Московский Государственный Технический Университет им. Н. Э. Баумана

Диссертация магистра по направлению «Проектирование и технология производства ЭС»

Комплекс поиска и обработки гипертекстовой информации в распределенных источниках данных

Смирнов А. С.
ИУ4, 2006 г.

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

Цели и задачи работы

Цель работы:

Исследование методов поиска информации и разработка информационной системы поиска и обработки гипертекстовой информации в распределенных источниках данных

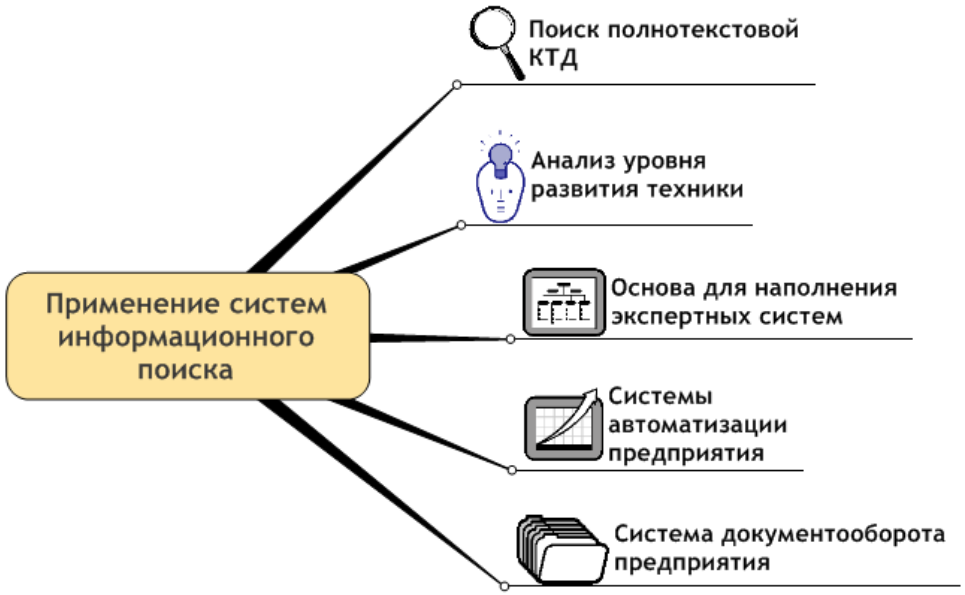
Решаемые задачи:

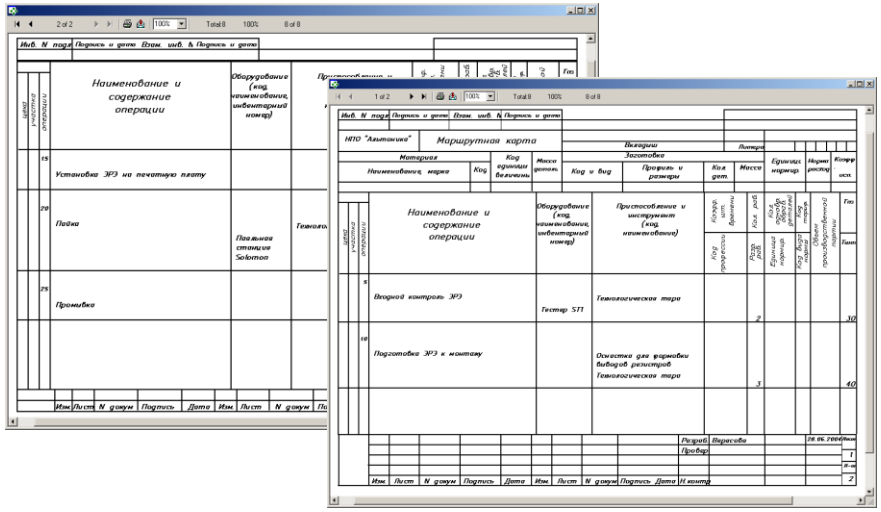
- Анализ и систематизация задач, решаемых информационно-поисковыми системами.
- Систематизация и сравнительный анализ существующих информационно-поисковых систем
- Анализ математических методов расчета релевантности документа пользовательскому запросу и исследование применимости различных стратегий поиска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

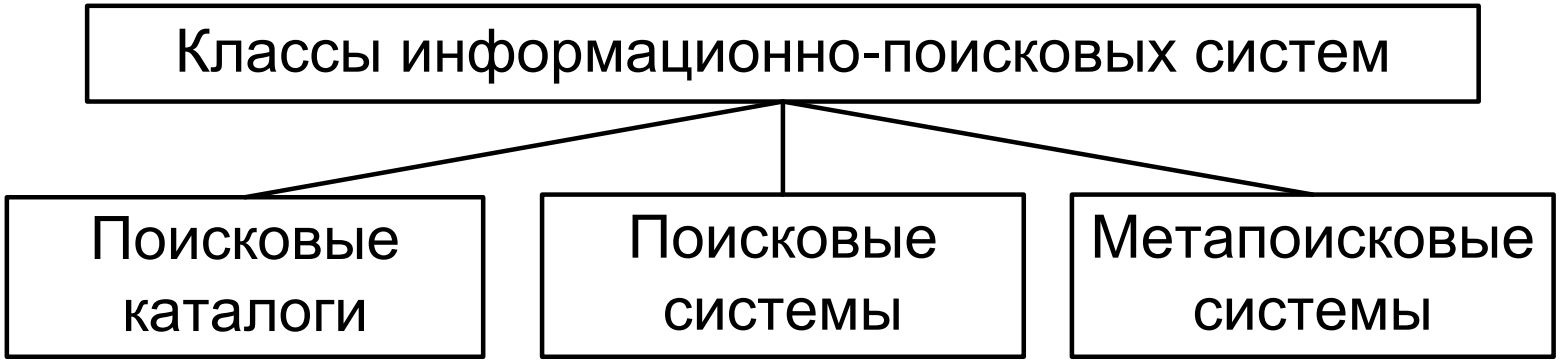
Существующие проблемы

- Большие объемы (до нескольких терабайт) полнотекстовой конструкторско-технологической документации:
  - Спецификации
  - Ведомости
  - Записки
  - Таблицы
  - Расчеты
  - Технологические инструкции
  - Карты ТП
  - …
- Большие объемы сопутствующей документации – контракты, соглашения, инструкции по эксплуатации, …
- Отсутствие единой системы хранения документации
- Низкая компьютерная подготовленность персонала
- Необходимость повышения скорости поиска технических решений

