



Разработка прототипа автоматизированной системы поиска дубликатов документов для цифровых научных библиотек

Романов Максим Владимирович 11-502

Научный руководитель:

Елизаров Александр Михайлович

Проблема

Проблема проверки уникальности научных документов и нахождения их дубликатов в контексте электронных научных библиотек

1. Новый документ
2. Проверка на дубликаты
3. Добавление/отклонение документа

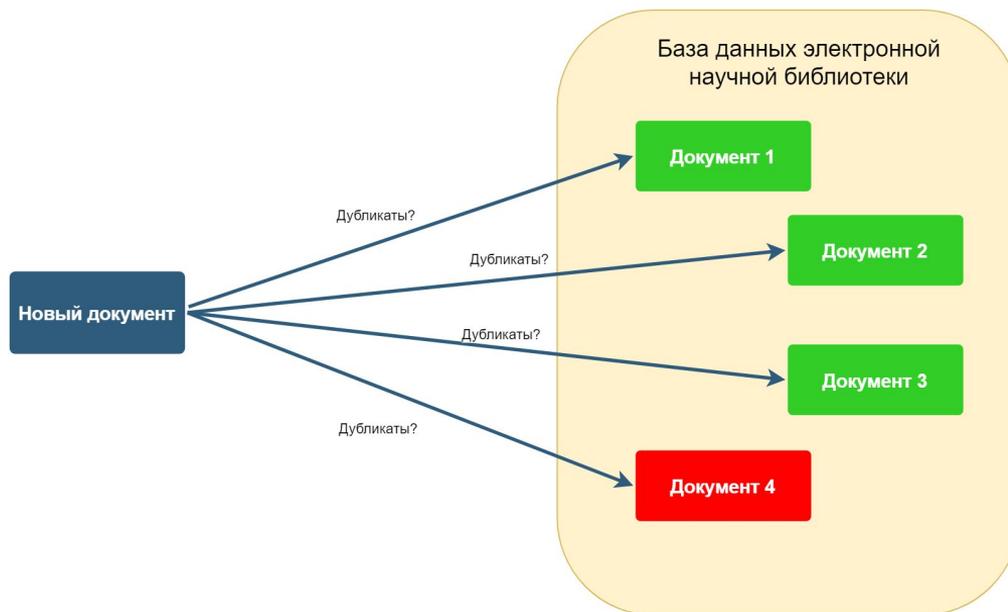


Рис. 1. Добавление нового документа

Цель и задачи

Цель: разработка сервиса поиска дубликатов в электронных научных библиотеках.

Задачи:

1. Исследовать способы организации данных в электронных научных библиотеках
2. Рассмотреть существующие алгоритмы поиска нечетких дубликатов текста и определить наиболее подходящий данной задаче
3. Разработать систему поиска дубликатов в электронных научных библиотеках

Существующие решения

Алгоритм “шинглов”:

- Физическое представление данных
- Точность ~91%
- Неустойчив к мелким изменениям
- Неустойчив к перестановкам слов

Отсутствие возможности добавления документов в базу данных сервиса

Предлагаемое решение

1. Алгоритм TF–RIDF:

- Точность ~95%
- Учитывает статистику всей коллекции
- Устойчив к мелким изменениям
- Устойчив к перестановкам слов

2. Сбор данных:

- Интерактивная индексация библиотек
- Добавление/расширение данных

Технологии

- Серверная часть:
 - Язык программирования – Java
 - Сервер – Spring Boot
 - Многопоточность – Concurrent, Guava
 - Агрегация данных – Stream API
 - Доступ к базе данных – Spring-jdbc
- Клиентская часть:
 - Разметка – HTML
 - Скрипты – Javascript
- База данных:
 - СУБД – PostgreSQL



Рис. 2. Технологии

Результаты (I часть)

Индексация документов электронных научных библиотек:

- Рекурсивный обход ссылок
- Диапазон ссылок

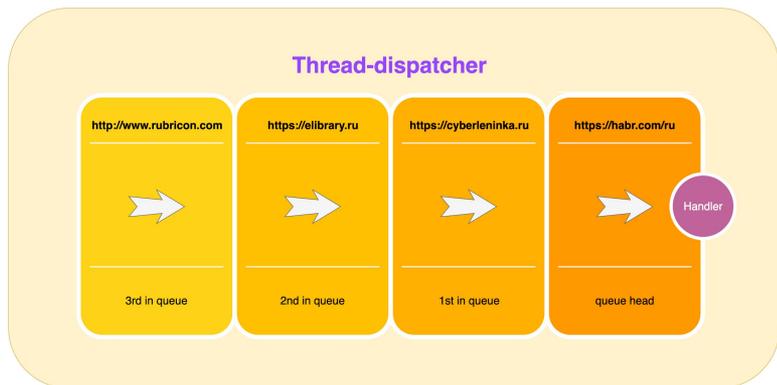


Рис. 4. Очередь индексации

The screenshot shows a web interface for library indexing. It features a table with the following data:

Queue head	
Library	https://habr.com/ru/
Parsed documents	560
All found documents	1200
Status	PROCESSING

Below the table, there is a section for "Index library" with an input field "Enter URL" and an example "https://habr.com/ru/".

Next is a "Clarification" section with an input field "Enter clarification" and an example "post". A blue "Index" button is located below this section.

The final section is "Remove from queue" with an input field containing the ID "40b30de0-7e5f-11e9-aeb1-0776811caaae" and an example "037bdd80-7e55-11e9-a45c-0b74479b39e1". A blue "Remove" button is located below this section.

Рис. 3. Интерфейс индексации библиотек

Результаты (II часть)

Проверка документов на наличие дубликатов:

- Сбор слов
- Вычисление значимости слов
- Сравнение контрольных сумм

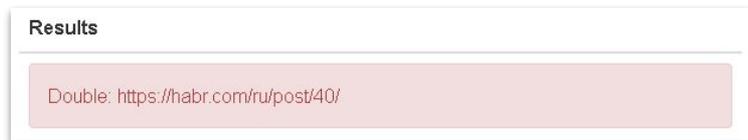
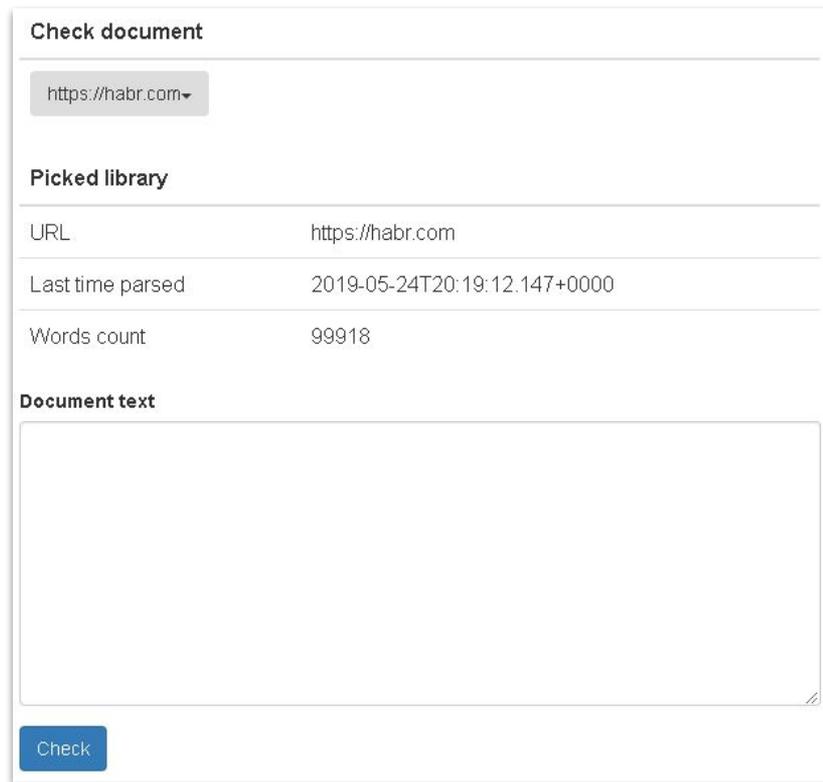


Рис. 6. Дубликат найден



Рис. 7. Дубликатов не найдено



The interface for checking a document. It includes a text input field with the URL 'https://habr.com', a 'Picked library' section with a table of metadata, and a 'Document text' section with a large empty text area. A blue 'Check' button is located at the bottom left.

