

Информационный поиск

Поиск информации и информационный поиск (Information Retrieval)

- Поиск информации в Интернет – это повседневная деятельность многих людей
- Поиск информации и общение – это наиболее популярные виды использования компьютеров
- Приложения, использующие поиск информации, - везде вокруг нас
- Сфера науки, которая исследует методы поиска информации, называется информационный поиск (*information retrieval (IR)*)

Информационный поиск

- *“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968)*
- Общее определение, которое может быть применимо ко многим типам приложений обработки и поиска информации
- Основной фокус информационного поиска с 50-х годов – на тексты и документы

Что такое документ?

- Примеры
 - Интернет-страницы, электронные письма, книги, новости, посты форумов, патенты и многое другое
- Общие свойства
 - Значительное текстовое содержание
 - Некоторая структура:
 - заголовок, автор, дата - для статей;
 - тема, отправитель, адресат - для писем

Документы vs. записи базы данных

- Записи базы данных (структурированные таблицы) состоят из хорошо определенных полей и атрибутов
 - e.g., банковские записи балансы, номера счетов, имена, адреса, даты рождения, номера социального обеспечения
- Легко сопоставлять запросы и поля таких баз данных (хорошо определенная семантика)
- Текст более сложный – неструктурированная информация

Документы и записи базы данных

- Запрос к базе данных по банкам
 - *Найти записи с банковским балансом > \$50,000 в отделениях, расположенных в Amherst, MA.*
 - Сопоставление выполняется сравнением со значением соответствующего поля
- Запрос к поисковой машине
 - *Банковские скандалы в России*
 - Этот текст может быть сопоставлен с целым новостным документом

Сравнение текстов

- Сопоставление текста запроса с текстом документа и определение того, что такое хорошее сопоставление – базовый вопрос информационного поиска
- Точное сопоставление слов - недостаточно
 - Много различных способов сказать одно и то же на естественном языке
 - e.g., *преступность в Сибири*
 - Некоторые документы подходят к запросу лучше, чем другие

Измерения информационного поиска

- Информационный поиск – это больше чем поиск по текстам, и больше, чем просто интернет-поиск
 - Хотя эти вопросы являются центральными!
- Поиск осуществляется на основе разных типов данных, разных типов приложений и разных задач

Другие исходные данные (нетексты)

- Поиск по нетекстовым данным
 - видео, фото, музыка, речь
- Их содержание также трудно описывать и сравнивать
 - Текст может использоваться для описания (теги)
- Подходы, созданные для классического информационного поиска являются приемлимыми и для нетекстовых данных

Задачи, связанные с поиском информации

- Ad-hoc поиск
 - Найти релевантный документ в ответ на произвольный запрос
- Фильтрация
 - Отобразить нужные пользователю документы
- Классификация
 - Проставить рубрики документам
- Ответы на вопрос
 - Дать ответ на заданный вопрос
- Визуализация выдаваемой информации
 - Аннотации (рефераты) и др.

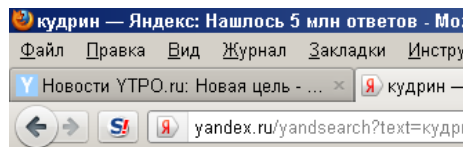
Измерения информационного поиска

| Содержание | Приложения | Задачи |
|--------------|-------------------|--------------------|
| Text | Web search | Ad hoc search |
| Images | Vertical search | Filtering |
| Video | Enterprise search | Classification |
| Scanned docs | Desktop search | Question answering |
| Audio | Forum search | |
| Music | Literature search | |

Вертикальный поиск

- Вертикальный поиск - тематический поиск в Интернет: книги, недвижимость, новости, карты и др.
- Современные поисковые системы – включают элементы вертикального поиска: некоторые запросы направляют на соответствующие вертикали
- Яндекс: поисковые колдунщики
- Вертикали Яндекса: погода, конвертер валют, википедия, новости, маркет, картинки, видео, музыка, вакансии и др.
- <http://help.yandex.ru/search/?id=1111313>

Вертикальный поиск по НОВОСТЯМ



Яндекс
Нашлось
5 млн ответов

[Поиск](#) [Почта](#) [Карты](#) [Маркет](#) [Новости](#) [Словари](#) [Блоги](#) [Видео](#) [Картинки](#) [ещё](#)

кудрин

☐ в найденном ☐ в Москве

[расширенный поиск](#)

Найти

[Войти](#)

[Мои находки](#)

[Помощь](#)

[Настройка](#)

[Регион: Москва](#)

[Все объявления](#)

[Уход Алексея Кудрина](#)

"Единая Россия" сделала Кудрина мишенью для критики. Читайте на сайте www.firstnews.ru



[Чубайс: «Последствия отставки Кудрина могут оказаться драматическими»](#)



Гендиректор ОАО "Российская корпорация нанотехнологий" Анатолий Чубайс заявил, что отставка вице-премьера, министра финансов России Алексея Кудрина может иметь драматические последствия. Решение об отставке главы Минфина, "какими бы причинами оно не было вызвано, создает серьезные риски для страны", отметил А. Чубайс в своем блоге.