Области применения информационно-поисковых систем



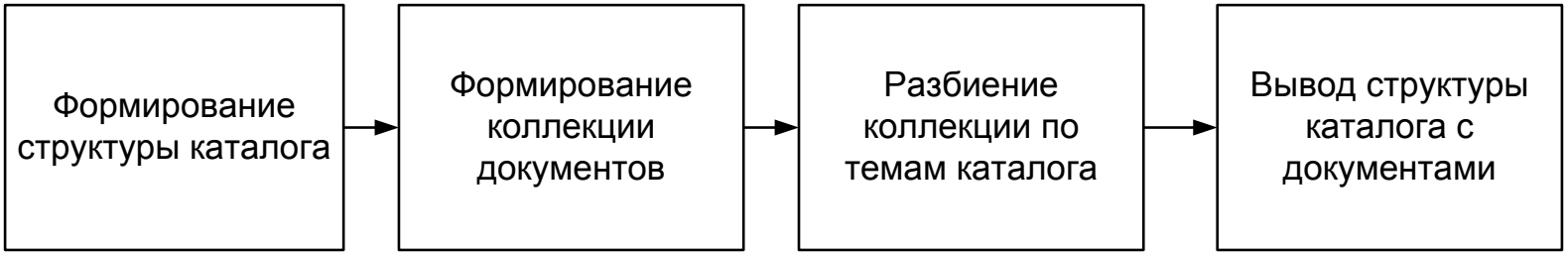




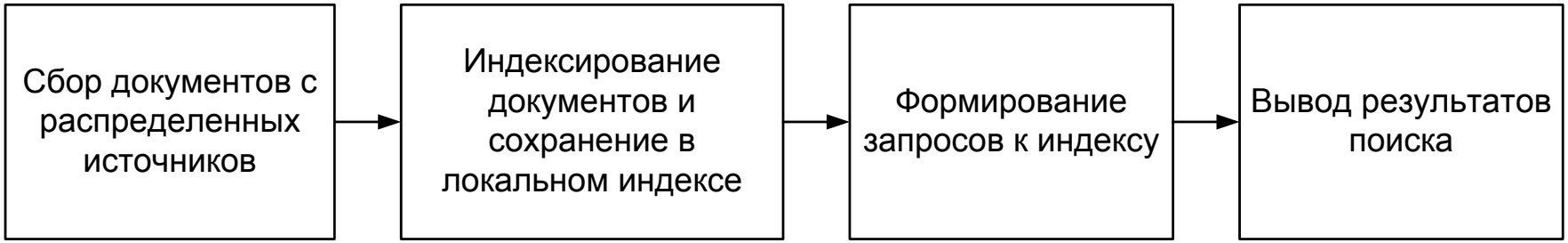
Сравнительный анализ классов информационно-поисковых систем

Тип ИПС	Особенности применения	Достоинства	Недостатки
Поисковые каталоги	Ориентированны на структурную организацию тематических коллекций с иерархией документов по тематическим коллекциям.	- быстрый поиск сведений по определённой крупной теме; - содержат качественную информацию, отобранную разработчиками	- не могут дать исчерпывающих сведений по определённой тематике; - Несовершенная структура каталога
Поисковые системы	Ориентированны на поиск слабоструктурированной информации. Особенностью является отсутствие тематической организации	- Большой объем доступных документов; - Предоставляют более широкие возможности по поиску информации; - Информация постоянно обновляется	- Более сложны к использованию - Отсутствует разбиение документов по группам и тематике
Метапоисковые системы	Ориентированны на интеграцию результатов поиска от различных поисковых систем.	- Возможность поиска сразу в нескольких поисковых системах - Агрегирование результатов из нескольких поисковых систем	- Отсутствует разбиение документов по тематике - Несовершенные алгоритмы ранжирования результатов

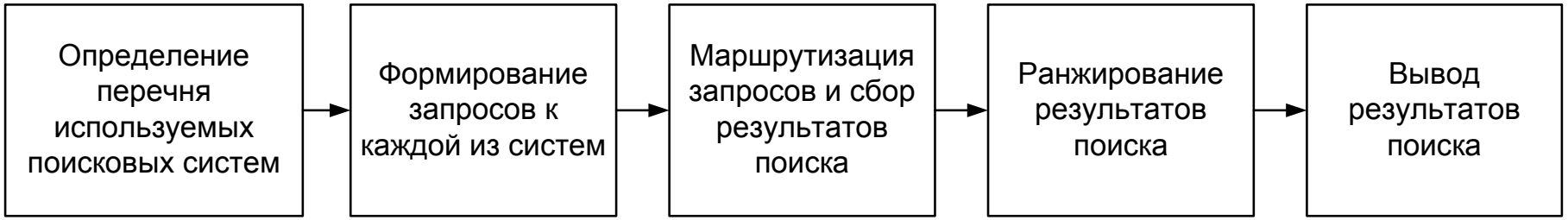
Поисковый каталог



Поисковая система



Метапоисковая система



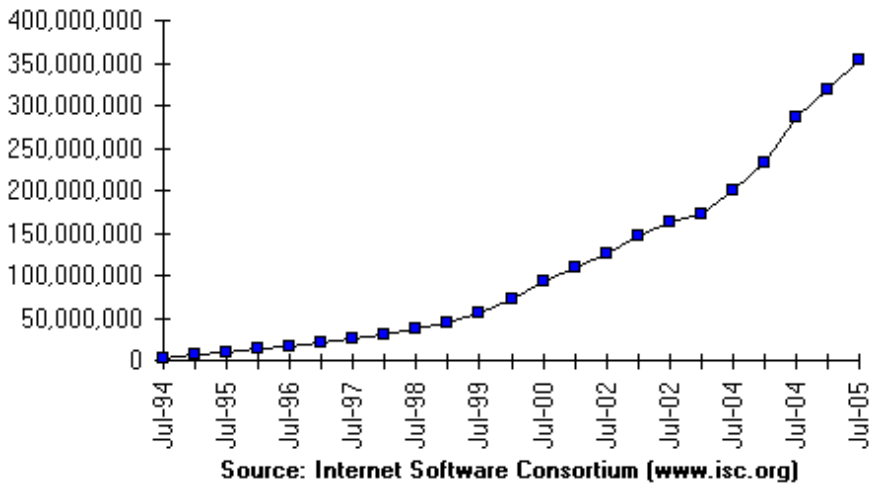
Тенденции развития ИТ-инфраструктуры:

- Переход к использованию Web-технологий
- Внедрение систем управления документами
- Увеличение роли сети Интернет в деятельности предприятия
- Интеграция с информационными системами других предприятий

Механизмы повышения доступности:

- Единая схема наименования для поиска ресурсов в Web (URI)
- Протоколы для доступа к именованным ресурсам через Web (HTTP)
- Гипертекст для простого перемещения по ресурсам (HTML)

Рост числа серверов



Особенности сети Интернет с точки зрения информационного поиска:

- Размер
- Динамика
- Взаимосвязи
- Свободная публикация
- Избыточность
- Неконтролируемое качество
- Пользователи
- Доступ
- Многоязычность

