Дипломная работа по направлению «Проектирование и технология производства ЭС»

Система интеллектуального поиска информации в электронных архивах конструкторско-технологической документации

Петухов А. М. ИУ4, 2007 г.

Научный руководитель: доцент, к. т. н. Власов А. И.

Цели и задачи работы

Цель работы:

Целью настоящей работы является разработка информационно-поисковой систем для использования на радиотехнических предприятиях, исследование математических методов информационного поиска, разработка и экспериментальное исследование информационно-поисковой системы поиска текстовой информации.

Решаемые задачи:

- Анализ и систематизация задач, решаемых информационно-поисковыми системами.
- Систематизация и сравнительный анализ существующих информационно-поисковых систем
- Анализ математических методов расчета релевантности документа пользовательскому запросу и исследование применимости различных стратегий поиска для распределенных источников данных
- Исследование применения статистических подходов к кластеризации документов
- Исследование и выбор методов аппаратной и программной реализации комплекса поиска полнотекстовой информации и разработка системы в виде аппаратно-программного комплекса
- Экспериментальное исследование эффективности предложенных алгоритмов и методов реализации комплекса

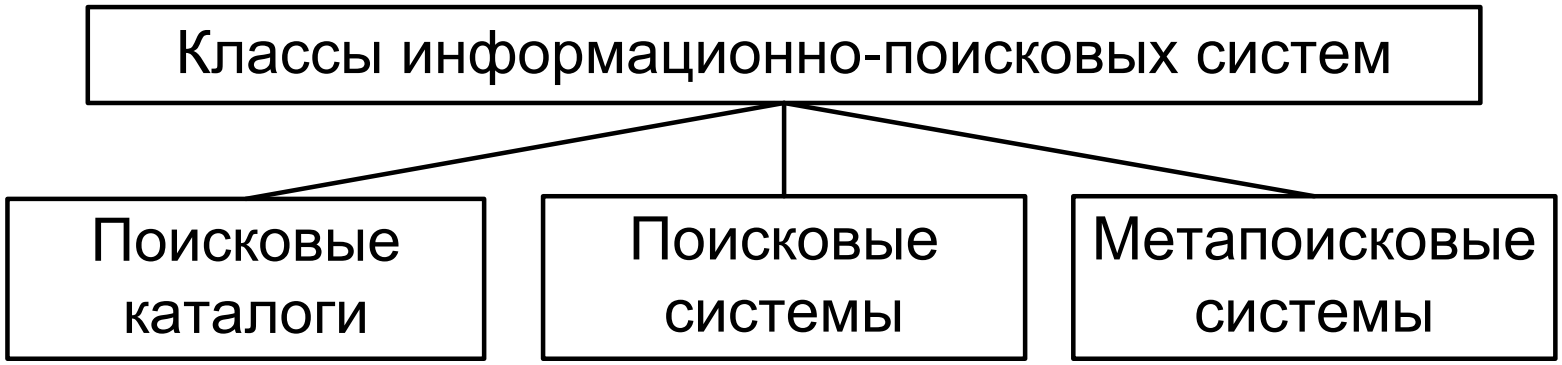
Постановка задачи

Существующие проблемы

Управление современными предприятиями не может быть эффективным без использования новых информационных технологий обеспечения руководителей производства необходимыми средствами оперативного управления финансово-хозяйственной деятельностью, оптимального планирования ресурсов и управления технологическими процессами, интегрированными в единую информационно-управляющую систему.

- **Большие объемы полнотекстовой конструкторско-технологической документации**
- **Большие объемы сопутствующей документации – контракты, договора, руководства по эксплуатации, описания покупных изделий**
- **Отсутствие единой системы хранения документации**
- **Необходимость повышения скорости проектирования и поиска новых технических решений**

Классификация ИПС



- **Поисковые каталоги** - ориентированны на структурную организацию тематических коллекций с иерархией документов по тематическим коллекциям.

