1. Основные понятия информационного поиска

- Сфера науки, которая исследует методы поиска информации, называется информационный поиск (information retrieval (IR))

Документ - материальный объект, содержащий информацию в зафиксированном виде и специально предназначенный для её передачи во времени и пространстве.

Общие свойства
- Значительное текстовое содержание
- Некоторая структура:
    - заголовок,автор, дата - для статей;
    - тема, отправитель, адресат - для писем

Примеры:Интернет-страницы, электронные письма, книги, новости, посты форумов, патенты и многое другое

Записи базы данных (структурированные таблицы) состоят из хорошо определенных полей и атрибутов

Примеры: банковские записи балансы, номера счетов, имена, адреса, даты рождения, номера социального обеспечения

(опционально: сравнение документов и записей)

1. (запись) Легко сопоставлять запросы и поля таких баз данных (хорошо определенная семантика)

2. (документ) Текст более сложный – неструктурированная информация

Сопоставление текста запроса с текстом документа и определение того, что такое хорошее сопоставление – **базовый вопрос информационного поиска**.

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

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

Важные понятия:

- **Relevance – релевантность**
    - 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)
- **Evaluation - оценка качества**
    - Экспериментальные процедуры и меры для сравнения результатов работы систем с ожиданиями пользователей
    - Методы оценки качества поиска сейчас используются во многих областях
    - Типично используются тестовые коллекции документов, запросов, и оценки релевантности
    - *Полнота и точность – простые примеры оценки качества*
- **Users and Information Needs** – потребность пользователя, 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*

- **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

2. Виды поисковых систем по охвату и направленности. Особенности разных типов поисковых систем

По охвату

1. Интернет-поиск
2. Корпоративный поиск

Сравнение

- Интернет-поиск
  - Собирает результаты по общедоступному Интернету
  - Проблема 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

По типу

- Патентный поиск
- Медицинский поиск
- Поиск научных публикаций
- Поиск по законодательству
- Поиск по химическим документам
- …

3. Особенности научного поиска

Проблемы:

- Поиск научной литературы
- Рост публикаций
- Устаревание
- Проблема литературы до интернетовской эпохи
- Цитаты, которые можно использовать как ссылки в Интернет-поиске
  - Большое количество ссылок на работу – фактор значимости работы
  - Наукометрия

Причины цитирования в научной работе:

- Affirmational – базируется на цитируемой работе
- Assumptive **–** цитаты основоположников
- Conceptual – используются понятия, определения
- Contrastive – сравнение с близкими, но не совпадающими подходами
- Methodological – используется метод, оборудование, инструменты
- Negational – подвергает сомнению
- и др.

Анализ ссылок в литературе

• Количество ссылок на работу
  − Импакт-фактор работы
• Сходство работ на основе ссылок
  − Если работа A цитирует работы B и C, то, возможно, имеется сходство между работами B и C
  − Если работы B и C цитируют одну и ту же работу A, то возможно есть сходство между этими работами
  − Если таких совпадений много, то связь между работами

Системы поиска научной литературы

- Scopus
- CiteCeer
- ScienceDirect
- Microsoft Academic Search
- Google Scholar
- IEEE Xplore

4. Основные этапы обработки текстов в поисковой машине

1. Извлечение текстов
   Идентифицирует и сохраняет тексты для индексирования
2. Трансформация текстов
   Трансформирует документы в индексные термы (parser, стоп-слова, стемминг, анализ ссылок, извлечение информации, классификация по темам)
3. Создание индексов
   Берет индексные термы и создает индексы для быстрого поиска (статистика по документам, определение весов, инвертирование, распределенное хранение индекса)

5. Основные этапы обработки запроса в поисковой машине

1. Взаимодействие с пользователем
   Поддерживает создание и уточнение запроса, показ результатов (интерфейс и парсер для языка запроса, трансформация запроса: спеллчек, подсказки, расширение запроса; выдача результатов: подсветка, релевантная реклама и пр)
2. Ранжирование
   Использует запрос и индексы, чтобы породить ранжированный лист документов (для каждого документа есть вес соответствия запросу; оптимизация выполнения запроса; распределенное выполнение)

3. Оценка качества
   Мониторит и измеряет качество поиска

6. Что такое графематический анализ? Что такое лемматизация

Графематический анализ - это программа начального анализа естественного текста, представленного в виде цепочки ASCII символов, вырабатывающая информацию, необходимую для дальнейшней обработки Морфологическим и Синтаксическим процессорами.

Графематический анализ:

1. Разделение текста на слова, разделители
2. Выделение устойчивых оборотов, не имеющих словоизменительных вариантов
3. Выделение предложений
4. Выделение абзацев
5. Выделение дат
6. Определение языка слова (русский, нерусский)
7. Определение формата написания слова (прописные, строчные буквы)

Морфологический анализ текста осуществляет приведение словоформ, встречающихся в тексте, к нормальному (словарному) виду и определяет морфологические характеристики словоформы

Лемматизация – тип морфологического анализа – приведение к нормальной форме.
Лесной, лесного, лесному->лесной
леса -> лес
Танцующая -> танцевать

Упрощенная процедура: лемматизация
Лемматизация – приведение слова к нормальной форме (лесные -> лесной)
Стемминг –выделение псевдоосновы (лесной -> лес) Обратная процедура: морфологический синтез.

7. Как работает словарный морфологический анализ?