Picked library	
URL	https://habr.com
Last time parsed	2019-05-24T20:19:12.147+0000
Words count	99918

Рис. 5. Интерфейс проверки документа

Производительность

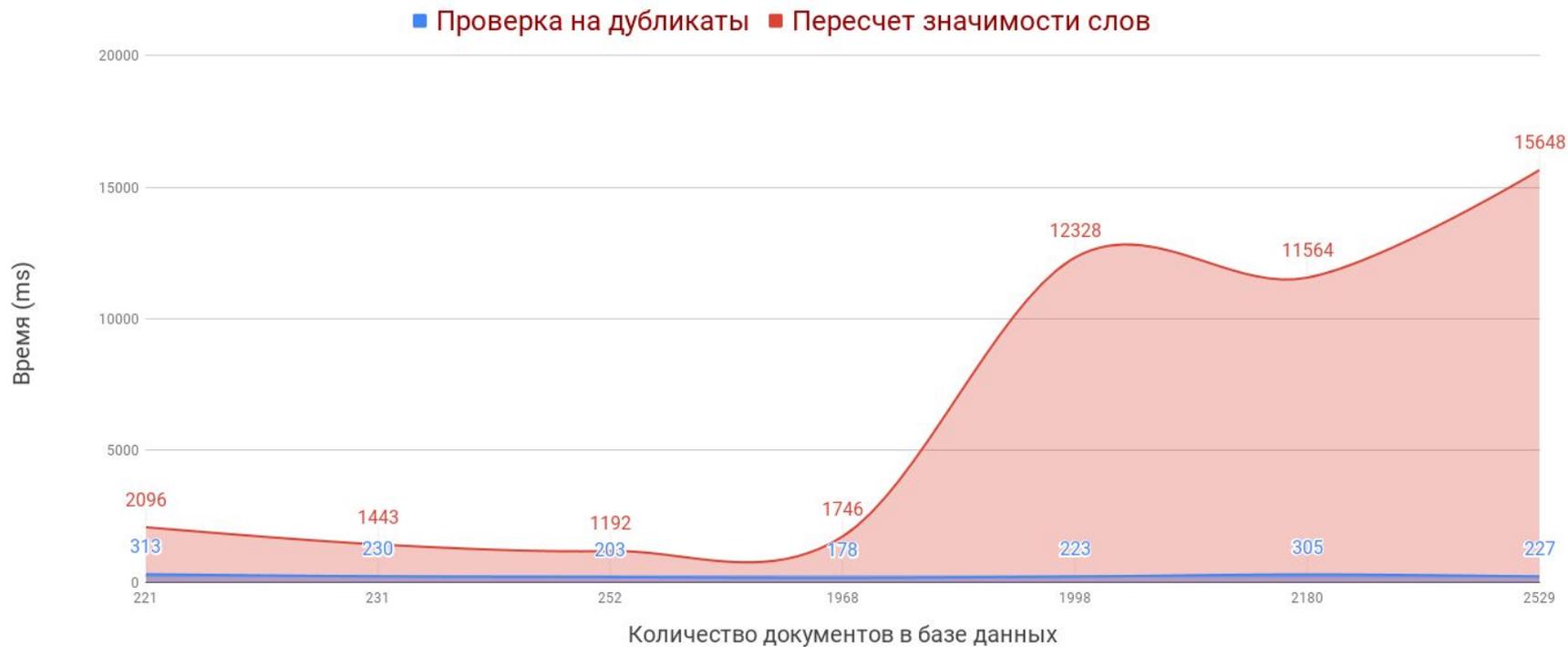


Диаграмма 1. Тест производительности

Выводы

Свойства системы:

- Алгоритм TF–RIDF
- Индексация электронных научных библиотек
- Быстрая проверка на дубликаты ~200ms



Разработка прототипа автоматизированной системы поиска дубликатов документов для цифровых научных библиотек

Романов Максим Владимирович 11-502

Научный руководитель:

Елизаров Александр Михайлович