Введение в поисковые системы

Лекция N 5 курса
“Современные задачи теоретической информатики”

Юрий Лифшиц
yura@logic.pdmi.ras.ru

ИТМО

Осень’2005

План лекции

1. **Архитектура поисковых систем**
2. **Алгоритмы поисковых систем**
   - PageRank
   - Собственные смыслы
3. **Поисковая оптимизация**

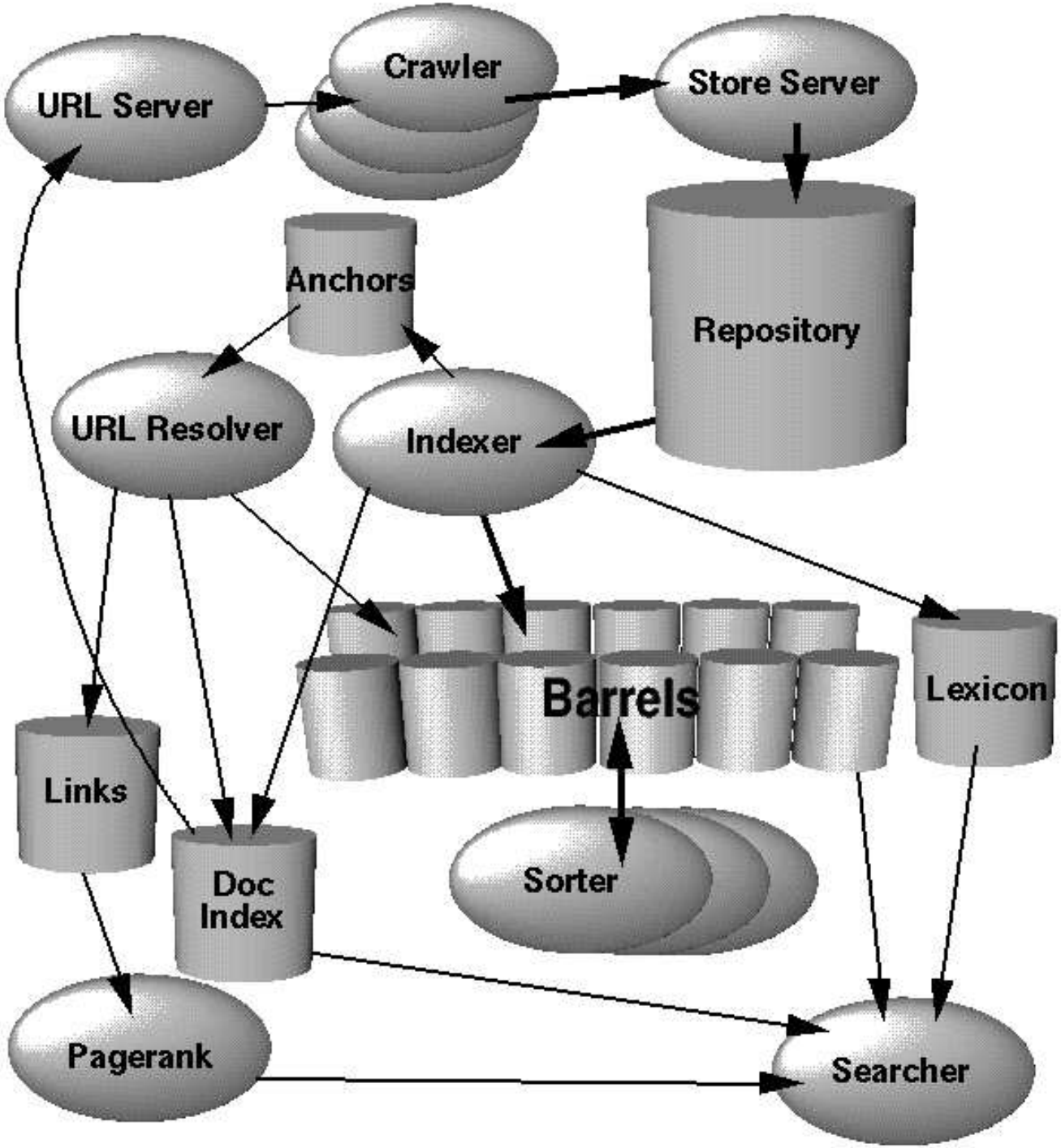
План лекции

1. **Архитектура поисковых систем**
2. Алгоритмы поисковых систем
   - PageRank
   - Собственные смыслы
3. Поисковая оптимизация