Методы морфологического анализа:

1. *Словарный:*

- со словарем словоформ
  Каждой словоформе поставлена в соответствие основа или лемма
- со словарем основ

2. бессловарный (фактически – со словарем псевдоокончаний) + анализ по аналогии («предсказание»)

Схема морфологического анализа со словарем:
Для неслужебных слов:

1. Выделить возможные окончания слова длиной от 0 до 3 символов
2. Для каждого полученного окончания определить код окончания по таблице окончаний и номер флективного класса по словарю основ (лемм)
3. Если номер флективного класса и номер окончания найдены, то проверить их согласованность по морфологической таблице
4. Если согласованность подтверждается, то сохранить данный вариант

Проблемы:

1. Дают максимально полный анализ словоформы ?чем плохо?
2. На реальных текстах дают сбои (опечатки, уникальные слова)
3. Не существует абсолютно полных словарей – лексика языка непрерывно пополняется
4. Для примера – невозможно включить в словарь всю существующую терминологию, имена, фамилии и т.д.

8. Как морфологические анализаторы обрабатывают слова, отсутствующие в словаре

Методы морфологического анализа:

1. словарный
   • со словарем словоформ
   Каждой словоформе поставлена в соответствие основа или лемма
   • со словарем основ
2. бессловарный (фактически – со словарем псевдоокончаний)
   + анализ по аналогии («предсказание»)

Предсказание в морфологическом анализе:

- Функциональное назначение предсказания – морфологический анализ слов (словоформ), отсутствующих в словаре
- Метод предсказания – выявление аналогий со словоформами, распознаваемыми имеющимся словарем

Используется метод предсказаний: *Метод предсказания* – выявление аналогий со словоформами, распознаваемыми имеющимся словарем.

Алгоритм:

1. предсказание префиксального образования
   - Попытка найти существующую словоформу языка, которая максимально совпадала бы справа со входным словом.
   - Если левая часть (потенциальный префикс) не длиннее M символов (пяти), а правая часть (совпавшая с известной словоформой) не короче N символов

(четырех), то слово разбирается по образцу известной словоформы. [евро]технологию, [супер]коньками.

2. предсказание по концовке, взятой из известных словоформ
   - Отделяются инвертированные концовки известных словоформа – длины K (пять букв),
     Сопоставляются с морфологическими характеристиками:
     «анием» - как «ср. род, ед. ч., тв. пад.»

   - Такая строка заносится в исходный лексикон, если она встречается:
     - не менее L раз (трех) и
     - чаще конкурентов в пределах одной части речи

   - ВСЕГДА предусматривается разбор именем существительным, хотя бы неизменяемым.

Предсказание по префиксу:

- попытка найти существующую словоформу языка, которая максимально совпадала бы справа со входным словом.
- Если левая часть (потенциальный префикс) не длиннее M символов (пяти), а правая часть (совпавшая с известной словоформой) не короче N символов(четырех), то слово разбирается по образцу известной словоформы.
  [евро]технологию, *[супер]коньками*

Предсказание по концовке известной словоформы:
Отделяются инвертированные концовки известных словоформа – длины K (пять букв), сопоставляются с морфологическими характеристиками:

- *«анием»* - как «ср. род, ед. ч., тв. пад.»

Такая строка заносится в исходный лексикон, если она встречается:

- не менее L раз (трех) и
- чаще конкурентов в пределах одной части речи

ВСЕГДА предусматривается разбор именем существительным, хотя бы неизменяемым.

9. Что такое постморфологический анализ. Основные методы.

1. = предсинтаксический анализ
2. Предназначен для устранения морфологической омонимии (многозначности) слов
   - Выбор правильной леммы
   - Уточнение морфологических характеристик

Основные методы

1. Написание правил,
2. Статистические методы, прежде всего, на основе морфологически размеченного корпуса

Примеры правил:

- Удаление признаков служебных частей речи для однобуквенных слов, за которыми следуют точки
- Удаление омонимов слова «уже», соответствующих прилагательным, если за ним не стоит запятая или слово в родительном падеже
- Удаление омонимов слова «сорока», если после слова следует числительное (сорок пять)
- Обработка предлогов: удаление у слова, следующего за предлогом, всех омонимов, не соответствующих падежам, которыми обычно управляет данный предлог

Морфологическая разметка корпуса:

1. Морфологический анализ всех словоформ текста (информация о морфологических (грамматических) характеристик в виде тегов)
2. Снятие неоднозначностей (или исправление ошибок)
3. Добавление информации о результатах в электронное представление текста

Снимаем омонимию с помощью скрытых марковских моделей:
- встречаемость каждого тега в определенном месте цепочки зависит только от предыдущего тега;
- то, какое слово находится в том или ином месте цепочки, полностью определяется тегом (а не, допустим, соседними словами).
Таким образом, порождение правильно построенной цепочки тегов уподобляется действию конечного автомата, где дуги помечены тегами с приписанными им вероятностями, а слова – это наблюдаемые реализации тегов. Состояния определяются парой «текущий тег + предыдущий тег».