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 $\vec{d}$ и $\vec{q}$

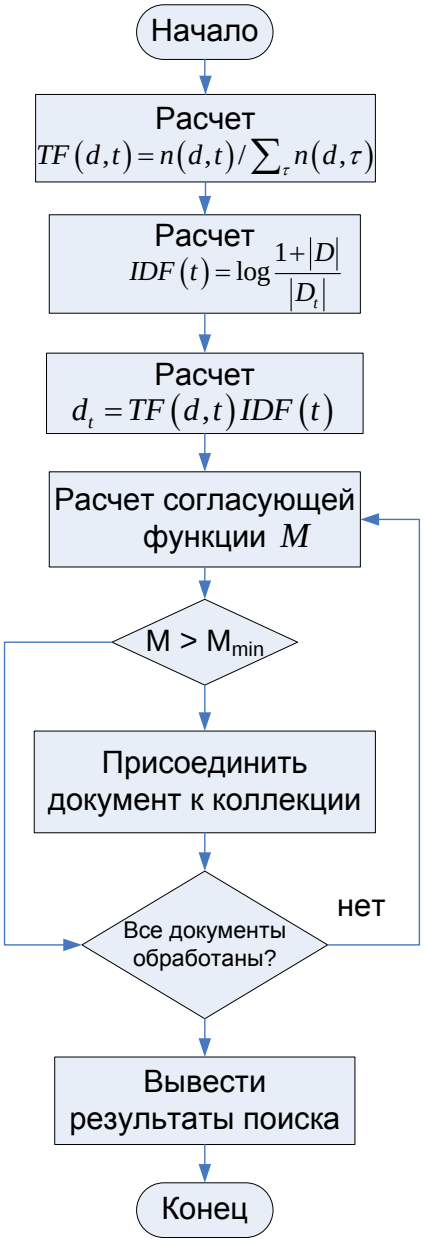
Примеры согласующих функций

- Коэффициент Диса

$$M = \frac{2|D \cap Q|}{|D| + |Q|}$$

- Косинусная корреляция

$$r = \frac{(Q \cdot D)}{\|Q\| \cdot \|D\|} = \cos\theta$$



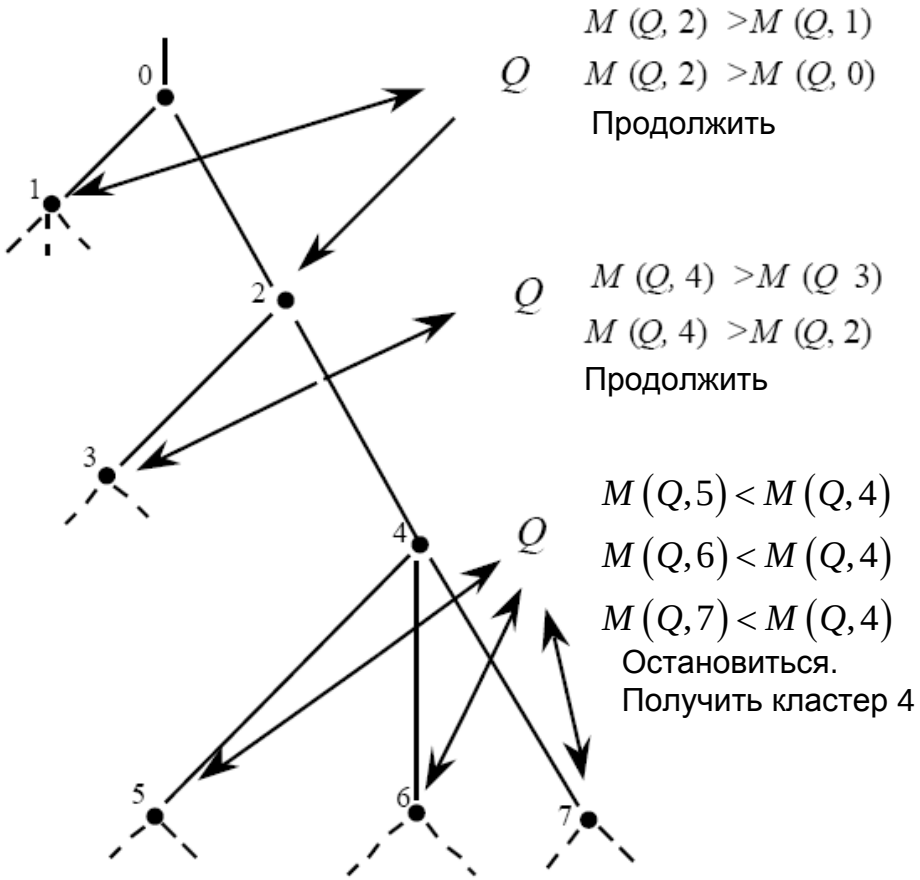
Стратегии поиска

-*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



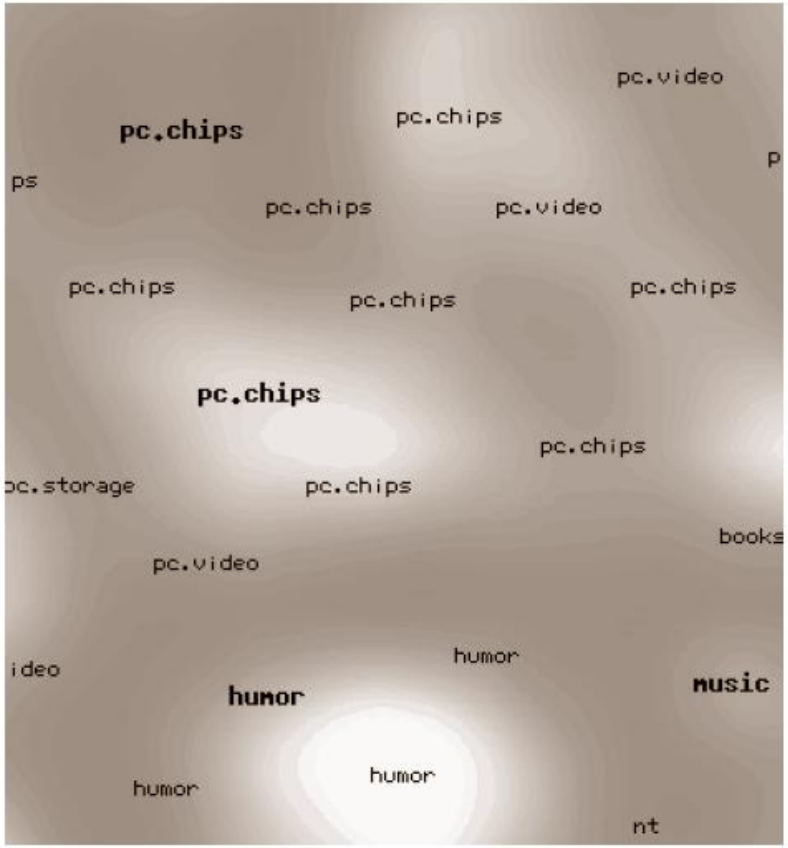
Кластерная гипотеза

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 поиска



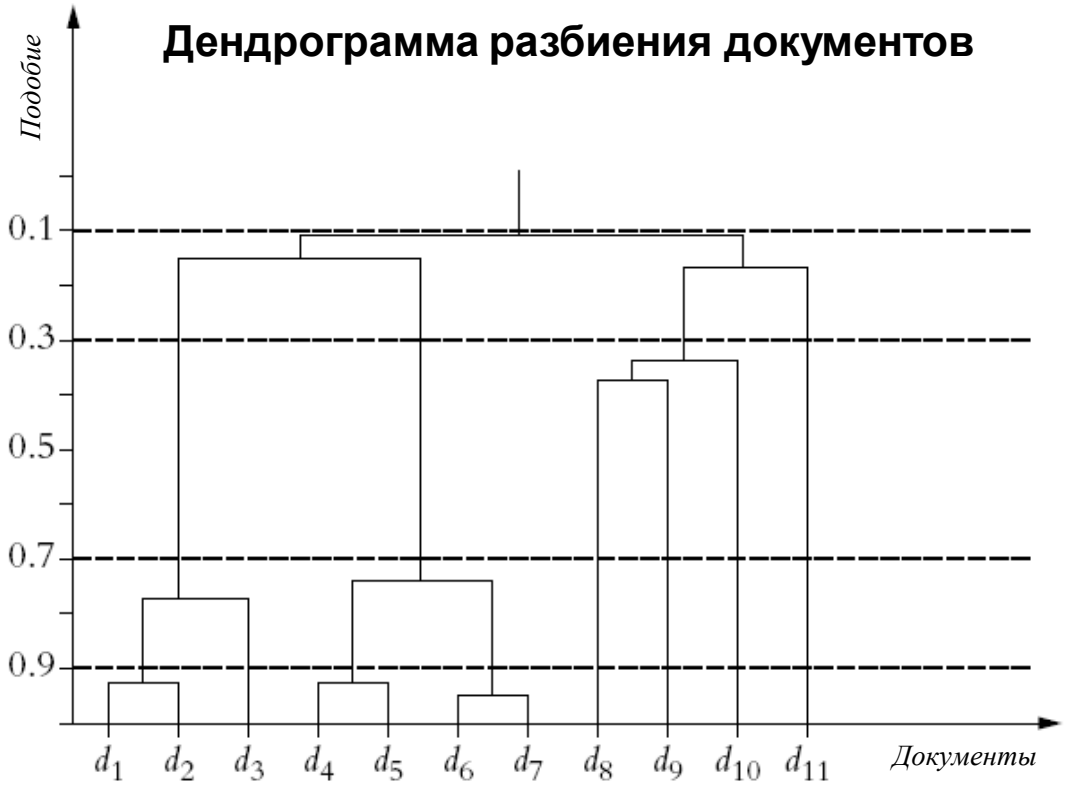
Сравнительное описание подходов

Подход	Описание	Примеры алгоритмов
Подход декомпозиции (разбиения)	В основе данного подхода лежит идея о разбиении коллекции документов на k поднаборов или кластеров $D_1,\ldots,D_k$ так, чтобы уменьшить внутрикластерное расстояние $\sum_i \sum_{d_1,d_2 \in D_i} \delta(d_1,d_2)$ или максимизировать внутрикластерное сходство $\sum_i \sum_{d_1,d_2 \in D_i} \rho(d_1,d_2)$.	- собирательная кластеризация - алгоритм К средних