Анатомия поисковой системы

Любая поисковая система содержит три базовые части:

- Робот (он же краулер, спайдер или индексатор)
- Базы данных
- Клиент (обработка запросов)


URL Server
Crawler
Store Server
Anchors
Repository
URL Resolver
Indexer
Barrels
Lexicon
Links
Doc Index
Sorter
Pagerank
Searcher

Прямой и обратный индекс

Прямой индекс — записи отсортированы по документам

- Номер документа
- Отсортированный список слов
- Для каждого слова: первые несколько вхождений, частота вхождений, формат вхождений

Прямой и обратный индекс

Прямой индекс — записи отсортированы по документам

- Номер документа
- Отсортированный список слов
- Для каждого слова: первые несколько вхождений, частота вхождений, формат вхождений

Обратный индекс — записи отсортированы по словам

- Номер слова
- Отсортированный список документов
- Для каждого документа: вся информация о вхождении

Релевантность

Характеристики, влияющие на позицию в списке ответов?

Релевантность

Характеристики, влияющие на позицию в списке ответов?

- Наличие слов на сайте

Релевантность

Характеристики, влияющие на позицию в списке ответов?

- Наличие слов на сайте
- Частота слов

Релевантность

Характеристики, влияющие на позицию в списке ответов?

- Наличие слов на сайте
- Частота слов
- Форматирование

Релевантность

Характеристики, влияющие на позицию в списке ответов?

- Наличие слов на сайте
- Частота слов
- Форматирование
- Близость слов друг к другу

Релевантность

Характеристики, влияющие на позицию в списке ответов?

- Наличие слов на сайте
- Частота слов
- Форматирование
- Близость слов друг к другу
- Количество ссылок с других страниц на данную

Релевантность

Характеристики, влияющие на позицию в списке ответов?

- Наличие слов на сайте
- Частота слов
- Форматирование
- Близость слов друг к другу
- Количество ссылок с других страниц на данную
- Качество ссылок

Релевантность

Характеристики, влияющие на позицию в списке ответов?

- Наличие слов на сайте
- Частота слов
- Форматирование
- Близость слов друг к другу
- Количество ссылок с других страниц на данную
- Качество ссылок
- Соответствие тематик сайта и запроса

Релевантность

Характеристики, влияющие на позицию в списке ответов?

- Наличие слов на сайте
- Частота слов
- Форматирование
- Близость слов друг к другу
- Количество ссылок с других страниц на данную
- Качество ссылок
- Соответствие тематик сайта и запроса
- Регистрация в каталоге, связанном с поисковой системой

Как работает клиент?

1. Разбирает запрос на слова

Как работает клиент?

1. Разбирает запрос на слова
2. Переводит слова в их идентификаторы

Как работает клиент?

1. Разбирает запрос на слова
2. Переводит слова в их идентификаторы
3. Для каждого слова находит в обратном индексе список документов, его содержащих

Как работает клиент?

1. Разбирает запрос на слова
2. Переводит слова в их идентификаторы
3. Для каждого слова находит в обратном индексе список документов, его содержащих
4. Одновременно бежит по этим спискам, ища общий документ

Как работает клиент?

1. Разбирает запрос на слова
2. Переводит слова в их идентификаторы
3. Для каждого слова находит в обратном индексе список документов, его содержащих
4. Одновременно бежит по этим спискам, ища общий документ
5. Для каждого найденного документа вычисляет степень релевантности

Как работает клиент?

1. Разбирает запрос на слова
2. Переводит слова в их идентификаторы
3. Для каждого слова находит в обратном индексе список документов, его содержащих
4. Одновременно бежит по этим спискам, ища общий документ
5. Для каждого найденного документа вычисляет степень релевантности
6. Сортирует образовавшийся список по релевантности

Качество поиска

Как оценить качество поиска?

- **Полнота:** отношение количества найденных релевантных документов к общему количеству релевантных документов

Качество поиска

Как оценить качество поиска?

- **Полнота:** отношение количества найденных релевантных документов к общему количеству релевантных документов
- **Точность:** доля релевантных документов в общем количестве найденных документов

Качество поиска

Как оценить качество поиска?

- **Полнота:** отношение количества найденных релевантных документов к общему количеству релевантных документов
- **Точность:** доля релевантных документов в общем количестве найденных документов
- **Benchmarks:** показатели системы на контрольных запросах и специальных коллекциях документов

Качество поиска

Как оценить качество поиска?

