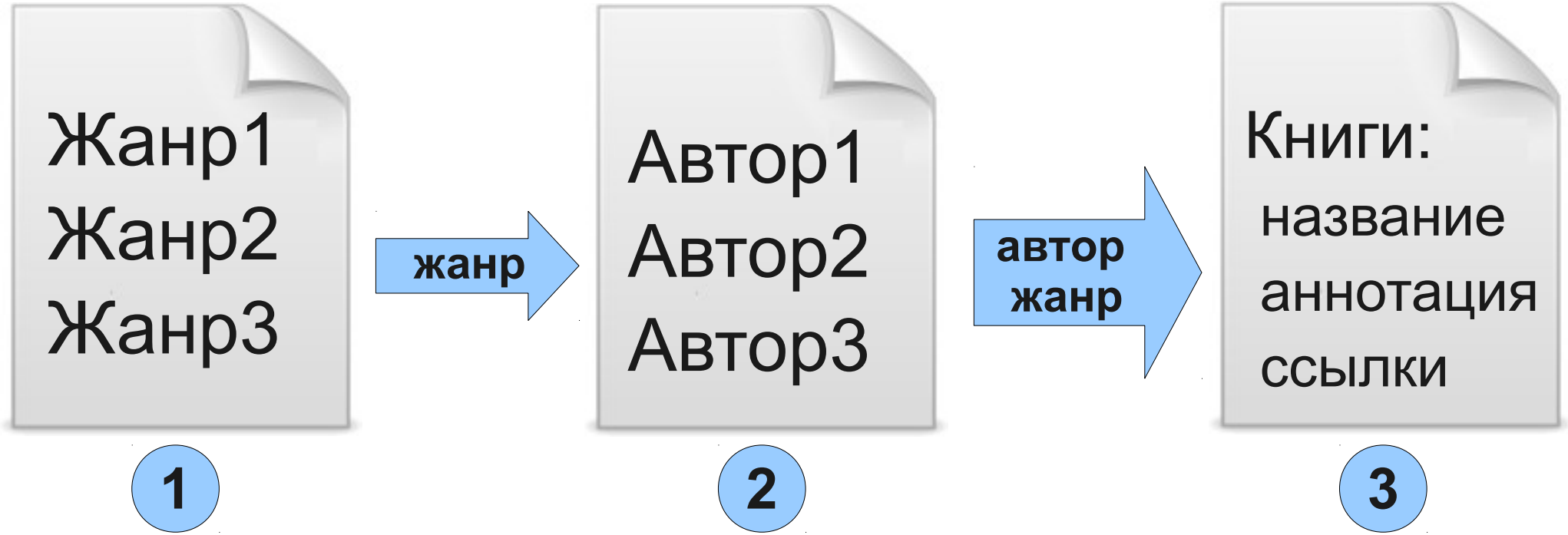
Кратко об OPDS

- OPDS = **O**pen **P**ublication **D**istribution **S**ystem
- На базе языка разметки Atom
- Для информации об электронных документах
- Структура: Acquisition Feeds, Navigation Feeds
- Поддержка каталогов: FBReader, Stanza

eBookSearch
ebook search service

Analyser

Архитектура парсера



Состав парсера:

- порядок обхода
- местоположение данных на странице

eBookSearch
ebook search service

Analyser

Особенности:

- Standalone
- Агрегация любых данных
- Легкое добавление нового сайта
- Тесты для проверки актуальности парсера
- Удобная система отладки обхода
- Интеллектуальная обработка

eBookSearch
ebook search service

Analyser

Индексируемые сайты:

- http://lib.ru
- http://nehudlit.ru
- http://magazines.russ.ru
- http://flibusta.net
- http://manybooks.net
- http://smashwords.com
- http://feedbooks.com
- http://data.fbreader.org

больше

100 000

книг

~2 000

выпусков журналов

eBookSearch
ebook search service

Classifier

Определение жанров по аннотациям.