Подход геометрического включения	В основе данного подхода лежит способность человеческого глаза воспринимать шаблоны и кластеры, заключенные в точках, лежащих в двух или трехмерном пространстве. Методы, используемые при данном подходе, являются по природе эвристическими, нет никакой общей гарантии, что все коллекции документов будут обрабатываться одинаково хорошо.	- Самоорганизующиеся карты Кохонена - Многомерное масштабирование и алгоритм FastMap - Скрытое семантическое индексирование
Вероятностный подход к кластеризации	Вероятностный подход рассматривает коллекцию документов как сгенерированную случайным процессом, используя определенный набор распределений. Для данной коллекции необходимо определить число распределений и параметры, определяющие эти распределения. Оценка этих распределений может рассматриваться как задача кластеризации.	- Смешанные модели и максимизация математического ожидания - Смешанная модель с множественными причинами - Вероятностное скрытое семантическое индексирование

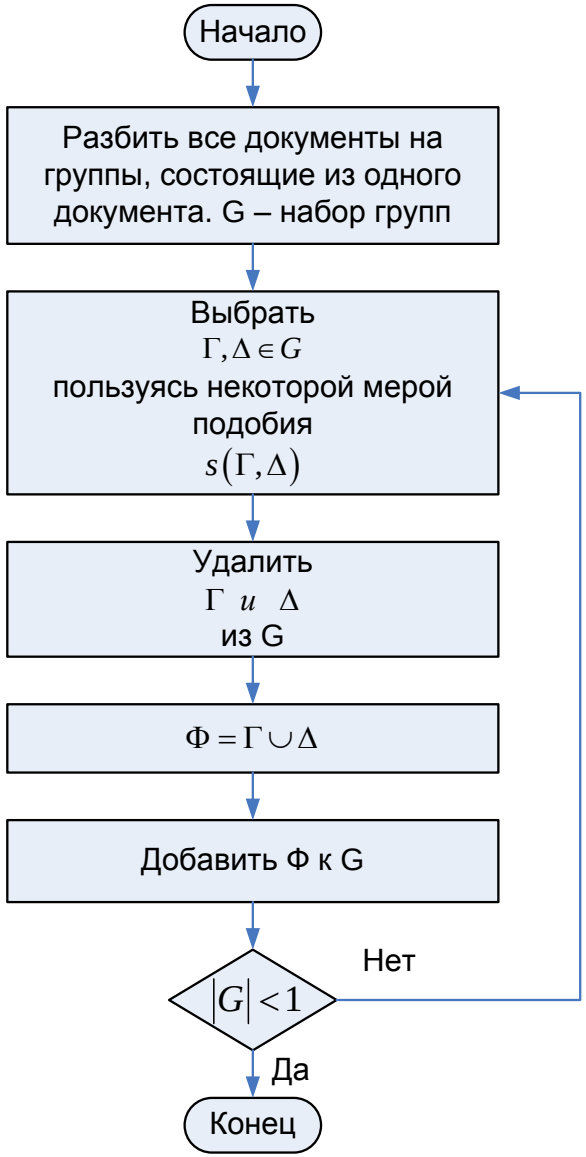
Все документы в коллекции разбиваются на отдельные группы и итеративно производится объединение групп с большим подобием:

$$s(\Phi) = \frac{1}{\binom{|\Phi|}{2}} \sum_{d_1, d_2 \in \Phi} s(d_1, d_2) = \frac{2}{|\Phi|(|\Phi|-1)} \sum_{d_1, d_2 \in \Phi} s(d_1, d_2)$$