[Эхо Москвы](#) 14:18 [Финмаркет](#) 12:03 [Взгляд.ру](#) 10:36

[Все сообщения](#) 26

[Путин утвердил новое распределение обязанностей в правительстве](#) 295 сообщений

[Ясин: Отставка Кудрина грозит ростом расходов и дефицита бюджета](#) 2050

сообщений

news.yandex.ru 4 часа назад

[Кудрин, Алексей Леонидович — Википедия](#)

[Биография](#) [Экономическое мировоззрение](#) [Библиография](#)

Алексе́й Леони́дович Ку́дрин (12 октября 1960, Добеле, Латвийская ССР) — российский государственный деятель, министр финансов Российской Федерации (с мая 2000)...

ru.wikipedia.org > [Кудрин](#) [копия](#) [ещё](#)

[Яндекс.Директ](#)

[Кудрин ушёл в отставку!](#)

Причины ухода министра финансов. Последствия для экономики. Новости часа www.bfm.ru

[Министр финансов Кудрин](#)

отправлен в отставку! Что за этим последует? Подробности: www.zagolovki.ru

[Почему Кудрин не хочет работать в](#)

команде Медведева? Подробности читайте на "Голосе Америки" www.voanews.com

[Отставка Кудрина](#)

Чем грозит отставка госбюджету России? Узнайте подробнее на: investcafe.ru

[Разместить объявление по запросу «кудрин»](#) — 15 519 запросов в месяц

[Видео «кудрин»](#)

Важные понятия в информационном поиске

- Relevance – релевантность
- Evaluation - оценка качества
- Users and Information Needs –
потребность пользователя,
информационная потребность

Релевантность

- Что это?
- Простое (и упрощающее) определение:
Релевантный документ содержит информацию, которую искал пользователь, когда задавал запрос поисковой машине
- На релевантность оказывают влияние много различных факторов: задача, контекст, опыт пользователя, новизна, стиль
- Тематическая релевантность (отражение заданной темы) vs. пользовательская релевантность (все остальные факторы)

Релевантность и модели поиска

- Модели поиска отражают «взгляд» на релевантность
- Ранжирующие алгоритмы, используемые в поисковых машинах базируются на моделях поиска
- Большинство моделей описывают статистические свойства текстов (а не лингвистические)
 - Простые признаки текстов такие, как слова в отличие от синтаксического разбора и учета предложений
 - Лингвистические признаки могут быть частью статистической модели

Оценка качества поиска (evaluation)

- Экспериментальные процедуры и меры для сравнения результатов работы систем с ожиданиями пользователей
- Метода оценки качества поиска сейчас используются во многих областях
- Типично используются тестовые коллекции документов, запросов, и оценки релевантности
- *Полнота и точность – простые примеры оценки качества*

Пользователи и информационная потребность

- Оценка качества поиска – является “пользователецентричной”
- Ключевые слова – это слишком бедное описание действительных информационных потребностей
- Взаимодействие и контекст – важны для понимания потребности пользователя
- Методы уточнения запроса: расширение запроса, предложение запроса, *relevance feedback*

Поисковые машины

ИПС

Информационно-поисковые
системы

Информационный поиск и поисковые машины

- Поисковая машина – это практическое приложение методов информационного поиска к большим текстовым коллекциям
- Интернет-поисковые системы – наиболее известны, но есть много других видов поисковых систем
 - *Open source поисковые системы для исследований*
 - Lucene, Lemur/Indri, Galago
- Практическая реализация предполагает решение дополнительных вопросов