10. Булевская модель информационного поиска. Преимущества и недостатки булевской модели поиска

1. Два возможных результата для сопоставления запроса и документа
    a. TRUE и FALSE
    b. Поиск по полному совпадению
    c. Простейшая форма ранжирования
2. Запрос может специфицироваться посредством Булевых операторов
    a. AND, OR, NOT
    b. Могут быть использоваться операторы близости (proximity)
3. Инвертированный индекс
    a. Для каждого терма храним список номеров документов, где этот терм встречается

Преимущества

1. Результаты предсказуемы, их легко объяснить
2. Могут быть встроены многие различные признаки

3. Эффективная обработка

Недостатки

1. Качество выдачи зависит исключительно от пользователя
2. Простые запросы дают слишком много документов (нет упорядочения)
3. Длинные запросы сложно составить

11. Как измеряется качество булевского поиска

Оценка булевского поиска

- *Специфика*: булевский поиск не имеет ранжирования (упорядочения)

1. Поисковая система разделяет коллекцию на два множества
    - Разделение по принципу выдано в ответ на запрос или нет
    - Разделение по принципу релевантен документ или нет
2. Вычисляются меры качества:
    - *Обозначения*

Доли различных видов документов среди всех

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- *Точность (Precision, P)*

Доля релевантных документов в выдаче

P(relevant | retrieved) = $P = \frac{tp}{tp + fp}$

- *Полнота (Recall, R)*

Доля выданных документов среди релевантных

P(retrieved | relevant) = $R = \frac{tp}{tp + fn}$

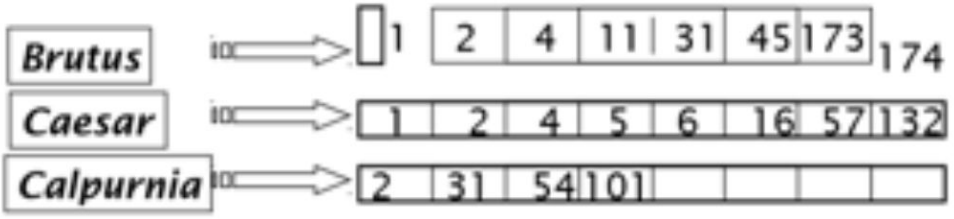
- *F-мера*

$F1 = \frac{2}{1/R + 1/P} = \frac{2PR}{P + R}$

12. Алгоритм сопоставления запроса с документами (Алгоритм Merge)

Для каждого терма t нужно хранить список документов, которые содержат t. Каждый документ идентифицируется номером документа —

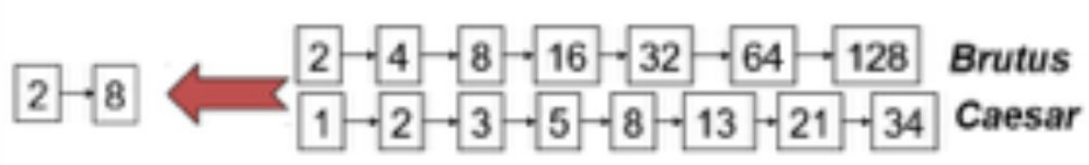
docID:



Обработка запроса: AND

• Запрос: «Brutus AND Caesar»

- Найти Brutus в словаре, извлечь все его docid. –Найти Caesar в словаре, извлечь все его docid. –“Merge” (пересечь) docid :



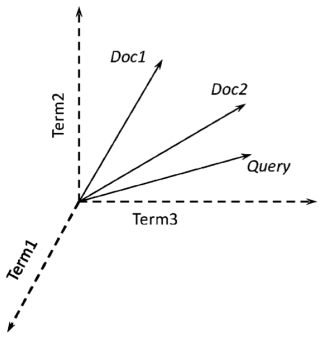

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

Если длины списков x и y, то пересечение O(x+y) операций
Необходимо: записи в списке должны быть отсортированы по docId

13. Что такое векторная модель информационного поиска?

Общие сведения

Рассмотрим v-мерное векторное пространство, v - количество слов. Слова - отдельные позиции (оси) в векторах. Каждый документ и запрос - вектор в N^v. Введем понятие близости документов. Предположение: чем меньше угол между векторами, тем больше они похожи.



Мера сходства документов
Для определения близости документов определяется, насколько одинаково направлены соответствующие вектора документов (то есть, насколько мал угол между ними). Используется косинусная мера:

$$q,\ d\ \in N^{|v|} \Rightarrow\ \rho(q,d)\ = cos(q,d)\ =\ \frac{q*d}{|q|*|d|} =\ \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2}\sqrt{\sum_i d_i^2}}$$

Если векторы q , d уже нормализованы, то

$$\rho(q,d)\ = cos(q,d)\ =\ \frac{q*d}{|q|*|d|} = \sum_i q_i d_i$$

Построение векторного пространства
Определим представление документа в векторном пространстве. Необходимо учесть долю пересечения слов запроса с документом.

14. Поясните смысл показателей *idf* и *tf.idf*. Способы вычисления tf.

Векторное представлением строится на основе двух факторов.

- *Первый фактор:* матрица частоты употребления терма в документе. Рассмотрим число вхождений терма в документ. Если всего v термов, каждый документ - вектор частот в N^v , обозначается tf (term frequency). Совокупность векторов частот порождает матрицу частот всех термов во всех документов.
- *Второй фактор:* подокументная частотность. Частотные слова менее информативны, чем редкие, самые частотные слова в документе - служебные. Сверхчастотные слова типа: предлоги, союза, которые есть во многих документах вообще иногда рассматриваются как стоп-слова и выбрасываются из документа.Чтобы учесть эту распространенность - вводится фактор, df - количество документов, в которых употреблялось это слово - подокументная частотность.