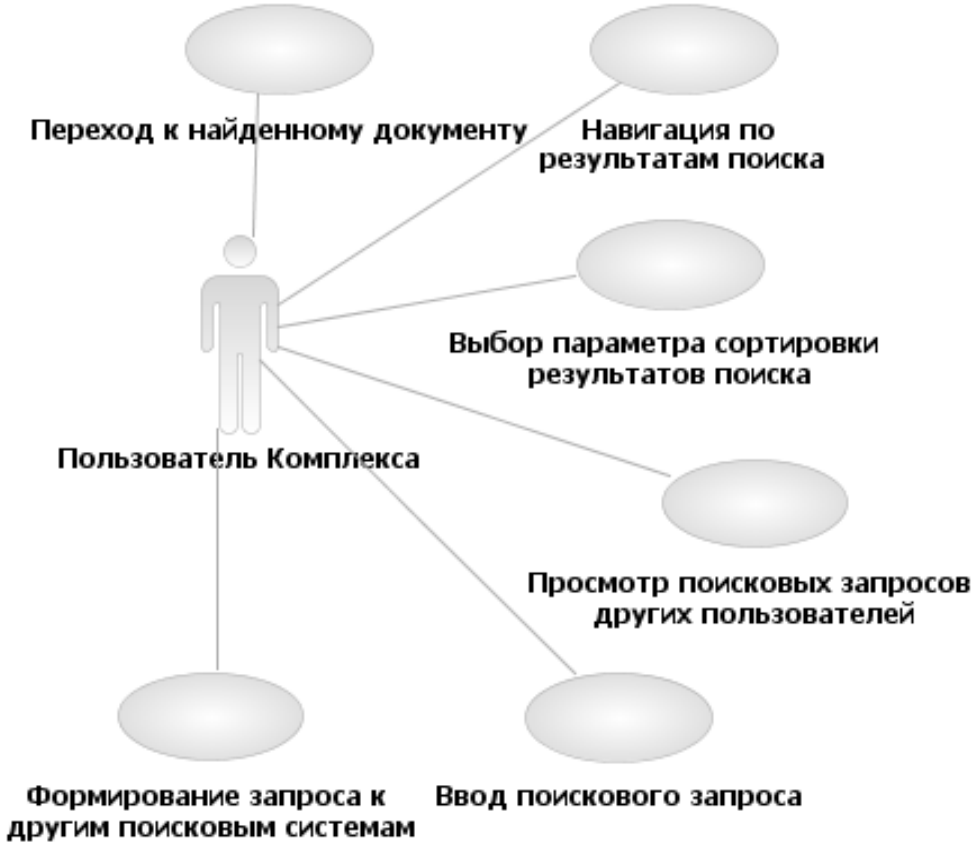
Дендрограмма разбиения документов



Алгоритм кластеризации

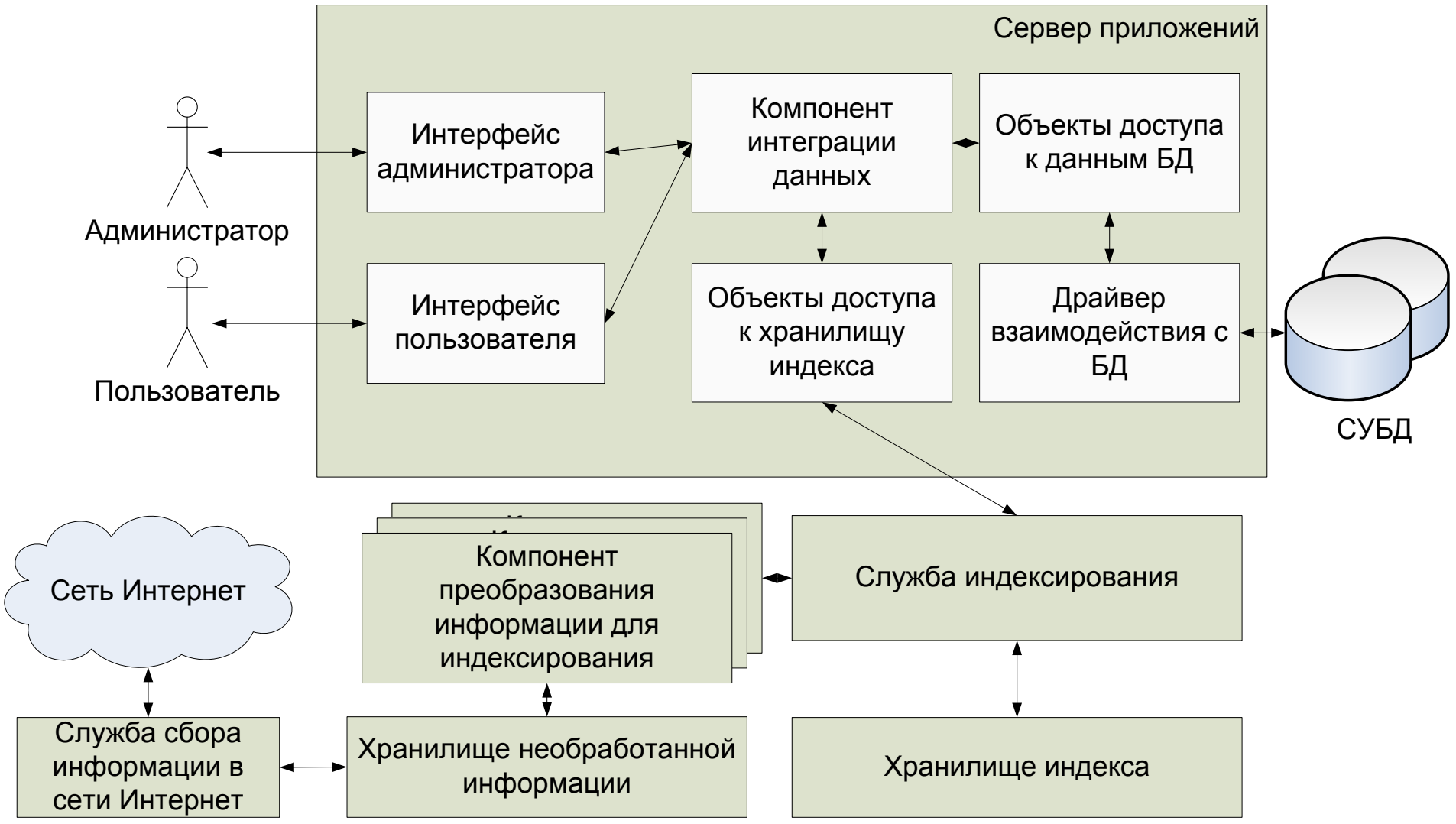


Пользователь Комплекса

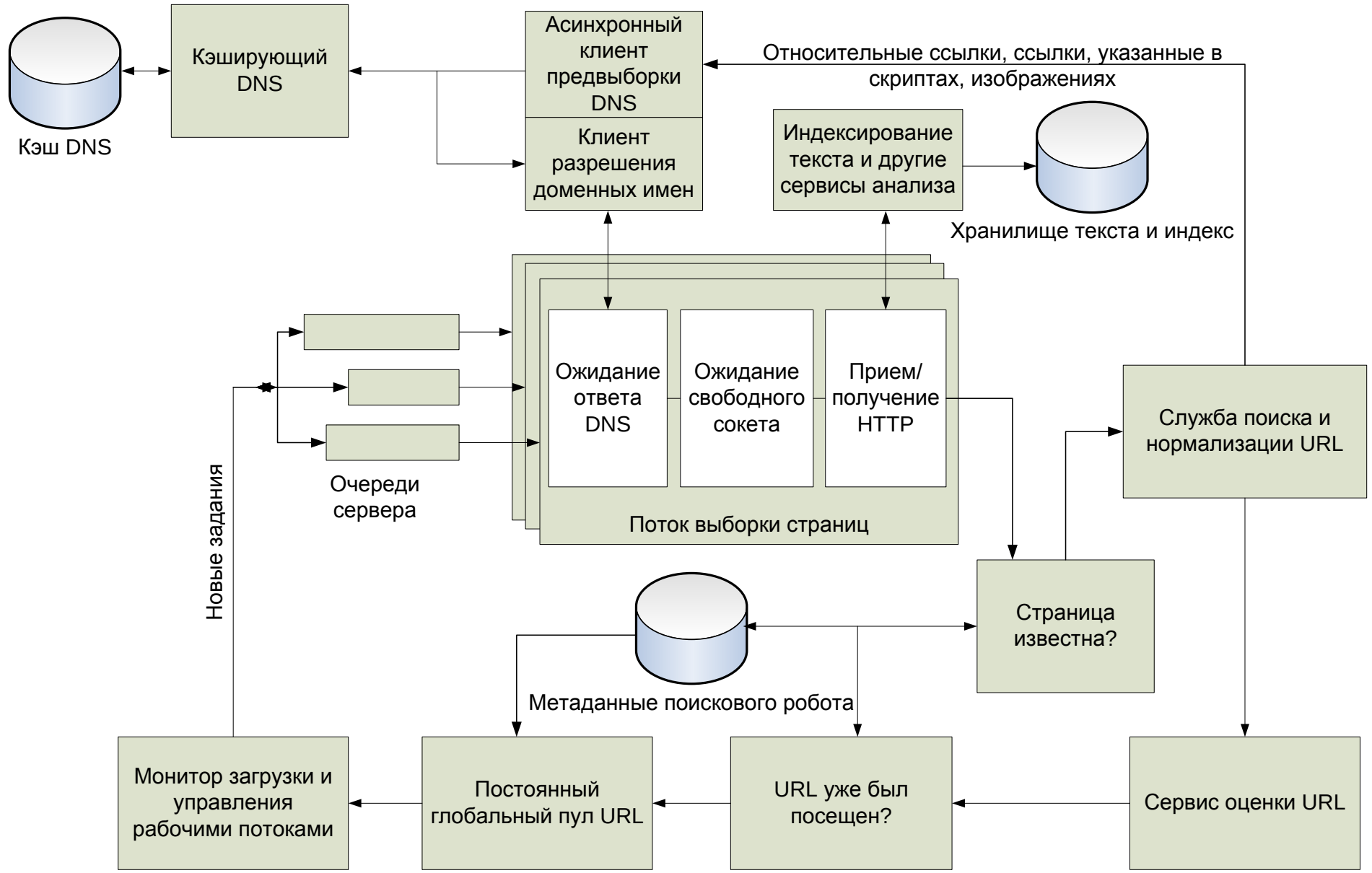


Функциональный администратор

Архитектурное построение Комплекса