Информационный поиск и поисковые машины-2

Информационный поиск

Релевантность

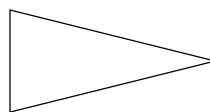
*-Эффективное
ранжирование*

Оценка качество

*-Тестирование и
измерение*

Потребности
пользователя

*-Взаимодействие с
пользователем*



Поисковые машины

Исполнение запроса

*-Эффективный поиск и
индексирование*

Включение новых данных

-Покрытие и свежесть

Масштабируемость

*-Рост с данными и
пользователями*

Адаптивность

-Настройка на приложения

Специфические проблемы

-например, спам

Особенности работы поисковых машин

- Выполнение запроса (performance)
 - Измерение и улучшение эффективности поиска
 - Уменьшение времени ответа, увеличение скорости индексирования
- *Индексы – это структуры данных, которые необходимы, чтобы уменьшить время ответа системы*
 - Важнейший вопрос для поисковых систем

Особенности работы поисковых машин - 2

- Динамические данные
 - «Коллекции» данных для наиболее востребованных приложений постоянно меняются: обновляются, удаляются, пополняются
 - Например, веб-страницы
 - Acquiring or “crawling” the documents is a major task
 - Типичные меры: покрытие (сколько проиндексировано) и новизна (*freshness*) (насколько недавно проиндексировано)
 - Необходимо одновременно менять индексы и обрабатывать запросы

Особенности работы поисковых машин-3

- Масштабируемость
 - Миллионы пользователей и терабайты документов
 - Используется распределенная обработка
- Адаптивность
 - Изменение и настройка компонентов поисковой машины, таких как алгоритм ранжирования, методы индексирования, интерфейсы для различных приложений