- **Поисковые системы** - ориентированны на поиск слабоструктурированной информации. Особенностью является отсутствие тематической организации

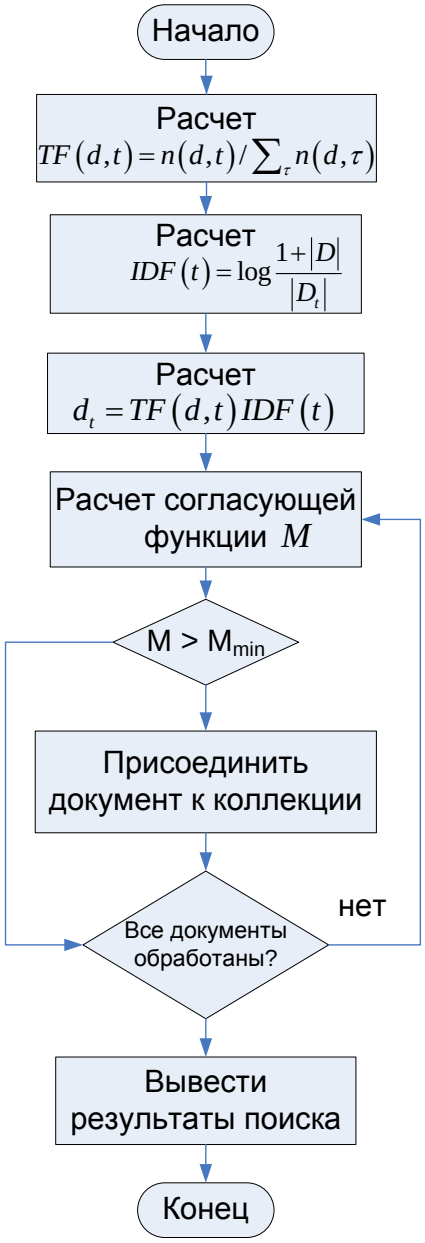
- **Метапоисковые системы** - ориентированны на интеграцию результатов поиска от различных поисковых систем.

Модель векторного пространства и оценка релевантности

Релевантность - мера логической близости результата поиска к запросу пользователя. При поиске документов в коллекции целью поиска является получение всех релевантных документов и неполучение нерелевантных

Документы - векторы в многомерном Евклидовом пространстве. Каждая ось в таком пространстве соответствует одному их ключевых слов индекса.

Для того чтобы можно было оценить релевантность того или иного документа запросу, необходимо оценить близость между векторами и $\vec{d}$ $\vec{q}$



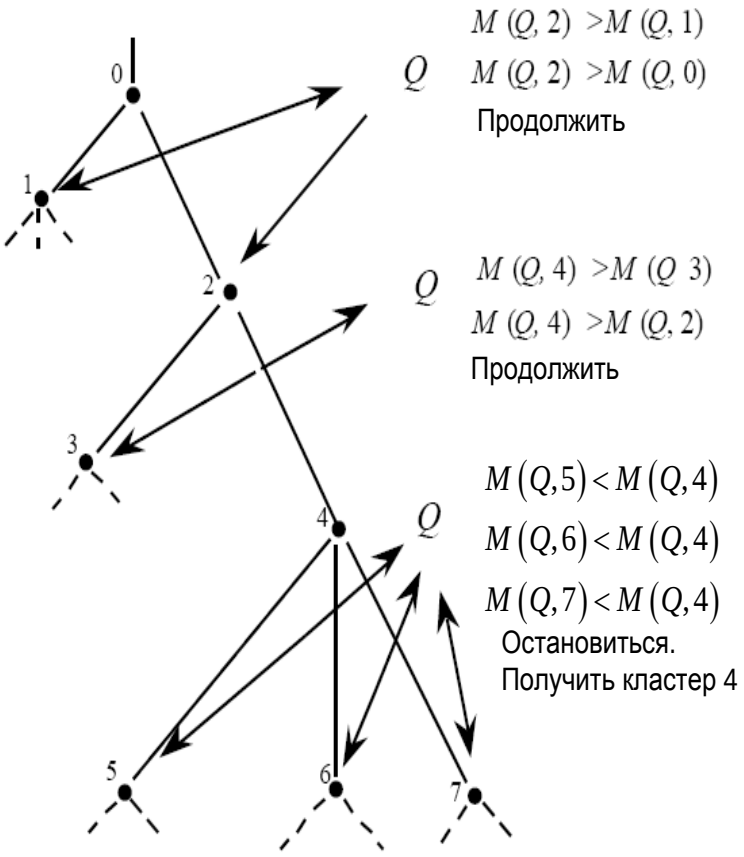
Стратегии поиска

-*Поиск, основанный на использовании Булевой алгебры* - стратегия поиска, основанная на использовании Булевой алгебры, заключается в получении только тех документов, которые «истинны» для пользовательского запроса.