Применение факторов для построения представления

- **Вес idf**. t - терм, df_t - подокументная частотность t: количество документов, содержащих t. Таким образом, df - обратная мера информативност, df<=N, N - число документов. Определяем idf (обратная подокументная частота):

$$idf_t = log_{10}(N/df_t)$$

Логарифм используется для смягчения разницы в употреблении более и менее частотных слов. Данная величина вычисляется для каждого слова во всех коллекции документов

- **tf.idf взвешивание.** Произведение tf на idf. Tf - количество входждений терма в документ, вычисляется отдельно для каждого документа и терма (первый фактор).Увеличивается при росте количества упоминаний в документе. Увеличивается для редких терминов в коллекции. Каждый документ представлен вещественным вектором размерностью число разных слов в коллекции.
- **Модификация 1 tf.idf.** $w_{t,d} = log(1 + tf_{t,d}) * idf_t$

Варианты весов

- Базовый – это сколько раз слово встретилось в документе (count),
- Логарифмирование count

$$w_{t,d} = \log(1 + tf_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

- Нормализация первого сомножителя
  - 1) tf/tf_{max} или $(\alpha + (1-\alpha)\ tf/tf_{max})$
  - 2) tf/|d|, где |d| - количество слов в документе (вариант, описанный в Википедии)

Okapi BM25

- BM – best match

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d / L_{\mathrm{ave}})) + \mathrm{tf}_{td}} \cdot \frac{(k_3 + 1)\mathrm{tf}_{tq}}{k_3 + \mathrm{tf}_{tq}}$$

- tf_{tq}: частота слова в запросе q
- k_3: параметр, контролирующий частоту термина в запросе
- Нет нормализации запроса по длине (поскольку поиск делается для фиксированного запроса)
- Параметры нужно настраивать на коллекции
- Если оптимизация не выполнялась, то в экспериментах получено, что величины k_1 и k_3 должны иметь значения в промежутке [1.2, 2], b = 0.75

46

15. Классическая процедура оценки качества информационно поиска

Классическая процедура оценки (Cranfield, Крендфилдские эксперименты) :

1. Составим список запросов и ограничим коллекцию документов
2. Для каждой пары запрос/документ выставим экспертную оценку “релевантности”

3. Будем рассматривать ответ системы как множество/последовательность оценок релевантности
4. На полученной последовательности/множестве оценок релевантности построим метрики

Оценка релевантности выдачи
- Информационная потребность выражается запросом
- Релевантность оценивается по отношению к информационной потребности, а не к словам запроса (даже, если все слова запроса присутствуют в тексте, текст может быть не релевантен)

Оценка ранжированных результатов:
- Современные системы выдают упорядоченные результаты
- Выдача может быть достаточно большой
- Релевантные документы должны выдаваться раньше нерелевантных
- Можно измерять точность на каждом уровне полноты

16. Что такое РОМИП, какие задачи в нем решаются?

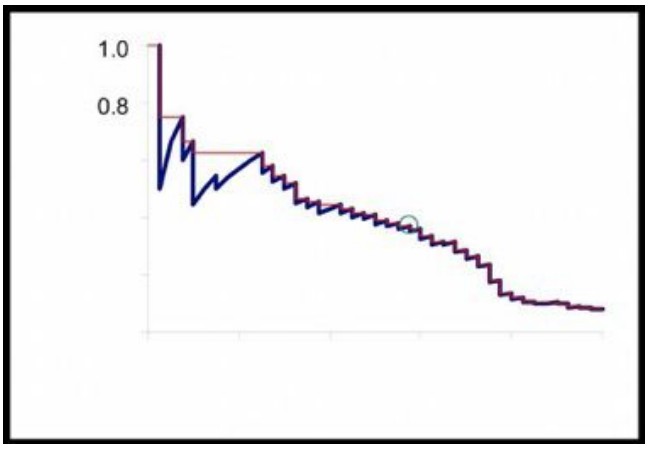
Российский семинар по оценке Методов Информационного Поиска (РОМИП) (2003) - это открытый семинар, проводимый ежегодно с 2003 года группой российских исследователей и разработчиков, занимающихся информационным поиском. Основная цель семинара — создание плацдарма для проведения независимой оценки методов информационного поиска, ориентированных на работу с русскоязычной информацией.

Структурно семинар представляет собой набор дорожек — секций, посвященных конкретным проектам с определенной задачей и правилами оценки. Оргкомитет формирует тестовые наборы данных, заданий и распространяет их участникам.

Примеры задач:
1. текстовый поиск (поиск по запросу)
2. текстовая классификация
    a. веб-сайты
    b. веб-страницы
3. контекстно-зависимое аннотирование (составление аннотации документа по запросу)
4. поиск изображения по образцу
5. выявление нечетких дубликатов изображений
6. построение текстовых меток по изображению
7. автоматический анализ тональности текста (2011г) (эмоциональная оценка, выраженная в тексте)

17. Что такое кривая полнота-точность?

Усреднение по запросам:
- Кривая полнота-точность для одного запроса не очень интересна
- Нужно построить кривую полнота-точность для совокупности запросов
    - Пока Кривая – это совокупность точек
    - Как интерполировать?