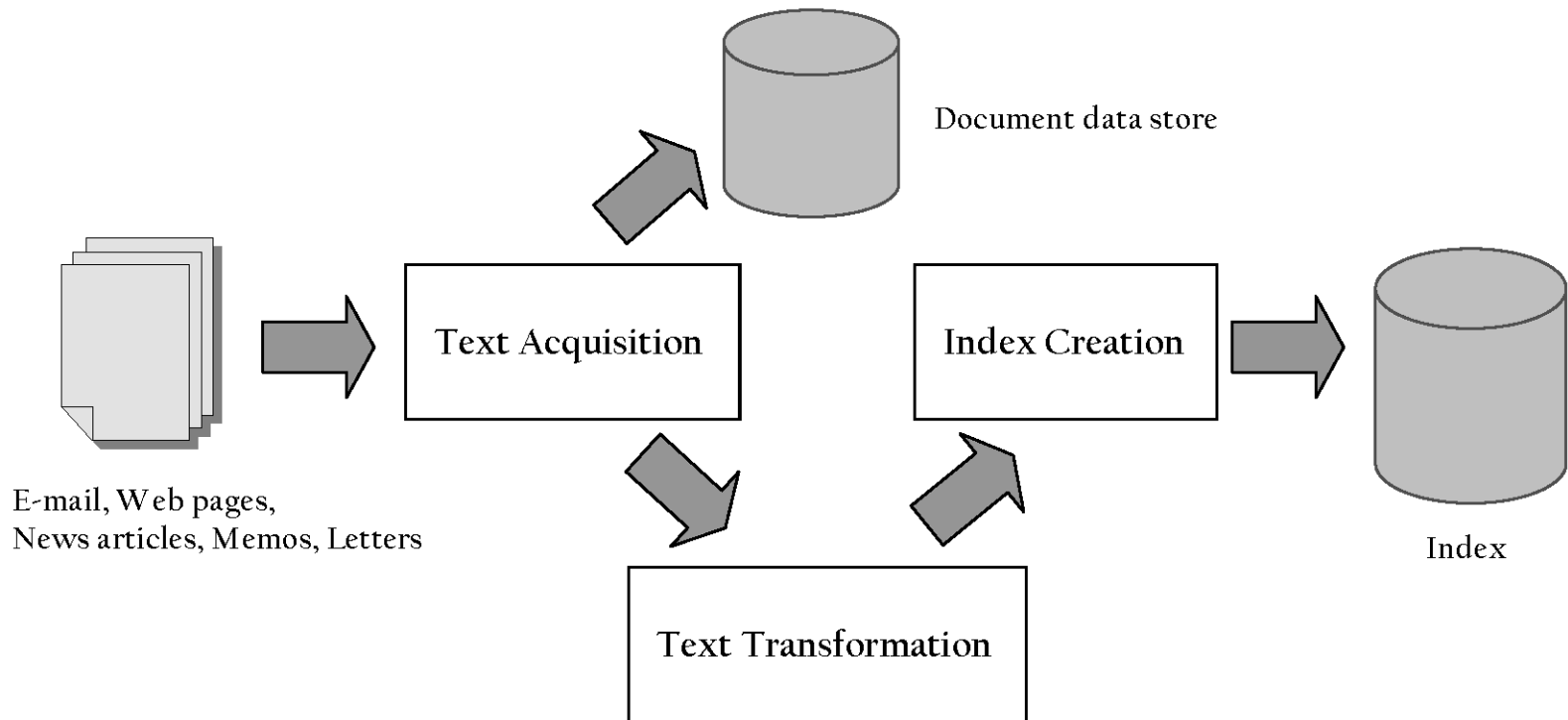
Поисковый спам

- Для веб поиска одним из важных направлений работы является поисковый спам
- Важно для качества поисковых результатов
- Много видов спама
 - Порождение текстов похожих на естественные
 - Ссылочный спам и др.
- Новая область информационного поиска - *adversarial IR*,
 - Спамеры - противники с различными целями

Архитектура поисковых машин

Основные компоненты
поисковых машин

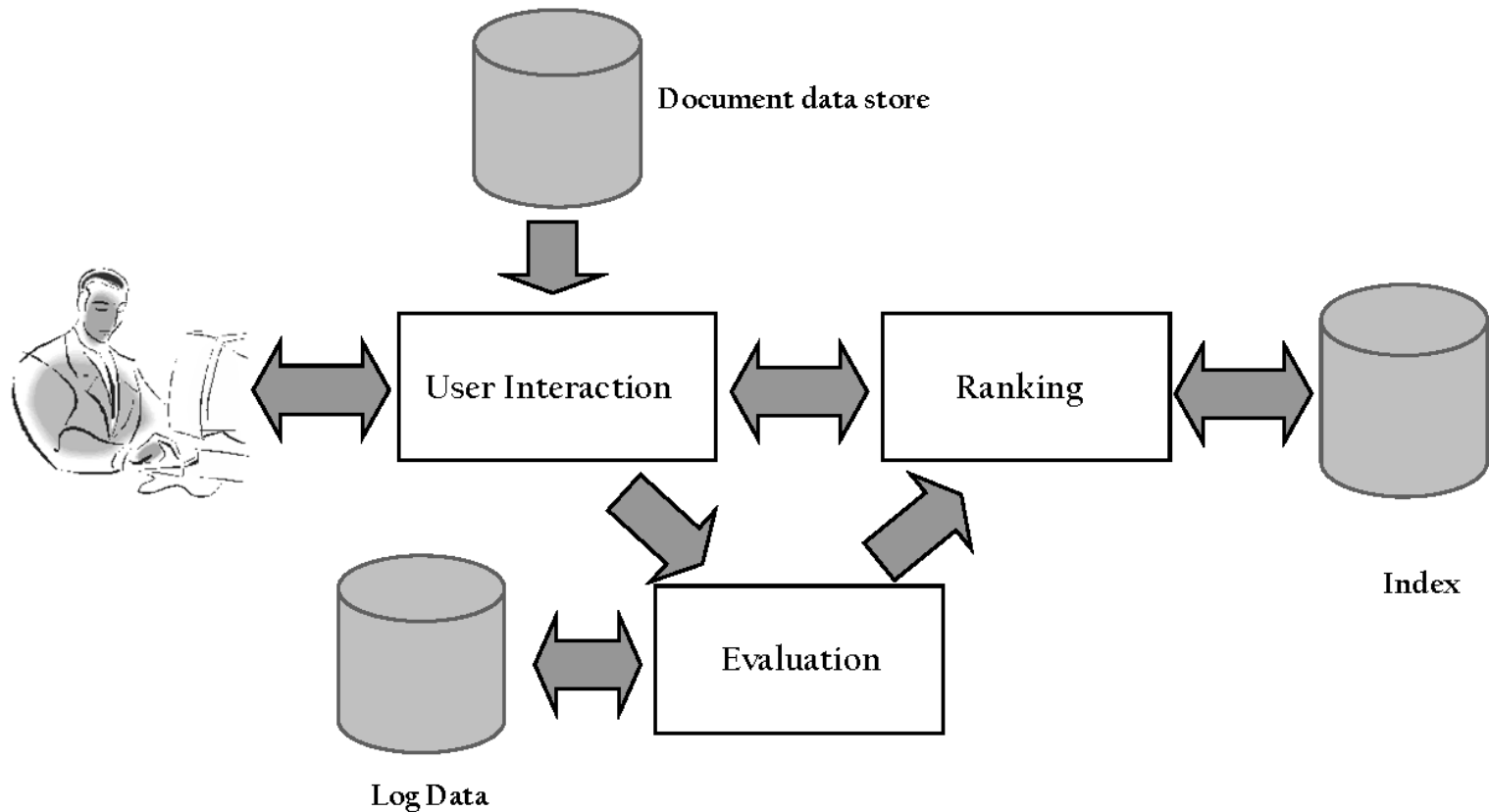
Процесс индексирования



Процесс индексирования-2

- Извлечение текстов
 - Идентифицирует и сохраняет тексты для индексирования
- Трансформация текстов
 - Трансформирует документы в индексные термины
- Создание индексов
 - Берет индексные термины и создает индексы для быстрого поиска