- Создание структуры жанров и поджанров.
- Классификация по поджанрам:
  - Определение частоты встречаемости слов и пар слов.
    - Использование стемминга
    - Отбрасывание stopwords
  - Определение вероятности принадлежности аннотации к каждому жанру.

eBookSearch
ebook search service

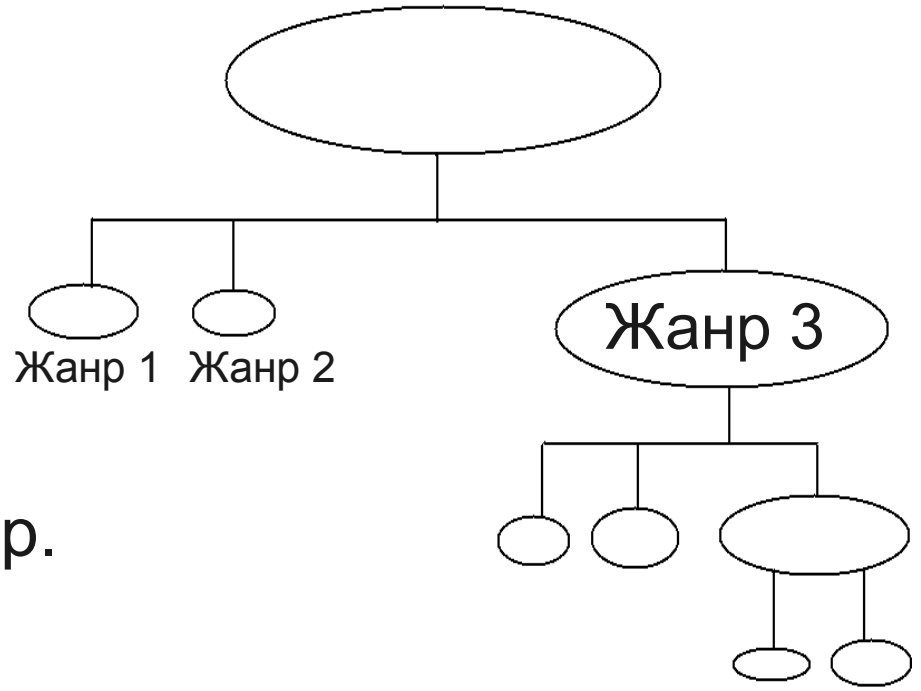
Classifier

Подбор рекомендаций

Для каждой аннотации есть вероятность принадлежности к каждому жанру.

Оптимизация:

- Разбивка всех аннотаций на группы по наиболее вероятным жанрам, пока не станет достаточно мало.
- В подгруппах — полный перебор.



eBookSearch
ebook search service

Справочник биохимика

by Досон Р. Эллиот Д. Эллиот У. Джонс К.

Язык: ru

Жанры: справочная литература химия наука, образование

Описание

Книга содержит данные о физико-химических и биологических свойствах биологически
методические указания для проведения стандартных биохимических процедур.

Загрузки: djvu

Похожие книги:

Механизмы реакций в органической химии
Основы органической химии лекарственных веществ
Биоэнергетика и линейная термодинамика необратимых процессов (стационарное состоя
Биохимия
Принципы структурной организации нуклеиновых кислот
Биохимия. Т. 1
Основы биохимии. Т. 1
Биохимия мембран. Рецепторы клеточных мембран
Биохимия мембран. Кинетика мембранных транспортных ферментов
Основы энзимологии

Пример рекомендации книг

eBookSearch
ebook search service

Classifier

Развитие

- Использование пользовательских рекомендаций
- Поиск дополнительной информации

eBookSearch
ebook search service

Epuber

Решаемые задачи:

- конвертация txt(html) файлов в ePub
- конвертация doc файлов в ePub

eBookSearch
ebook search service

Epuber

Конвертация txt и html файлов в ePub

- OPDS для lib.ru
- журналы в виде единого файла

Epuber

Конвертация doc файлов в epub

- используется утилита wvWare
- строится оглавление по тексту
- файл разбивается на главы и сжимается

eBookSearch
ebook search service

Ссылки

Ссылка на проект:
http://service.ebooksearch.webfactional.com/

Код проекта:
http://code.google.com/p/ebook-service/

Спасибо за внимание!