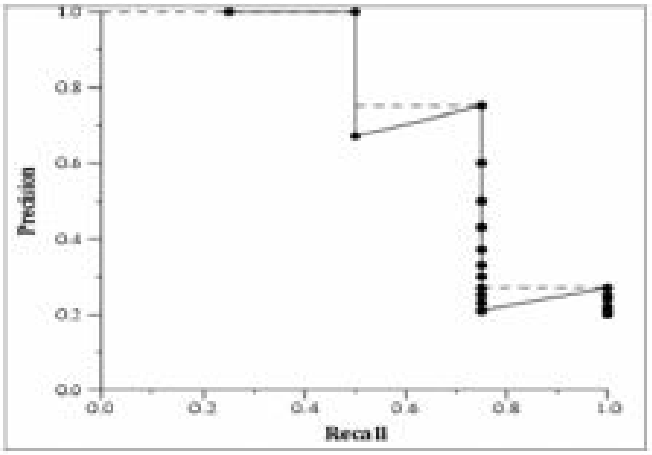


Кривая полнота-точность.
Интерполированная точность:
Идея: Если локально точность возросла с увеличением полноты, то засчитаем ее максимум…
Т.е. берем максимум точности справа на графике

18. Что такое 11-точечный график TREC?

11-точный график TREC:

- Значения полноты от 0 до 1 с шагом 0.1
- Интерполяция точности
- Если $r_t > recall(q_j)$, то $p(r_t, q_j) = 0$
- Если $r_t \leq recall(q_j)$, то $p(r_t, q_j) = max_n(precision(n))$, $n \geq pos(r_t, q_j)$



19. Что такое пулинг в информационном поиске? Сложности, связанные с пулингом

Пулинг VS Полнота
Для каждого запроса:

1. Собрать результаты систем участников глубины A
2. Выбрать из полученных результатов B первых
3. Удалить дубликаты
4. Проставить оценки релевантности

5. Не оцененные документы считать нерелевантными
6. Оценить весь ответ системы (с глубиной A)

Сложности:
- Взаимное усиление систем
- Недооценка систем, не участвовавших в оценке
- Получаемая оценка – оценка снизу
- Но: участники относительно в равных условиях

20. Оценка качества в поисковых машинах

1. Полноту невозможно измерить
2. K- первых документов
3. Релевантные документы должны показываться раньше
4. NDCG (Normalized Cumulative Discounted Gain)
5. Использование кликов пользователей
    a. A/B testing

21. Шкалы оценок. Мера NDCG

Пусть есть запрос p.
Каждому предложению ставится оценка, в соответствии с его релевантностью запросу по шкале оценок.
gi - оценка для i-го элемента выдачи
DCGp - характеристика выдачи
С помощью NDCG можно сравнивать на качество различные поисковики.

Шкалы оценок:
- В прошлом: TReC – бинарные
- Сейчас TReC:
    - Высоко релевантный
    - Релевантный
    - Нерелевантный
- РОМИП
    - Соответствует
    - Скорее соответствует
    - Возможно соответствует
    - Не соответствует
    - Не может быть оценен

Оценка качества выдачи по небинарным оценкам:
- Предположения
    - Лучше, если релевантные документы находятся в начале списка
    - Если есть несколько типов релевантных документов, то лучше, чтобы документы с высокими оценками были раньше в списке
- Существует наилучшее упорядочение расположения оценок от лучших к худшим

- В суммированной оценке выдачи каждая следующая позиция в списке должна давать меньший вклад, чем предыдущая

Cumulative gain: $CG_\lambda = \sum_{i=1}^{\lambda} g_i$

Discounted Cumulative Gain: $DCG_\lambda = g_1 + \sum_{i=2}^{\lambda} \frac{g_i}{log}$, g_i - значение релевантности для i-ого документа в выдаче

nDCG: Нормализация DCG по отношению к лучшему упорядочению по данному запросу:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

22. Что такое информационно-поисковые тезаурусы? Зачем они нужны? Где применяются сейчас

Тезаурус – словарь, в котором
- Единицы соответствуют близким по смыслу словам (понятия, дескрипторы и др.)
- Между единицами установлены формализованные отношения

Типы тезаурусов
- Информационно-поисковые тезаурусы
- Тезаурусы типа WordNet
- И др.

Информационно-поисковый тезаурус – нормативный словарь терминов предметной области, создаваемый для улучшения качества информационного поиска в данной предметной области
- Нормативные ключевые слова
- Национальные и международные стандарты
- Используются в ряде международных организаций и парламентский организаций
    - Европейский парламент – EUROVOC
    - ООН – UNBIS Thesaurus

Цели:
1. Перевод языка авторов на нормативный язык, используемый для индексации и поиска
2. Обеспечение последовательности в присваивании индексных терминов
3. Обозначение отношений между терминами
4. Облегчение информационного поиска

Применение
- Информационный поиск
    - Корпоративные или предметно-ориентированные системы

– Автоматическое расширение запроса
– Визуализация выдачи
• Автоматическая рубрикация текстов
– Несколько десятков рубрикаторов
• Автоматическая кластеризация текстов
• Автоматическое реферирование текстов
– Одного документа, многих документов, составление аналитических отчетов
• Системы мониторинга

23. Назовите методы расширения запросов пользователей при информационном поиске.