Обработка запроса



Обработка запроса

- Взаимодействие с пользователем
 - Поддерживает создание и уточнение запроса, показ результатов
- Ранжирование
 - Использует запрос и индексы породить ранжированный лист документов
- Оценка качества
 - Мониторит и измеряет качество поиска

Извлечение текстов. Краулер

- Идентифицирует и извлекает документы для поисковой машины
- Много типов – интернет, предприятие, компьютер
- Интернет-краулеры используют ссылки, чтобы найти документы
 - Должны найти огромное количество веб-страниц (покрытие) и сохранять их в актуальном состоянии
 - Краулеры сайтов
 - Тематические краулеры для вертикального поиска
- Краулеры документов для поиска по документам предприятия или компьютера
 - Используют ссылки и сканируют директории

Получение текстов-2

- Фиды
 - Потоки документов в реальном потоке времени
 - Новости, блоги, видео, радио, tv
 - RSS - стандарт
 - RSS читалка обеспечивает новые XML документы поисковой машине
- Конвертация
 - Конвертирует форматы в текст плюс мета-данные
 - HTML, XML, Word, PDF, и др. → XML
 - Конвертирует кодировки для различных языков
 - Например, в кодировку UTF-8

Получение текстов-3

- Хранилище документов
 - Хранит тексты, метаданные и другое содержание документов
 - Метаданные: тип, дата создания
 - Ссылки, текст ссылки
 - Обеспечивает быстрый доступ к содержанию документов
 - Порождение списка результатов
 - Эффективное хранение
 - не реляционная база данных

Преобразование текстов

- Анализатор (Parser)
 - Обработывает последовательность токенов в документе, распознает структурные элементы
 - Заголовки, ссылки, подзаголовки и др.
 - *Токенизатор распознает «слова» в тексте*
 - Обработка капитализации, кавычек, дефисов ..
 - *Обработка структуры, задаваемой HTML, XML*
 - *Теги:* `<h2> Overview </h2>`
 - Парсер использует синтаксис языка разметки идентифицировать структуру документа

Преобразование текстов-2

- Стоп-слова
 - Удаление наиболее частотных слов
 - Предлоги, союзы, артикли..
 - Может быть проблемой для некоторых запросов
- Стемминг (морф. анализ)
 - “computer”, “computers”, “computing”, “compute”
 - Обычно эффективен, но не для всех запросов
 - Разное действие для разных языков

Преобразование текстов-3

- Анализ ссылок
 - Ссылки и тексты ссылок (анкор ссылки - якоря)
 - Анализ ссылок важен для определения популярности сайта и сообщества, связанного с сайтом
 - Например, PageRank
 - Текст ссылки может значительно уточнить содержание связанных страниц, Significant impact on web search
 - Значительное влияние на интернет-поиск
 - Меньше значимость в других поисковых приложениях

Преобразование текстов-4

- Извлечение информации
 - Идентифицирует семантические классы индексных термов, которые важны для конкретных приложений индекс
 - Например, распознавание имен людей, географических мест, компаний, дат...
- Классификатор
 - Отнесение текста к категориям
 - Тематика, тональность, жанры и др.
 - Зависит от приложения

Создание индекса

- Статистика по документам
 - Собирает частоты и позиции слов и других признаков
 - Используется в ранжирующем алгоритме
- Определение весов
 - Вычисляет веса для индексных термов
 - Используется в алгоритме ранжирования
 - например, вес *tf.idf*
 - Комбинирование частоты слова в документе и инверсной поддокументной частоты слова в коллекции

Создание индекса-2

- Инвертирование
 - Преобразует матрицу документ-терм в данные терм-документ, необходимые для индексирования
 - Сложно для большого числа документов
 - Формат инвертированного файла – для быстрой обработки запросов
 - Должен обрабатывать изменения индекса
 - Сжатие данных