- **Полнота:** отношение количества найденных релевантных документов к общему количеству релевантных документов
- **Точность:** доля релевантных документов в общем количестве найденных документов
- **Benchmarks:** показатели системы на контрольных запросах и специальных коллекциях документов
- **Оценка экспертов**

План лекции

1. Архитектура поисковых систем
2. **Алгоритмы поисковых систем**
   - PageRank
   - Собственные смыслы
3. Поисковая оптимизация

PageRank: постановка задачи

Хотим для каждой страницы сосчитать показатель ее “качества”.

PageRank: постановка задачи

Хотим для каждой страницы сосчитать показатель ее “качества”.

Идея [Брин, 1998]: Определить рейтинг страницы через количество ведущих на нее ссылок и рейтинг ссылающихся страниц

PageRank: постановка задачи

Хотим для каждой страницы сосчитать показатель ее “качества”.

Идея [Брин, 1998]: Определить рейтинг страницы через количество ведущих на нее ссылок и рейтинг ссылающихся страниц

Другие методы:

- Учет частоты обновляемости страницы
- Учет посещаемости
- Учет регистрации в каталоге — спутнике поисковой системы

Модель случайного блуждания

Сеть:

- Вершины
- Ориентированные ребра (ссылки)

Модель случайного блуждания

Сеть:

- Вершины
- Ориентированные ребра (ссылки)

Передвижение пользователей по сети

- Стартуем в случайной вершине
- С верятностью ε переходим в *случайную* вершину
- С вероятностью $1 - \varepsilon$ переходим по случайному исходящему ребру

Модель случайного блуждания

Сеть:

Вершины

Ориентированные ребра (ссылки)

Передвижение пользователей по сети

Стартуем в случайной вершине

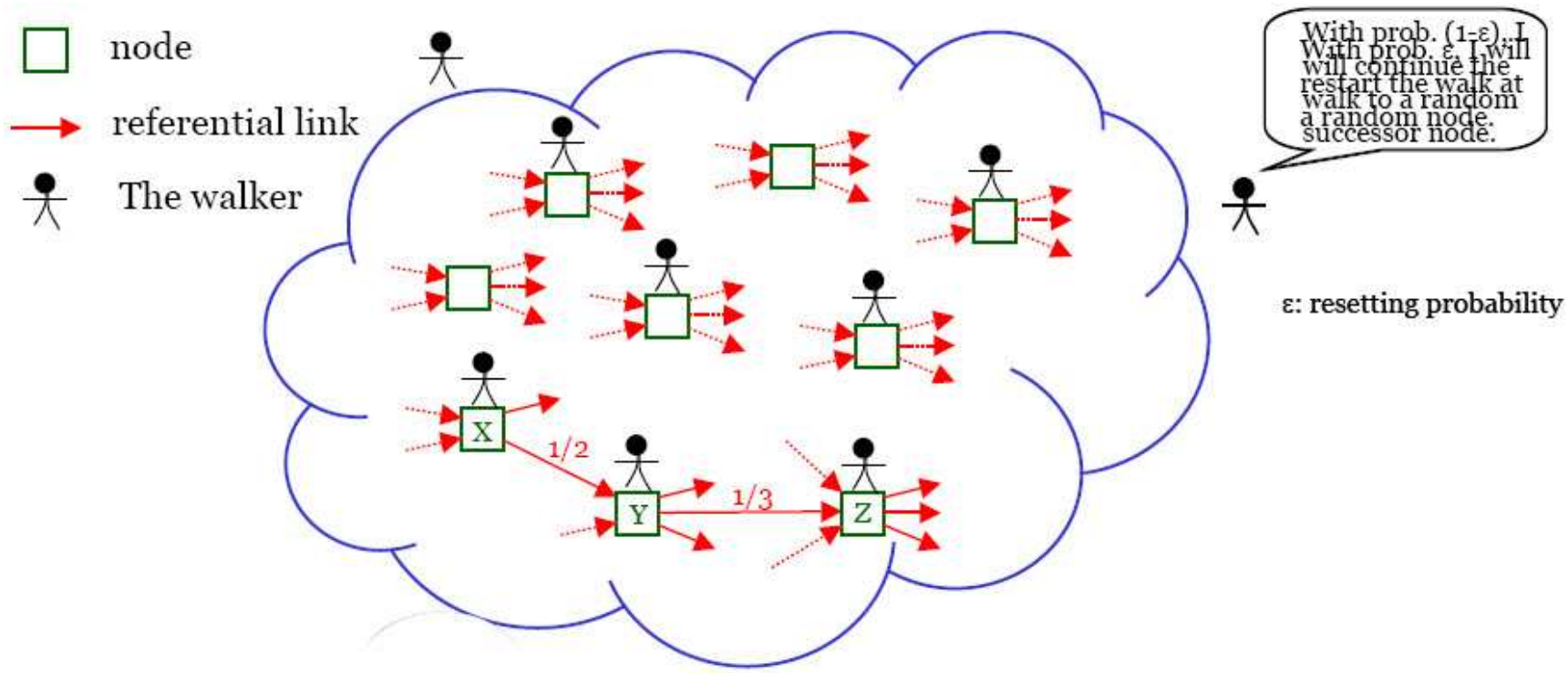
С верятностью ε переходим в *случайную* вершину