- Методы расширения запроса:
  - Глобальные методы
    - Ручные тезаурусы
    - Автоматически порождаемый тезаурус
  - Локальные методы (по конкретному запросу)
    - Relevance feedback (обратная связь по релевантности)
    - Pseudo Relevance feedback (обратная связь по псевдорелевантности

24. Что означает термин *relevance feedback*? Поясните основные принципы работы

Relevance FeedBack:

1. Пользователь оценивает документы в поисковой выдаче
   - Пользователь задает относительно простой, короткий запрос
   - Затем пользователь размечает часть результатов как релевантные и нерелевантные
   - Система вычисляет улучшает соответствие документов запросу на основе пользовательской разметки
   - Процедура может выполняться итеративно.
2. *Основная идея*: сформулировать хороший запрос трудно, если пользователь не знаком с коллекцией, поэтому – итеративное построение запроса

Pseudo Relevance FeedBack:

1. Pseudo-relevance feedback автоматизирует «ручную» часть реального relevance feedback.
2. Pseudo-relevance алгоритм:
   - Строит поисковую выдачу по запросу
   - Предполагает, что первые k документов - релевантны
   - Выполняет relevance feedback
3. В среднем хорошо работает
4. Но может получить очень плохие результаты для некоторых запросов

5. Несколько итераций могут вызвать «искажение запроса»

25. Алгоритм Роккио для *relevance feedback*

Алгоритм Роккио (Rocchio algorithm) — классический алгоритм для реализации метода RF. Он инкорпорирует модель обратной связи по релевантности в модель векторного пространства, описанную в разделе 6.3.

Теория. Мы хотим найти вектор запроса $\vec{q}$, максимально близкий к релевантным документам и минимально похожий на нерелевантные документы. Если C_r — это множество релевантных документов, а C_{nr} — нерелевантных, то целью наших поисков является следующий вектор.[1]

$$\vec{q}_{opt} = \arg\max_{\vec{q}} \left[sim(\vec{q}, C_r) - sim(\vec{q}, C_{nr}) \right] \quad (9.1)$$

• Проблема: мы не знаем все релевантные документы
Особенности:
• Соотношение α vs. β/γ : Если у нас много оцененных документов, то лучше более высокие β/γ.
• Некоторые веса в модифицированном векторе запроса становятся отрицательными
– Отрицательные веса слов игнорируются (устанавливаются равными 0)
• На практике используется:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

• Dr = множество известных релевантных doc векторов
• Dnr = множество известных нерелевантных doc векторов
– Отличны от Cr и Cnr
• qm = модифицированный вектор запроса; q0 = исходный вектор запроса;
α,β,γ: веса
• Новый запрос «сдвигается» по направлению к релевантным документам и «уходит» от нерелевантных документов

26. Порождение и применение автоматического тезауруса для расширения запросов

При автоматической обработке текстов человека - посредника между текстом и описанием его содержания в виде дескрипторов нет. Есть только автоматический процесс и Тезаурус, который должен содержать и те знания, которые содержатся в традиционных информационно-поисковых тезаурусах, и те знания (насколько это возможно), которые использует индексатор для определения основной темы текста.

Именно поэтому традиционные тезаурусы, разработанные для ручного индексирования, трудно использовать при автоматическом индексировании. Разработка тезауруса для автоматического индексирования (далее - АИ тезауруса) характеризуется прежде всего необходимостью описания значительно большего количества слов и словосочетаний, встречающихся в текстах данной предметной области. АИ тезаурус должен не только включать термины, которые представляют важные понятия в текстах данной предметной области, но также охватывать широкий круг более специфических терминов, обнаружение которых в конкретном тексте сделает этот текст релевантным запросу по понятиям более высокого уровня. Например, должны быть описаны не только дескриптор РЫБА и его основные подразделения, такие как МОРСКИЕ РЫБЫ, АНАДРОМНЫЕ РЫБЫ и т.п., но и значительное количество конкретных видов рыб, с тем чтобы текст, обсуждающий проблемы вылова минтая, мог бы быть получен при поиске по термину рыба.
Синонимические ряды понятий должны быть значительно богаче, чем совокупности вариантов дескриптора в тезаурусе для ручного индексирования, поскольку синонимы должны описывать различные способы выражения данного понятия в тексте для автоматического процесса, а не для человека. Ряды синонимов включают в себя не только существительные и именные группы, а также прилагательные, глаголы и глагольные группы. Расширение терминологической базы АИ-тезауруса ведет к необходимости описания многозначных терминов.
Расширение понятийной базы тезауруса ведет к увеличению и усложнению функций отношений между понятиями тезауруса (концептуальными отношениями): возникает необходимость логического вывода отношений, поскольку описать отношения всех дескрипторов со всеми близкими дескрипторами АИ-тезауруса становится трудоемким занятием и затрудняет проверку таких описаний.

27. Вероятностная модель информационного поиска: основная идея, различие с векторной моделью

1.3. Вероятностная модель

В 1977 году Робертсон (Robertson) и Спарк-Джоунз (Sparck-Jones) обосновали и реализовали вероятностную модель. Релевантность в этой модели рассматривается как вероятность того, что данный документ может оказаться интересным пользователю. При этом подразумевается наличие уже существующего первоначального набора релевантных документов, выбранных пользователем или полученных автоматически при каком-нибудь упрощенном предположении. Вероятность оказаться релевантным для каждого следующего документа рассчитывается на основании соотношения встречаемости термов в релевантном наборе и в остальной части коллекции.

Документом будем считать множество слов без учета частоты встречаемости слова в документе. Можно также представить множество в виде обычного булевского вектора $D = \{d_1, \ldots, d_n\}$, где n — количество всех термов, а d_i может принимать значения из множества $\{0, 1\}$. Запросом будем считать множество слов.

Соответствие документа запросу будем строить следующим образом: представим себе, что для каждого фиксированного запроса Q_k у нас имеются распределения вероятностей на всех документах «быть релевантным» и «быть нерелевантным» запросу Q_k. Обозначается это соответственно как $P(R|Q_k, D)$ и $P(\overline{R}|Q_k, D)$. Тогда функцией соответствия будем считать отношение двух этих величин:

$$\frac{P(R|Q_k, D)}{P(\overline{R}|Q_k, D)}.$$

Теперь вспомним теорему Байеса:

$$P(a|b) = P(b|a)\frac{P(a)}{P(b)},$$

где

- $P(a)$ — априорная вероятность гипотезы a;
- $P(b)$ — вероятность наступления события b;
- $P(a|b)$ — вероятность гипотезы a при наступлении события b (апостериорная вероятность);
- $P(b|a)$ — вероятность наступления события b при истинности гипотезы a.