Создание индекса-3

- Распределенное хранение индекса
 - Распределяет индексы по многим компьютерам и/или многим дата-центрам
 - Необходимо для быстрой обработки запросов
- Много вариантов
 - Подокументное распределенное хранение, распределенное хранение термов, повтор данных

*Особое направление исследований:
Distributed IR – распределенный
информационный поиск*

Взаимодействие с пользователем


- Ввод запроса
 - Интерфейс и парсер для языка запросов
 - Большинство интернет запросов – простые
- Язык запросов нужен для описания сложных запросов и результатов трансформации запросов (работа т.н. колдунщиков запросов)
 - Булевские запросы
 - Специализированные языки запросов для информационно-поисковых систем (Indri, Galago)
 - Сходны с SQL языками, используемыми в базах данных

Язык запросов Яндекса


Как пользоваться поиском?

[Базовые возможности](#)

 [Результаты поиска](#)

 [Поисковые подсказки](#)

[Диалоговые подсказки](#)

 [Поисковые колдунчики](#)

[Исправление запроса](#)

[Расширенный поиск](#)

[Сниппеты профилей ВКонтакте](#)

 [Вопросы и ответы](#)

Будьте бдительны!

[Вирус подмены страниц](#)

[Изменение домашней страницы в браузере](#)

[Предупреждение о потенциально опасных сайтах](#)

Памятка по использованию языка запросов

| Пример | Значение |
|-----------------------------------|--|
| "К нам на утренний рассол" | Слова идут подряд в точной форме |
| "Прибыл * посол" | Пропущено слово в цитате |
| полгорбушки & мосол | Слова в пределах одного предложения |
| снаряжайся && добудь | Слова в пределах одного документа |
| технический прогресс +антирес | Поиск документов, в которых обязательно встречается определённое слово |
| глухаря куропатку кого-нибудь | Поиск любого из слов |
| не сможешь << винить | Неранжирующее "и": выражение после |

Полезные сервисы

[Мои находки](#)

[Виджет](#)

[Яндекс.Поиска](#)

[Мобильный](#)

[Яндекс](#)

[Семейный](#)

[поиск](#)

[Для](#)

[слабовидящих](#)

[Аскетичный](#)

[поиск](#)

[Пожаловаться](#)

[на спам](#)

[Сервис для](#)

[вебмастеров](#)

[Блог](#)

[Яндекс.Поиска](#)

Взаимодействие с пользователем-2

- Трансформация запросов
 - Улучшает исходный запрос
 - Спеллчекинг
 - Подсказка запроса
 - Автоматическое расширение запроса – пополнение его дополнительными словами
 - *Relevance feedback* - автоматизированная технология с участием пользователя
 - Пользователь размечает релевантные документы

Взаимодействие с пользователем-3

- Выдача результатов
 - Строит поисковую выдачу (SERP)
 - Порождает сниппеты, чтобы отразить соответствие документа запросу
 - *Подсвечивает важные слова*
 - Показывает релевантную рекламу – основной источник прибыли Интернет-поисковых систем
 - Может обеспечивать кластеризацию результатов и другие виды визуализации

Ранжирование

- Присваивание веса соответствия документа запросы
 - Веса использует алгоритм ранжирования
 - Базовый компонент поисковой машины
 - Базовое вычисление веса $\sum q_i d_i$
 - q_i и d_i – веса слова запроса и документа
 - Много вариантов алгоритмов вычисления весов и ранжирования