-*Последовательный поиск* – вычисление N значений согласующих функций и выбор документов

-*Поиск на основе кластеров* – производится аналогично последовательному поиску, но расчет согласующей функции производится не для каждого документа, а для кластера

Поиск на основе кластеров



Применение кластеризации в задачах информационного поиска

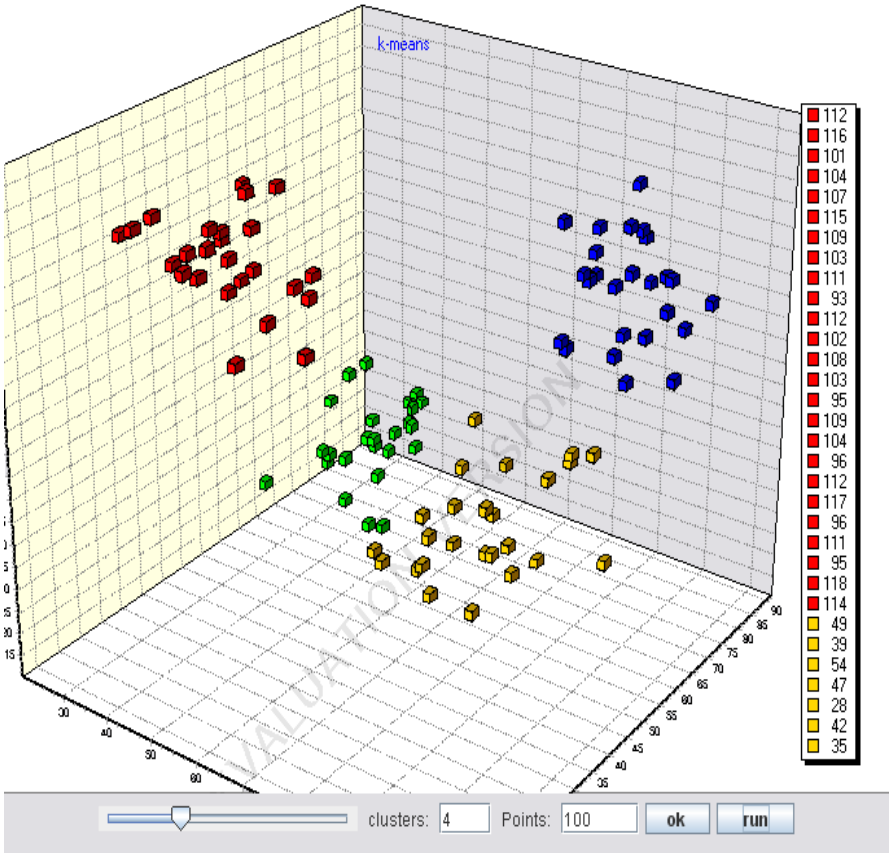
Кластерная гипотеза

При кластеризации коллекции, если пользователь заинтересован в документе *d*, он также будет заинтересован в остальных документах кластера, к которому принадлежит *d.*

Применение кластеризации

- Визуализация результатов поиска
- Поиск подобных документов
- Реализация кластерной стратегии поиска
- Повышение точности результатов поиска
- Подготовительный этап для построения информационных каталогов
- Пополнение информационных каталогов
- Формирование каталога поисковых терминов