Применим ее для числителя и знаменателя дроби, стоящей в функции соответствия:

$$P(R|Q_k, D) = \frac{P(D|R, Q_k)P(R|Q_k)}{P(D|Q_k)};$$

$$P(\overline{R}|Q_k, D) = \frac{P(D|\overline{R}, Q_k)P(\overline{R}|Q_k)}{P(D|Q_k)};$$

$$\frac{P(R|Q_k, D)}{P(\overline{R}|Q_k, D)} = \frac{P(D|R, Q_k)P(R|Q_k)}{P(D|Q_k)} \frac{P(D|Q_k)}{P(D|\overline{R}, Q_k)P(\overline{R}|Q_k)} = \frac{P(R|Q_k)}{P(\overline{R}|Q_k)} \frac{P(D|R, Q_k)}{P(D|\overline{R}, Q_k)}.$$

Заметим, что первый сомножитель $\frac{P(R|Q_k)}{P(\overline{R}|Q_k)}$ одинаков для всех документов, так как в нем не фигурирует D, и мы его дальше можем не рассматривать. Предполагая независимость всех слов (это очень сильное, и на практике неверное предположение), второй сомножитель можно представить в виде произведения:

$$\frac{P(D|R, Q_k)}{P(D|\overline{R}, Q_k)} = \prod_{i=1}^{n} \frac{P(x_i = d_i|R, Q_k)}{P(x_i = d_i|\overline{R}, Q_k)},$$

где x_i — случайный документ, а d_i — число; $P(x_i = d_i|R, Q_k)$ — вероятность того, что i-й терм будет одновременно присутствовать или отсутствовать у случайного документа, релевантного нашему запросу, так же, как и в документе D. В произведении $\prod_{i=1}^{n} P(x_i = d_i|R, Q_k)$ будут именно те вероятности, которые описывают сам документ D, и оно будет равно $P(D|R, Q_k)$ в предположении независимости всех слов.

28. Что такое языковые статистические модели?

- Вероятностное распределение на множестве словарных последовательностей:
    a. p("*Мама мыла раму*") ≈ 0.001;
    b. p("*Рама мыла маму*") ≈ 0.0000000000001;
    c. p("*Матрица Грамма в унитарном пространстве эрмитова"*) ≈ 0.00001.

- Может быть использована для порождения текста, если рассматривать как случайный процесс семплирования слов из данного вероятностного распределения. Поэтому также можно встретить термин *генеративная модель языка*.

- Зависит от коллекции, тематики, типа модели.

29. Языковая модель информационного поиска

- Как вероятна каждая последовательность?
    a. P (w1, w2, w3,.. wn)
    b. P(w5| w1, w2, w3, w4)

- Языковая модель – математическая модель, которая вычисляется вероятность последовательности слов или условную вероятность следования слова в контексте

30. Вопросно-ответные системы: постановка задачи. основные компоненты, особенности тестирования.

Вопросно-ответная система (QA-система; от англ. QA — англ. Question-answering system) — информационная система, способная принимать вопросы и отвечать на них на естественном языке, другими словами, это система с естественно-языковым интерфейсом.

Этапы работы:
1. Анализ вопроса
    - определяется тип вопроса (вопрос времени, места, количества и другие) и тип ответа

- Формируется запрос в информационно-поисковую систему

2. Поиск релевантных документов или абзацев информационно-поисковой системой
   - формируется упорядоченный список наиболее релевантных документов (абзацев)
   - выбирается первых n (например, n=100-1000) документов (абзацев) для дальнейшей обработки
3. Анализ полученных документов
   - содержит ли документ требуемый тип ответа
   - Анализ предложений: близость слов ответа и вопроса, сходство синтаксических структур и т.п.
4. Извлечение ответа заданного типа
   - суммирование ответов от разных документов

31. Классификация вопросов в вопросно-ответных системах. Типы вопросов и типы ответов

Типы вопросов:

1. Фактоидные вопросы - краткий ответ
2. Нефактоидные (нарративные) вопросы - ответ включает одно или более предложений

Типы вопросов:

- Вопрос типа почему
  - по какой причине, на каком основании, с какой целью, какова причина, какова цель,
  - зачем, для чего, из-за чего, благодаря чему, и т.п. все предлоги из типа фрагм.
  - что + (вызвать/ вызывать/ обуславливать/ обусловить/ определить/ определять/ повлечь/ повлиять/ породить/ порождать/ привести/ приводить/ способствовать)
- Вопрос типа когда
  - в какое время (_ / случиться / происходить / быть / произойти /совершиться / случаться / приключиться / приключаться ; с какого времени, до какого времени / к какому сроку / к какому времени)
- Вопрос типа «где»
- Вопрос типа «как»
- Вопрос типа «кто»
- Вопрос типа «что такое»
- Вопрос типа «куда»
- Вопрос типа «сколько»
- Вопрос типа «какой параметр»
- Вопрос типа «что»

- Вопрос типа «как называется»
- Вопрос на сравнение

Типы ответов на вопрос “когда”:

- Ответ типа ФРАГМЕНТ
    - в начале
    - в конце
    - в то время как …
- Ответ типа СПИСОК
    - день
    - ночь
    - Утро…
- Ответ типа ЧИСЛО
    - Даты, точное время

Таксономия ответов:

- 6 классов верхнего уровня
- Сокращение, сущность, определение, человек, место, число
- 50 классов второго уровня
- Место: город, страна, гора
- Человек: группа, титул, персона,
- Сущность: животное, тело, цвет, валюта

32. Особенности обработки фактоидных вопросов

- Фактоидные вопросы - краткий ответ Тестирование на конференции TREC с 1999
    1. How long did the Charles Manson Murder trial last?
    2. Who is the first American in space?
    3. What does the Peugeot company manufacture?