Ранжирование-2

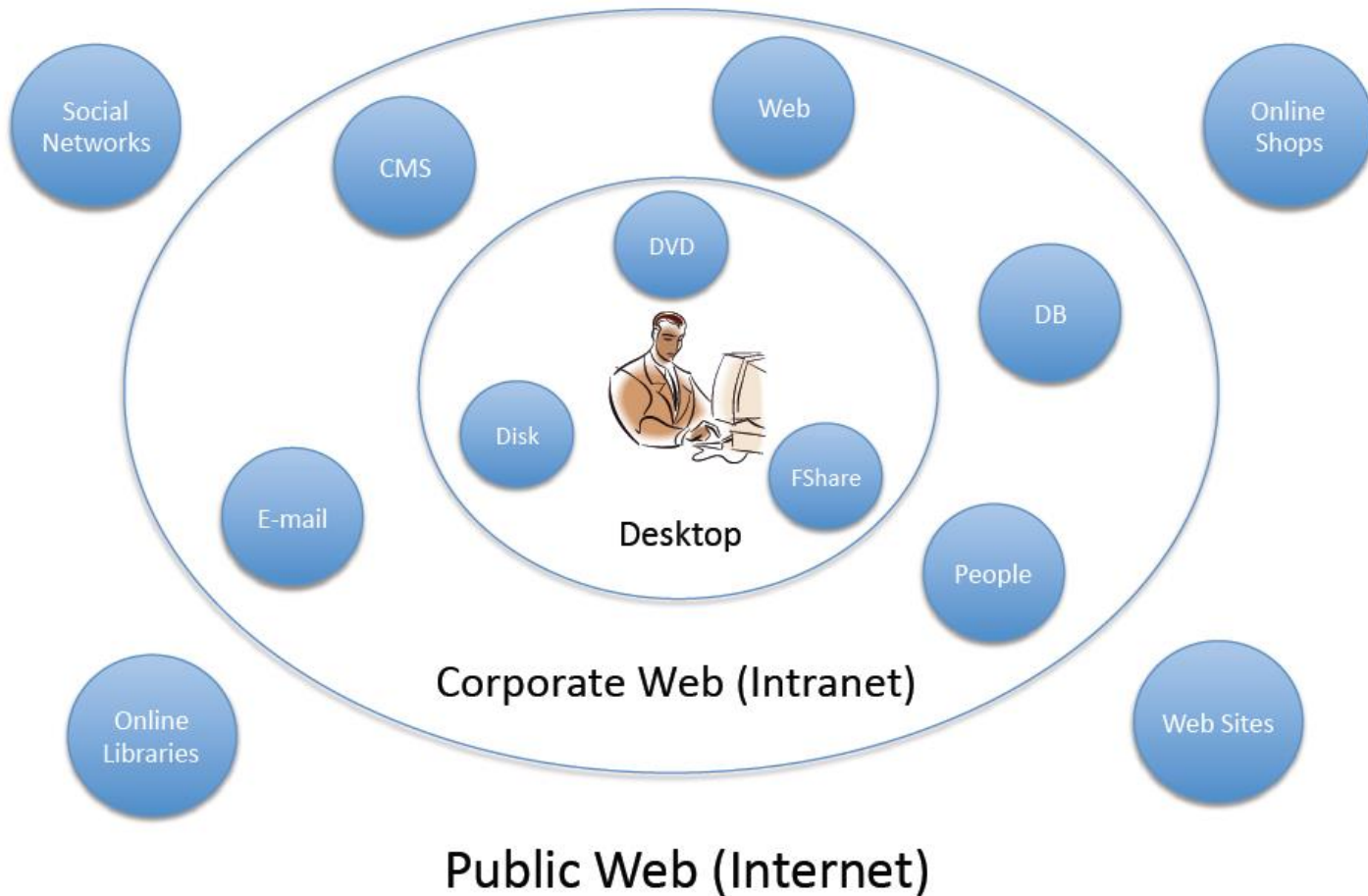
- Оптимизация выполнения запроса
 - Ранжирующие алгоритмы должны позволять эффективное исполнение
- Распределенное выполнение
 - Обработка запросов в распределенной среде
 - *Брокер запросов рассылает запросы и собирает результаты*
 - *Кэширование*

Поисковые системы разного уровня

Russir, 2009

Курс “Enterprise and Desktop Search”
(Дмитриев и др.)

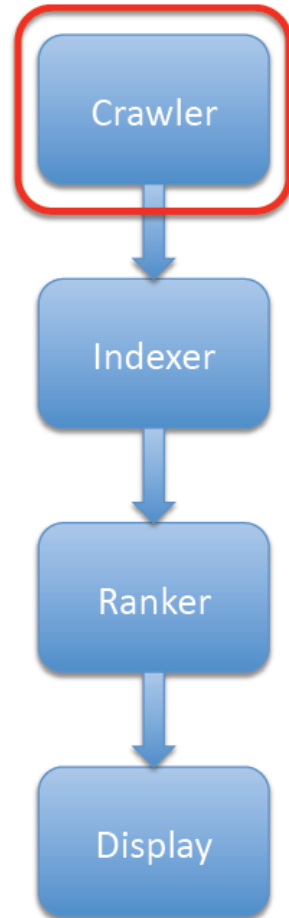
Search Environment of a Company Employee