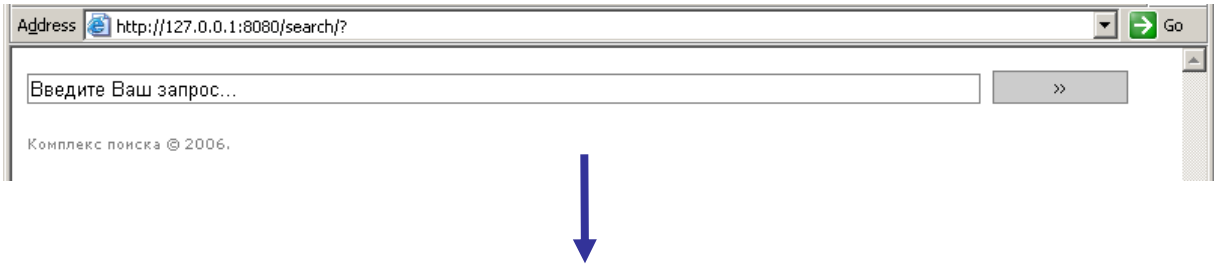
Архитектура службы сбора информации



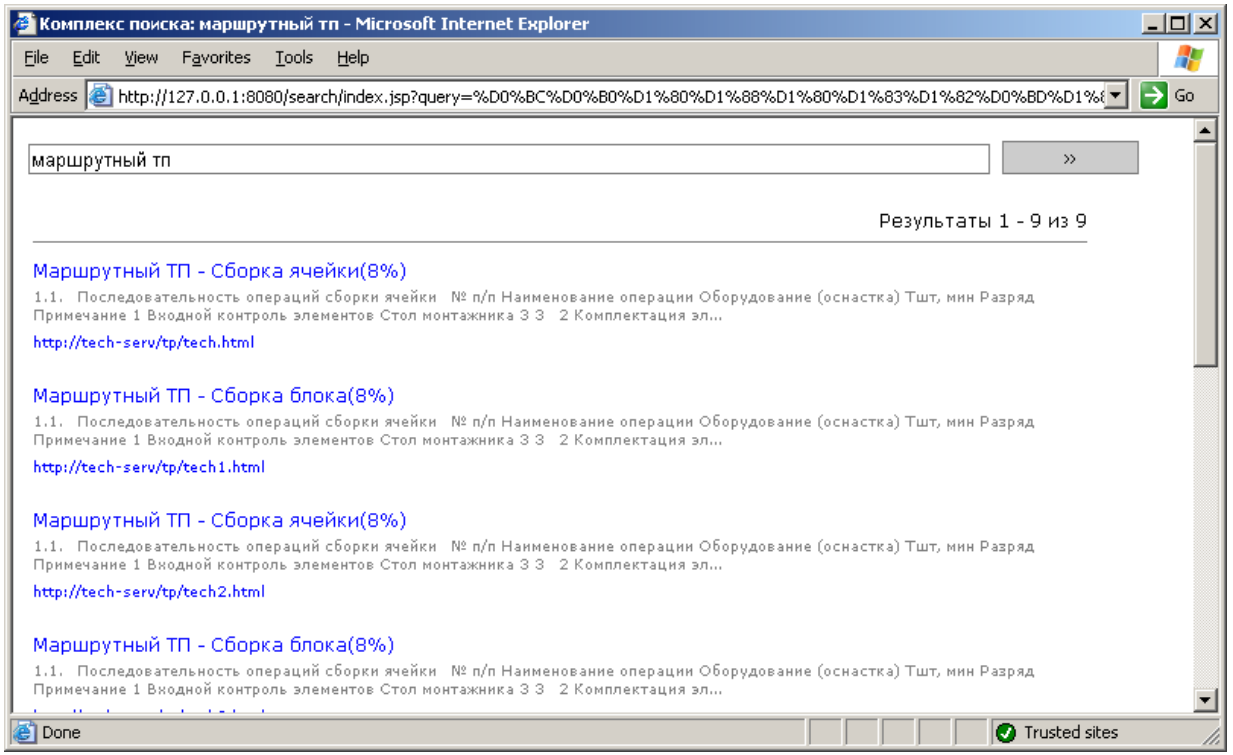
Технология реализации:
тонкий клиент
Язык программирования:
Java
СУБД:
Berkley DB Java Edition
Сервер приложений:
Apache Tomcat
Хранилище индекса:
Файловая система