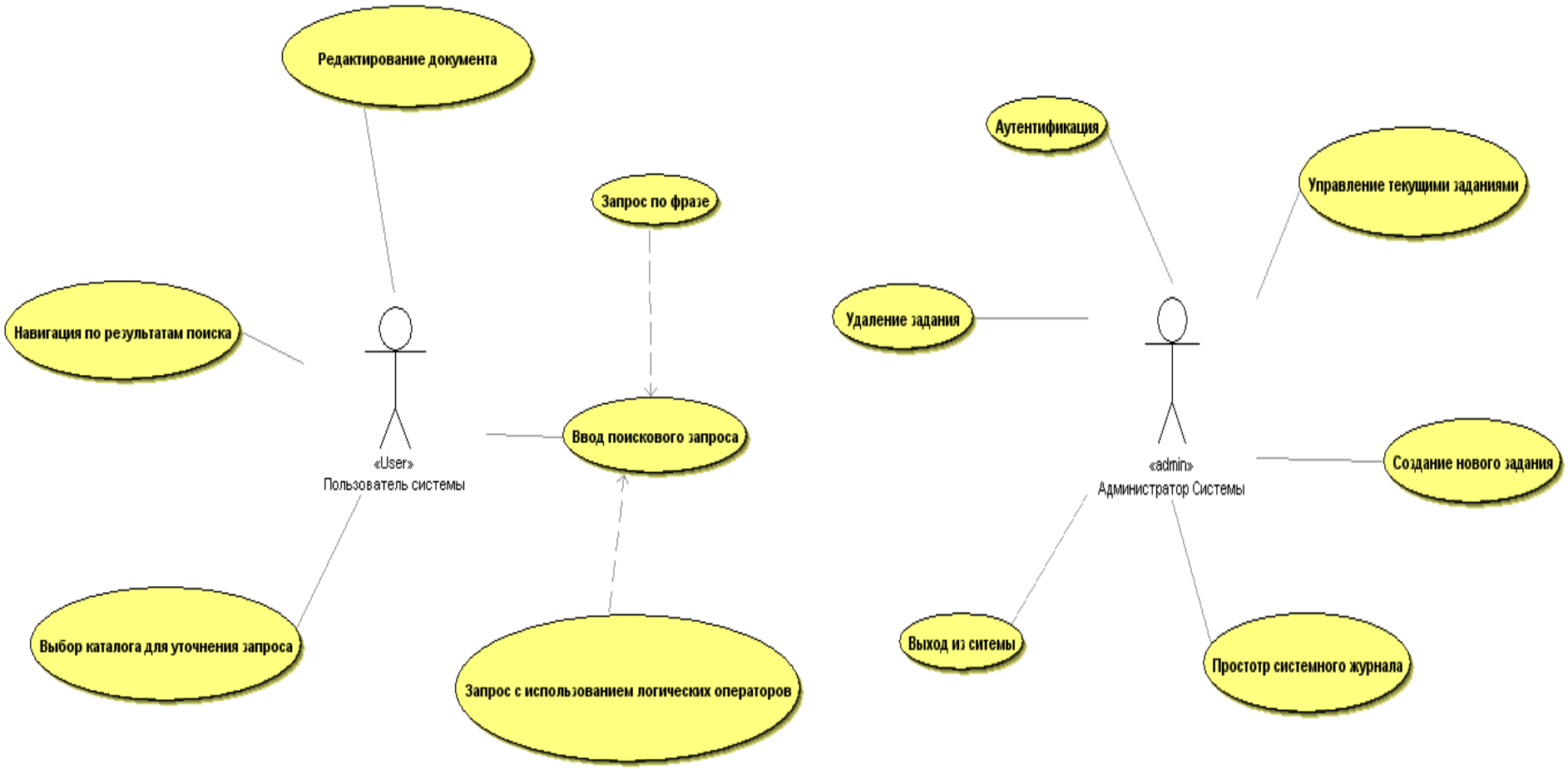
Визуализация результатов кластеризации



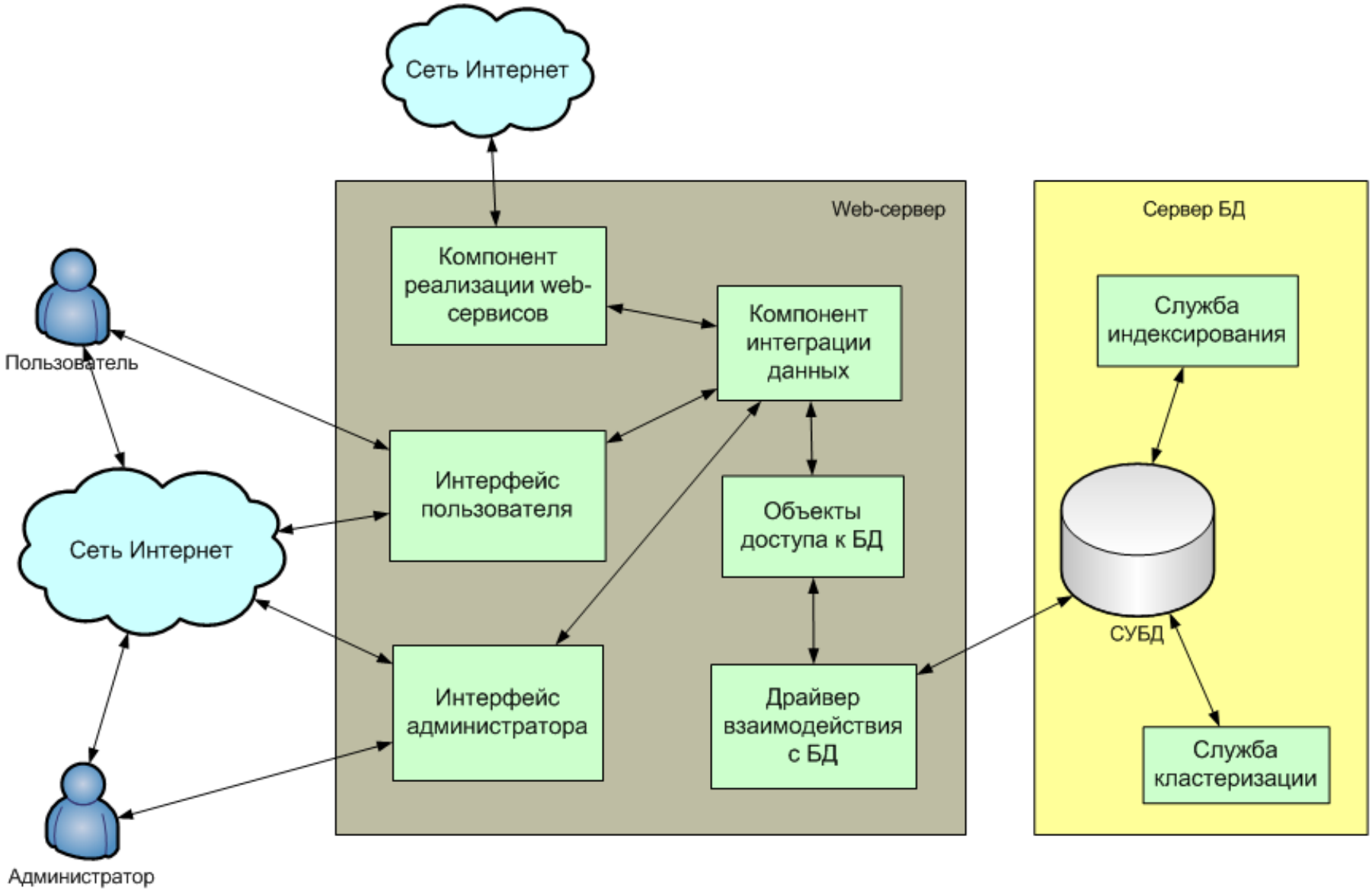
Функциональные требования к Системе

Пользователь системы

Администратор системы



Структурно-функциональная схема системы



Реализация Системы

Технология реализации:
тонкий клиент

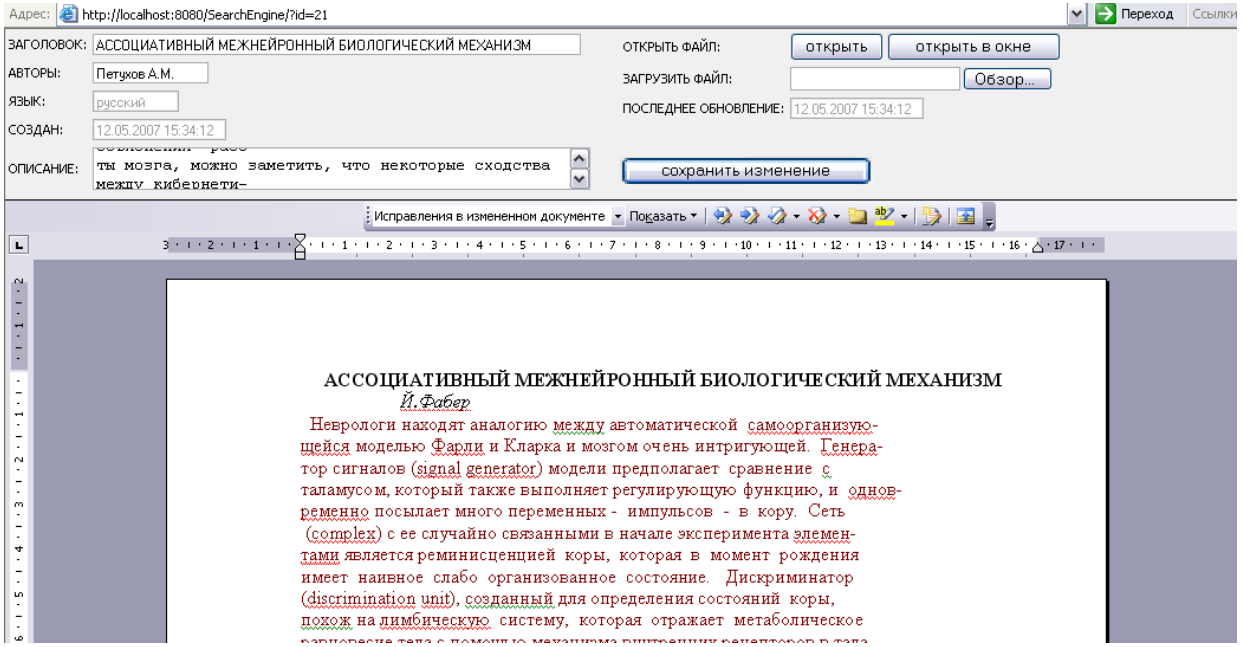
Язык программирования:
Java

СУБД:
Oracle 10g

Сервер приложений:
Apache Tomcat 5.5

Форма ввода запроса

Просмотр документа



Особенности реализации:
- Кроссплатформенность
- Расширяемость
- Масштабируемость
- Использование открытых стандартов
- Использование бесплатных компонентов и библиотек

Экспериментальное исследование Системы

Целью исследований является:

- Анализ динамических характеристик Комплекса и его поведения в условиях реальной нагрузки