Интернет-поиск vs. Корпоративный ПОИСК

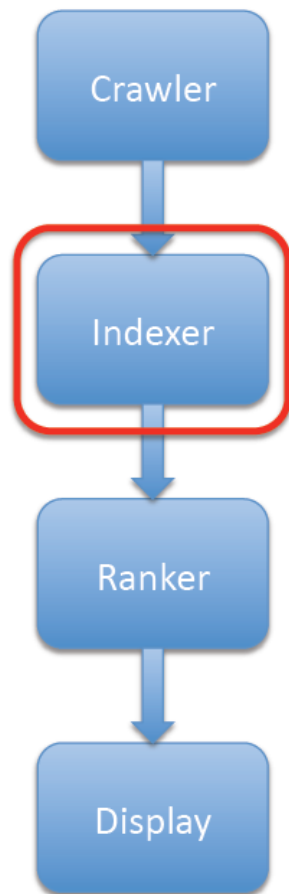
- Интернет-поиск
 - Собирает результаты по общедоступному Интернету
 - Проблема ранжирования ранжирования результатов
 - Большие объемы
 - Громадная индустрия – Интернет-реклама
 - Активные исследования: хорошее качество
- Корпоративный поиск
 - Собирает информацию разных форматов из совокупности хранилищ
 - Ранжирование документов разного типа
 - Относительно малый объем исследований
 - Хуже качество поиска – сложнее проводить сравнительные исследования
 - Активная сфера исследований

Differences between Web Search and Enterprise Search (Crawling)



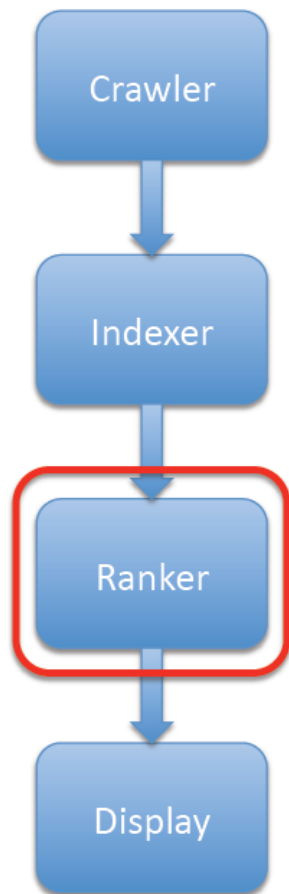
- Diverse information sources and formats, many are not “crawl-friendly”
 - Web pages, files, databases, etc.
 - “Compound” and “composite” documents
- A “click” may have undesirable side effects
 - Document deleted
 - Charge for accessing a 3rd party’s database
- Many security domains
- Hard to create a research test collection

Differences between Web Search and Enterprise Search (Indexing)



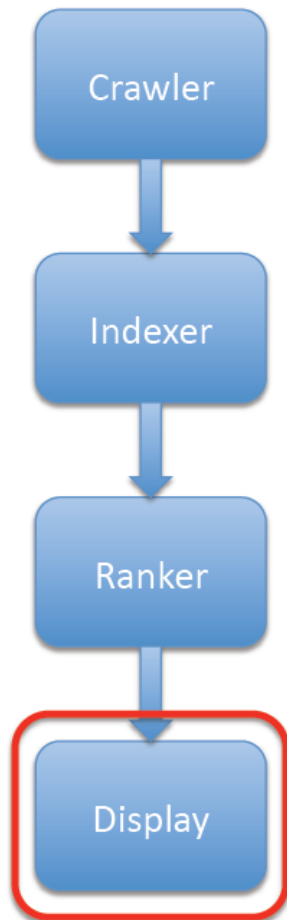
- Data is semi-structured and semantics is often known
 - Search for “objects” (people, rooms, printers, etc)
- Need to incorporate access-control info
- Vocabulary mismatch is a big problem
 - Need to use thesaurus
- Need to index special symbols/punctuation
- Need to efficiently support the kinds of queries generated by exploratory interfaces

Differences between Web Search and Enterprise Search (Ranking)



- Small set of correct answers (often just 1)
- Less hyperlinks and anchor text, and of poorer quality
- Poorer quality of content (pages are not created with a search engine in mind)
- User identity is often known
 - Can use user context
- Often need to retrieve ALL relevant documents
- No (intentional) spam
- Federation and blending is often necessary

Differences between Web Search and Enterprise Search (Display)



- Known identity and user history → personalized results presentation
- Search clients are not just browsers
 - Applications/Advanced search interfaces
- Since ranking is hard, need to provide exploratory interfaces / interactive search & browse experience, give the user more control

Заключение

- Информационный поиск
 - Потребности пользователя
 - Релевантность
 - Оценка качества поиска
- Типы поисковых систем
 - Интернет-поиск
 - Корпоративный поиск
 - Предметно-ориентированный поиск