С вероятностью $1 - \varepsilon$ переходим по случайному исходящему ребру

Предельные вероятности

Для каждого k можно определить $PR_k(i)$ как вероятность оказаться в вершине i через k шагов

Факт: $\lim_{k \to \infty} PR_k(i) = PR(i)$, то есть для каждой вершины есть предельная вероятность находится именно в ней


node
referential link
The walker
With prob. (1-ε), I will continue the walk to a random successor node.
With prob. ε, I will restart the walk at a random node.
ε: resetting probability
X
Y
Z
1/2
1/3

Основное уравнение PageRank

Пусть $T_1, \ldots, T_n$ — вершины, из которых идут ребра в i, $C(X)$ — обозначение для исходящей степени вершины X.

Утверждение: $PR(i) = \varepsilon + (1 - \varepsilon) \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)}$

Основное уравнение PageRank

Пусть $T_1, \ldots, T_n$ — вершины, из которых идут ребра в i, $C(X)$ — обозначение для исходящей степени вершины X.

Утверждение: $PR(i) = \varepsilon + (1 - \varepsilon) \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)}$

Кто может доказать?

Основное уравнение PageRank

Пусть $T_1, \ldots, T_n$ — вершины, из которых идут ребра в i,
$C(X)$ — обозначение для исходящей степени вершины X.

Утверждение: $PR(i) = \varepsilon + (1 - \varepsilon) \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)}$

Кто может доказать?

По определению $PR_k(i)$ верно следующее:

$PR_0(i) = 1/N$

$PR_k(i) = \varepsilon + (1 - \varepsilon) \sum_{i=1}^{n} \frac{PR_{k-1}(T_i)}{C(T_i)}$

Нужно просто перейти к пределу!

Основное уравнение PageRank

Пусть $T_1, \ldots, T_n$ — вершины, из которых идут ребра в i, $C(X)$ — обозначение для исходящей степени вершины X.

Утверждение: $PR(i) = \varepsilon + (1 - \varepsilon) \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)}$

Кто может доказать?

По определению $PR_k(i)$ верно следующее:

$PR_0(i) = 1/N$

$PR_k(i) = \varepsilon + (1 - \varepsilon) \sum_{i=1}^{n} \frac{PR_{k-1}(T_i)}{C(T_i)}$

Нужно просто перейти к пределу!

Практическое решение: вместо $PR(i)$ используют $PR_{50}(i)$, вычисленное по итеративной формуле.

PageRank как собственный вектор

Определим матрицу L:

Если нет ребра из i в j, то $l_{ij} := \varepsilon$

Если ребро есть, то $l_{ij} := \varepsilon + (1 - \varepsilon) \cdot \frac{1}{C(j)}$

PageRank как собственный вектор

Определим матрицу L:

Если нет ребра из i в j, то $l_{ij} := \varepsilon$

Если ребро есть, то $l_{ij} := \varepsilon + (1 - \varepsilon) \cdot \frac{1}{C(j)}$

Введем обозначения:

$\overline{PR_k} = (PR_k(1), \ldots, PR_k(N))$

$\overline{PR} = (PR(1), \ldots, PR(N))$

PageRank как собственный вектор

Определим матрицу L:

Если нет ребра из i в j, то $l_{ij} := \varepsilon$

Если ребро есть, то $l_{ij} := \varepsilon + (1 - \varepsilon) \cdot \frac{1}{C(j)}$

Введем обозначения:

$\overline{PR_k} = (PR_k(1), \ldots, PR_k(N))$

$\overline{PR} = (PR(1), \ldots, PR(N))$

Получаются соотношения:

$PR_k = L^k \, PR_0$

$PR = L \, PR$

PageRank как собственный вектор

Определим матрицу L:

Если нет ребра из i в j, то $l_{ij} := \varepsilon$

Если ребро есть, то $l_{ij} := \varepsilon + (1 - \varepsilon) \cdot \frac{1}{C(j)}$

Введем обозначения:

$\overline{PR_k} = (PR_k(1), \ldots, PR_k(N))$

$\overline{PR} = (PR(1), \ldots, PR(N))$

Получаются соотношения:

$PR_k = L^k \, PR_0$

$PR = L \, PR$

Векторная модель

Пусть есть коллекция *документов*, каждый является последовательностью *слов*.

Определим матрицу M по формуле $M_{ij} = TF_{ij} \cdot IDF_i$, где:

- Частота терма TF_{ij} — относительная доля слова i в тексте j
- Обратная встречаемость в документах IDF_i — величина, обратная количеству документов, содержащих слово i

Векторная модель

Пусть есть коллекция *документов*, каждый является последовательностью *слов*.

Определим матрицу M по формуле $M_{ij} = TF_{ij} \cdot IDF_i$, где:

- Частота терма TF_{ij} — относительная доля слова i в тексте j
- Обратная встречаемость в документах IDF_i — величина, обратная количеству документов, содержащих слово i

Физический смысл M_{ij} — степень соответствия слова i тексту j

Сингулярное разложение матриц

Определение:
Пусть A — матрица $m \times n$. Разложение $A = USV$ называется **сингулярным**, если U — ортогональная матрица $m \times m$, V — ортогональная матрица $n \times n$, а S — диагональная матрица $m \times n$

Сингулярное разложение матриц

Определение:
Пусть A — матрица $m \times n$. Разложение $A = USV$ называется **сингулярным**, если U — ортогональная матрица $m \times m$, V — ортогональная матрица $n \times n$, а S — диагональная матрица $m \times n$

Факт: Для каждой матрица A можно вычислить сингулярное разложение, причем числа, стоящие на диоганали S — корни из собственных чисел $AA^{\intercal}$

Сингулярное разложение матриц

Определение:
Пусть A — матрица $m \times n$. Разложение $A = USV$ называется **сингулярным**, если U — ортогональная матрица $m \times m$, V — ортогональная матрица $n \times n$, а S — диагональная матрица $m \times n$

Факт: Для каждой матрица A можно вычислить сингулярное разложение, причем числа, стоящие на диоганали S — корни из собственных чисел $AA^{\intercal}$

Факт: Если S' — матрица S, в которой оставили только k наибольших чисел, то $US'V$ — самое близкое приближение матрицы A имеющее ранг k.

Применение сингулярного разложения

- Размер матрицы документы-слова имеет порядок $10^{10} \times 10^6$

Применение сингулярного разложения

- Размер матрицы документы-слова имеет порядок $10^{10} \times 10^6$
- Построим приближение ранга 100

Применение сингулярного разложения

- Размер матрицы документы-слова имеет порядок $10^{10} \times 10^6$
- Построим приближение ранга 100
- Выделим 100 линейно-независимых строк, назовем их **обобщенными словами**

Применение сингулярного разложения

- Размер матрицы документы-слова имеет порядок $10^{10} \times 10^6$
- Построим приближение ранга 100
- Выделим 100 линейно-независимых строк, назовем их **обобщенными словами**
- Выделим 100 линейно-независимых столбцов, назовем их **обобщенными документами**

Применение сингулярного разложения

- Размер матрицы документы-слова имеет порядок $10^{10} \times 10^6$
- Построим приближение ранга 100
- Выделим 100 линейно-независимых строк, назовем их **обобщенными словами**
- Выделим 100 линейно-независимых столбцов, назовем их **обобщенными документами**
- Каждый документ можно представить как какую-то линейную комбинацию обобщенных документов, тоже верно и для слов

Применение сингулярного разложения

- Размер матрицы документы-слова имеет порядок $10^{10} \times 10^6$
- Построим приближение ранга 100
- Выделим 100 линейно-независимых строк, назовем их **обобщенными словами**
- Выделим 100 линейно-независимых столбцов, назовем их **обобщенными документами**
- Каждый документ можно представить как какую-то линейную комбинацию обобщенных документов, тоже верно и для слов