- Формирование рекомендаций по применению использующегося математического аппарата в задаче информационного поиска

Задачи

- Разработка плана экспериментальных исследований
- Разработка методики проведения эксперимента
- Построение экспериментального стенда
- Проведение экспериментальных исследований
- Оценка результатов экспериментов.

Характеристики тестовой коллекции

Тематика	Конструкторско-технологическое проектирование электронно-вычислительной техники
Количество документов	80 документов
Средний размер документа	4000 слов
Размер коллекции	10 Мб
Разметка документа	RTF
Язык	Русский

Характеристики экспериментального стенда

Центральный процессор	AMD Athlon XP 2800+ (1.6 ГГц)
ОЗУ	1.00 Гб DDR
HDD	60 Гб SATA
Контроллер ЛВС	10/100/1000 Base-T
Операционная система	MS Windows XP SP2
Сервер приложений	Apache Tomcat 5.5.16 Server

Заключение

Результаты

- Разработана классификация систем информационного поиска, рассмотрены принципы построения поисковых систем. Проведен сравнительный анализ существующих поисковых систем
- Рассмотрены основные математические модели представления документов, а также методы оценки релевантности документов
- Проведен анализ математических методов повышения точности поиска, рассмотрены методы кластеризации текстов и сформулированы рекомендации по их применению к задачам информационного поиска
- Разработана Система поиска и обработки полнотекстовой документации, позволяющий пользователям осуществлять поиск необходимых документов в распределенных источниках данных
- Проведены экспериментальные исследования разработанной Системы и предложены меры по повышению его производительности и надежности