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

Форма ввода запроса

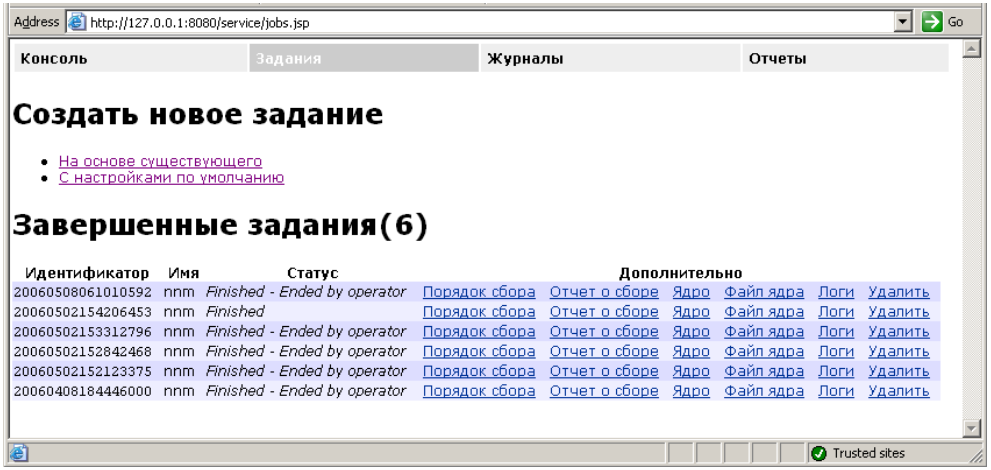


Просмотр результатов поиска

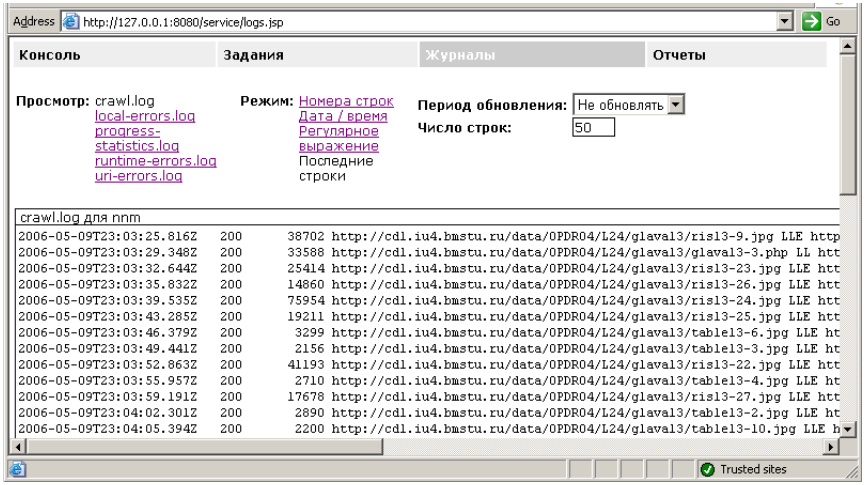


Снимки экранных форм

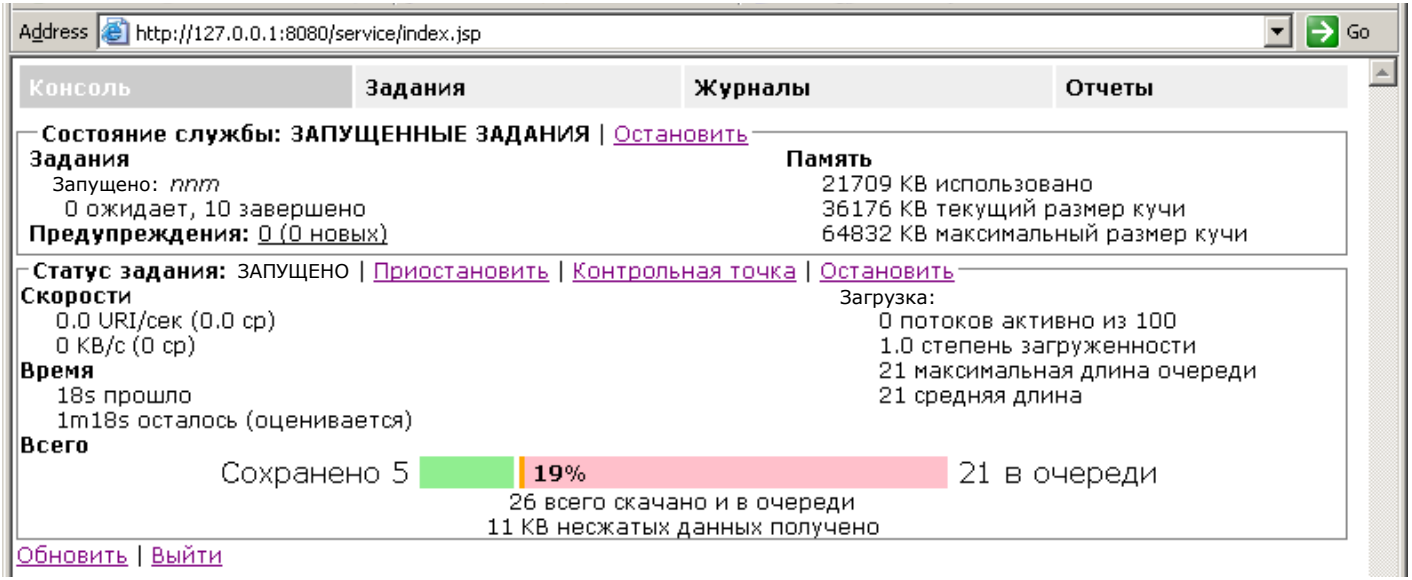
Создание нового задания



Просмотр журналов работы



Мониторинг выполнения задания



Целью исследований является:

- Определение оптимальных параметров использующихся алгоритмов
- 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
- Выбор массива сообщений
- Проведение экспериментальных исследований
- Оценка результатов экспериментов.

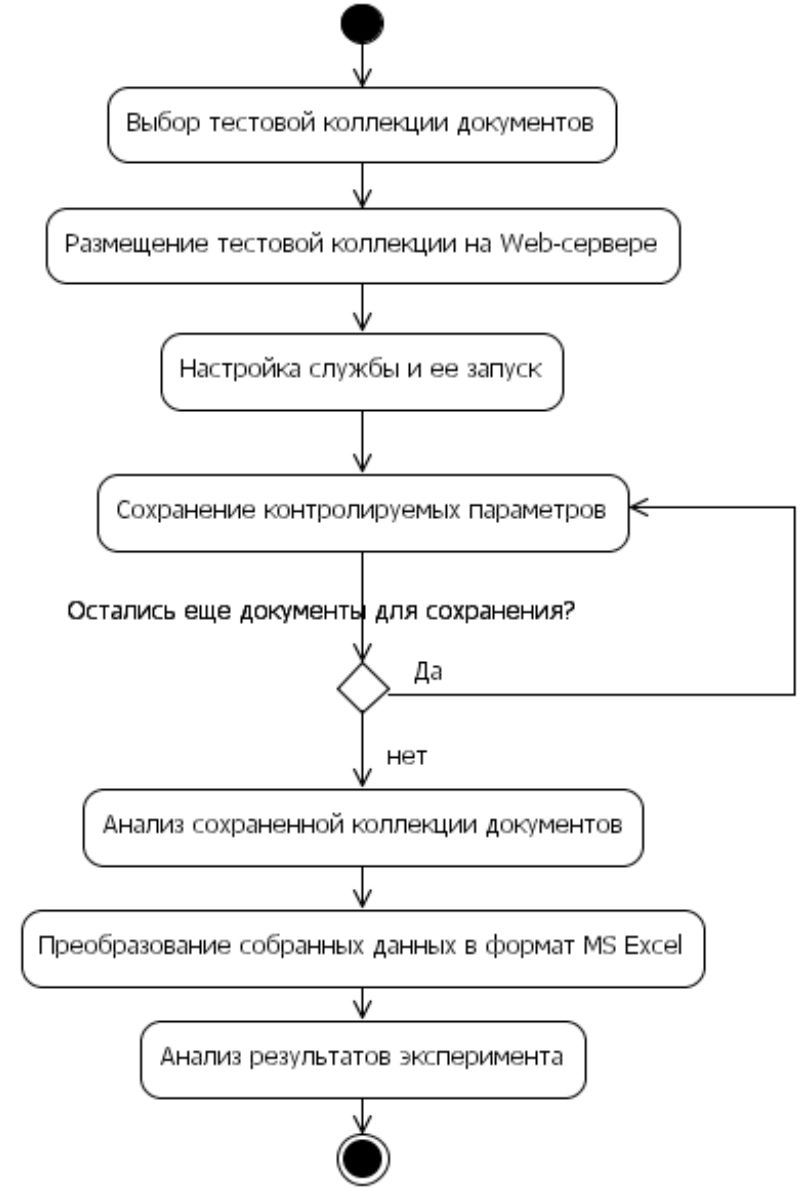
Характеристики тестовой коллекции

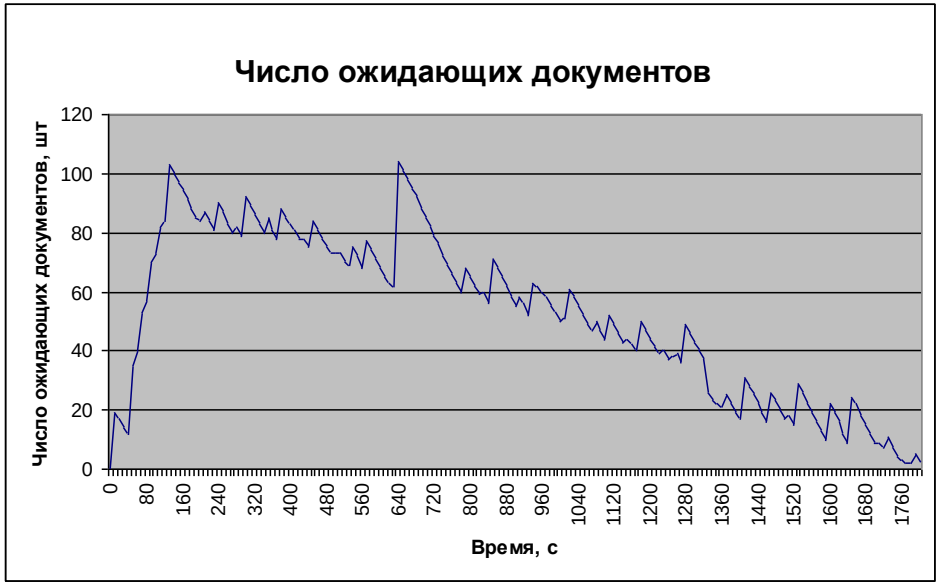
Тематика	Конструкторско-технологическое проектирование электронно-вычислительной техники
Количество документов	80 документов
Средний размер документа	4000 слов
Размер коллекции	10 Мб
Разметка документа	HTML
Язык	Русский