Применение сингулярного разложения

- Размер матрицы документы-слова имеет порядок $10^{10} \times 10^6$
- Построим приближение ранга 100
- Выделим 100 линейно-независимых строк, назовем их **обобщенными словами**
- Выделим 100 линейно-независимых столбцов, назовем их **обобщенными документами**
- Каждый документ можно представить как какую-то линейную комбинацию обобщенных документов, тоже верно и для слов

Обобщенные слова и документы также называют “собственными смыслами”

Использование собственных смыслов

Для каких задач применяются собственные смыслы:

- Поиск “похожих” слов, “похожих” документов

Использование собственных смыслов

Для каких задач применяются собственные смыслы:

- Поиск “похожих” слов, “похожих” документов
- Тематическая классификация документов

Использование собственных смыслов

Для каких задач применяются собственные смыслы:

- Поиск “похожих” слов, “похожих” документов
- Тематическая классификация документов
- Фильтрация документов

План лекции

1. Архитектура поисковых систем

2. Алгоритмы поисковых систем
   - PageRank
   - Собственные смыслы

3. **Поисковая оптимизация**

Поисковая оптимизация — это

Поисковая оптимизация

- Выводит ваш сайт в первую десятку ответов поисковых систем по ключевым запросам
- Без гарантии!
- Другие действия, направленные на привлечение целевой аудитории

Предварительные этапы

Первые шаги:

- Анализ тематического сегмента
- Анализ сайта
- Поиск ниши позиционирования
- Составление семантического ядра запросов

Оптимизация под аудиторию

Работа с содержанием:

- Рекомендации по контентному наполнению
- Рекомендации по предоставлению дополнительных сервисов
- Специальные тексты (SEO копирайтинг)

Продвижение

Оптимизация под поисковые запросы:

- Подготовка сайта к индексации (запрет на индексирование избыточной информации)
- Корректировка структуры сайта с учетом юзабилити
- Изменение архитектуры сайта для облегчения его индексации
- Работа над внутренними факторами (тэги "title"и "description", основной текст)
- Визуальное и архитектурное ("тэговое") выделение ключевых слов

Еще продвижение

Оптимизация под поисковые запросы (продолжение):

- Корректировка текстов с учетом ключевых слов, общей читаемости текста и его эмоционального восприятия
- Подготовка вариантов описаний сайта для регистрации в каталогах и обмена ссылками
- Работа над внешними факторами (регистрации в поисковиках, каталогах, повышение тематической авторитетности сайта: ссылки, новости, пресс-релизы, публикации)

Сопровождение

Дальнейшие шаги:

- Анализ достигнутых результатов и дальнейшая корректировка
- Выбор рекламных площадок и размещение рекламы

Ссылки по SEO

Где учиться:

- SearchEngines.Ru
- SearchEngineWatch.Com
- http://seotext.ru
- SEO in Wikipedia.Org
- Ralph Wilson Checklist
- Энциклопедия Интернет Рекламы

Эффективный поиск в интернете:

- http://logic.pdmi.ras.ru/~yura/search.html

Задачи

Докажите, что сходимость $\lim_{k \to \infty} PR_k(i) = PR(i)$ действительно имеет место

Последний слайд

Если не запомните ничего другого:

- Поисковые системы состоят из робота, системы управления базой данных и клиента

Последний слайд

Если не запомните ничего другого:

- Поисковые системы состоят из робота, системы управления базой данных и клиента
- В основе алгоритма определения релевантности лежит вычисление PageRank'а

Последний слайд

Если не запомните ничего другого:

- Поисковые системы состоят из робота, системы управления базой данных и клиента
- В основе алгоритма определения релевантности лежит вычисление PageRank'а
- Поисковая оптимизация — набор рекомендаций по выведению ваших сайтов в десятку ответов

Последний слайд

Если не запомните ничего другого:

- Поисковые системы состоят из робота, системы управления базой данных и клиента
- В основе алгоритма определения релевантности лежит вычисление PageRank'а
- Поисковая оптимизация — набор рекомендаций по выведению ваших сайтов в десятку ответов

Последний слайд

Если не запомните ничего другого:

- Поисковые системы состоят из робота, системы управления базой данных и клиента
- В основе алгоритма определения релевантности лежит вычисление PageRank'а
- Поисковая оптимизация — набор рекомендаций по выведению ваших сайтов в десятку ответов

Вопросы?