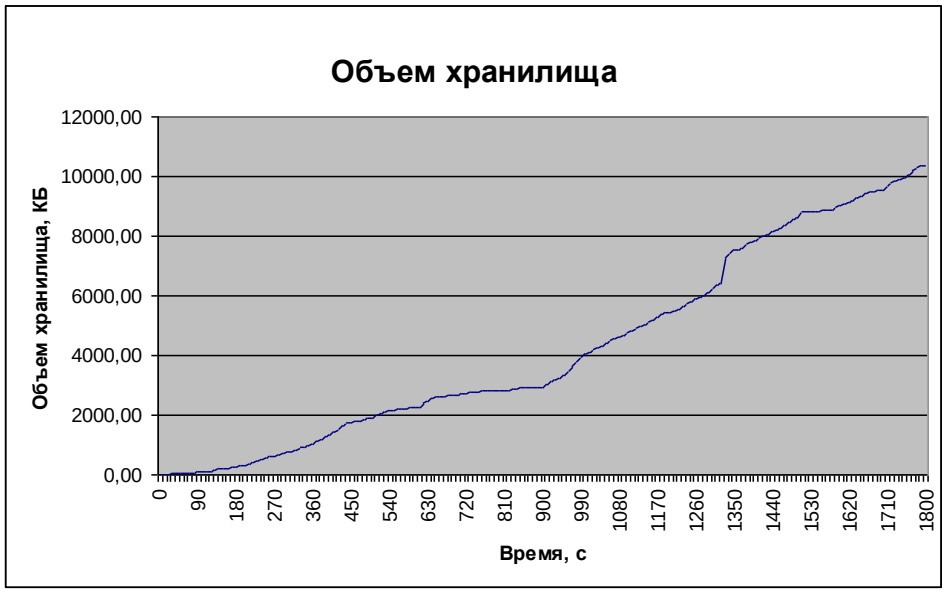
Характеристики экспериментального стенда

Центральный процессор	AMD Athlon XP 2800+ (2.08 ГГц)
ОЗУ	1.25 Гб DDR
Материнская плата	ASUS A7N8X Deluxe
HDD	160 Гб SATA
Контроллер ЛВС	10/100/1000 Base-T
Операционная система	MS Windows XP Prof. SP2
Виртуальная машина Java	Sun Java SDK 1.5
Сервер приложений	Apache Tomcat 5.5.16 Server

Исследование службы сбора информации







Точность оценки релевантности документа запросу определяется совокупностью двух показателей:

точность

$$P = \frac{|A \cap B|}{|B|}$$

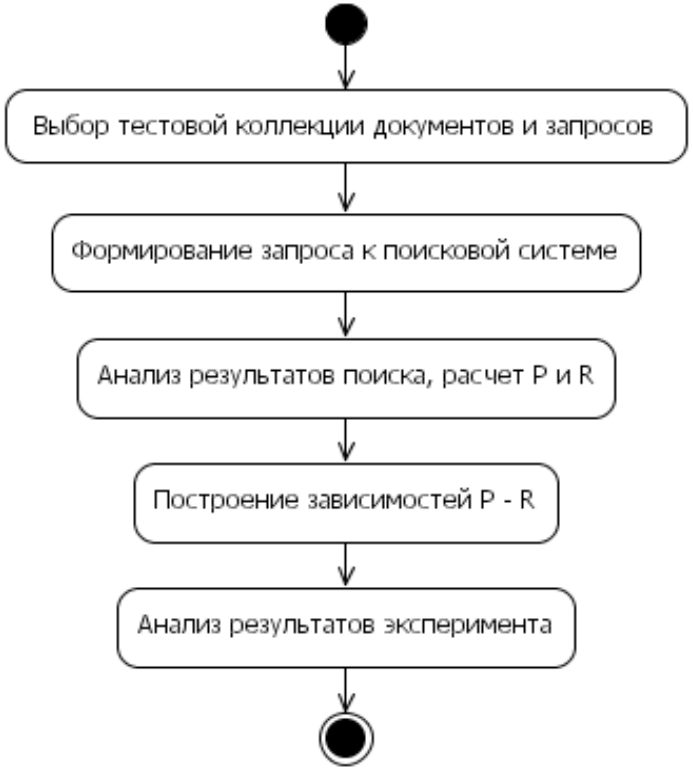
полнота

$$R = \frac{|A \cap B|}{|A|}$$

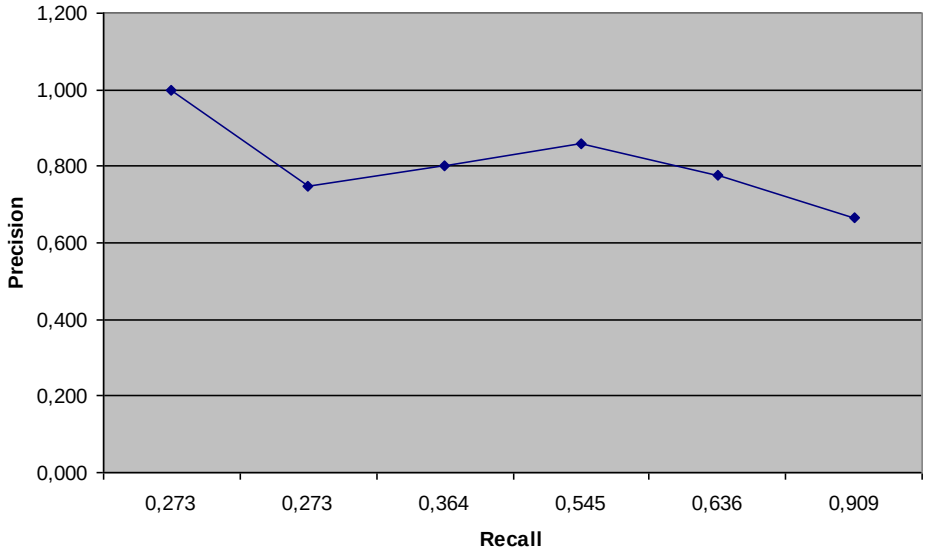
Разбиение массива документов

	Релевантные документы	Нерелевантные документы	
Документы, попавшие в результат поиска	$A \cap B$	$\overline{A} \cap B$	B
Документы, не попавшие в результат поиска	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$	$\overline{B}$
	A	$\overline{A}$	N

Алгоритм проведения эксперимента



Зависимость P - R



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
- Разработан Комплекс поиска и обработки гипертекстовой информации, позволяющий пользователям осуществлять поиск необходимой информации в распределенных источниках данных
- Проведены экспериментальные исследования разработанного Комплекса и предложены меры по повышению его производительности и надежности

Апробация

- Издано 5 публикаций на тему диссертации
- Результаты исследований и разработок докладывались на студенческих конференциях
- Проведенные исследования награждены дипломом по итогам открытого конкурса на лучшую работу студентов
- Комплекс внедрен в информационную инфраструктуру ЗАО «КРОК Инкорпорейтед» - более 30 хранилищ данных, более 1000 клиентских рабочих мест