- Нужно не только выдать правильный документ, но и выдать фрагмент, где есть ответ на вопрос или точный ответ и контекст к нему
- Формулировки запросов длинные
    - Лишние слова,
    - Более важные слова,
    - Учет синтаксической структуры

- В релевантном тексте могут использоваться синонимы или близкие по смыслу слова=> использование тезаурусов для расширения запросов

33. Особенности обработки нефактоидных вопросов.

- Нефактоидные (нарративные)вопросы
  a. Ответ включает предложение или несколько предложений**.**
  b. *Расскажите, пожалуйста, о туристических и транзитных визах в США. Что собой представляют визы, выдаваемые супругам, и визы, связанные с обучением? Сколько стоит оформление визы?*

34. Что такое PageRank? Зачем нужен, как вычисляется

Если PageRank дать точное определение то:

> *PageRank — это числовая величина, характеризующая «важность» веб-страницы. Чем больше ссылок на страницу, тем она становится «важнее». Кроме того, «вес» страницы A определяется весом ссылки, передаваемой страницей B. Таким образом, PageRank — это метод вычисления веса страницы путём подсчёта важности ссылок на неё.*

PageRank рассчитывается для каждой страницы, и если делать грамотную структуру сайта, его можно распределить равномерно или под нужные задачи по сайту.

- PR(A) = d + (1-d)(PR(T1)/C(T1) +…+PR(Tn)/C(Tn))
- **PR(Tn)** – исходная значимость страницы
- **C(Tn)** – количество исходящих ссылок со страницы
- **PR(Tn)/C(Tn)** – значимость страницы равномерно распределяется по исходящим ссылкам и переносится в значимость страницы A по входящим в нее ссылкам
- **d** – например, 0.15 значимость страницы, без входящих ссылок (коэффициент телепортации)

- **(1-d)(…)** – 0.85

35. Алгоритм HITS

Алгоритм HITS (*Hyperlink Induced Topic Search*), предложенный в 1999 году Джоном Клейнбергом, позволяет находить Интернет-страницы, соответствующие запросу пользователя, на основе информации, заложенной в гиперссылки.

Метрика HITS часто используется для ответа на широкую тему запросов и нахождения сообществ документов(*Tightly-Knit Community*), в Интернете. Идея алгоритма основана на предположении, что гиперссылки кодируют значительное количество скрытых авторитетных страниц.

Авторитетный документ (авторитетная страница, автор) — это документ, соответствующий запросу пользователя, имеющий больший удельный вес среди документов данной тематики, то есть большее число документов ссылаются на данный документ.

Хаб-документ (хаб-страница, посредник) — это документ, содержащий много ссылок на авторитетные документы.

Страница, на которую ссылаются многие другие точки должна быть хорошим «автором». В свою очередь страница, которая указывает на многие другие, должна быть хорошим «посредником». Основываясь на этом, в алгоритме HITS для каждой веб-страницы рассчитываются две оценки: оценка авторитетности и посредническая оценка. То есть для каждой страницы рекурсивно вычисляется её значимость как «автора» и «посредника».

Первым шагом в алгоритме HITS, является получение наиболее релевантных страниц в поисковом запросе. Это множество называется корневым набором и может быть получено путём принятия самых популярных страниц n, возвращаемых текстовым алгоритмом поиска. Базовый набор формируется путём увеличения корневого набора со всеми веб-страницами, которые с ним связаны и с некоторыми страницами, ссылающимися на него. Веб-страницы в базовом наборе и все гиперссылки между этими страниами, образуют сосредоточенный подграф. HITS вычисления выполняются только на этом подграфе.

Оценки авторитетного документа и посредника определены в терминах друг друга во взаимной рекурсии. Оценка авторитетности страницы вычисляется как сумма значений оценок посреднических страниц, которые указывают на эту страницу. Значение оценки посредника вычисляется как сумма оценок авторитетных страниц, на которые он указывает.

Алгоритм выполняет ряд итераций, каждая из которых состоит из двух основных этапов:

- **Обновление авторитетности**. Обновление авторитетной оценки каждой вершины подграфа, эквивалентное сумме посреднических оценок каждой из вершин, указывающих на них.

- **Хаб-обновление**. Обновление посреднической оценки каждой вершины подграфа, путём суммирования авторитетных оценок каждой из вершин, на которые они указывают.

Оценка авторитетности и посредническая оценка для вершины рассчитывается по следующему алгоритму:

- Начните с вершин, оценка авторитетности и посредническая оценка которых равна 1.
- Выполнение правила обновления авторитетности.
- Выполнение правила хаб-обновления.
- Нормализация значений путём деления каждой посреднической оценки на корень квадратный из суммы квадратов всех посреднических оценок, и деления каждой оценки авторитетности на корень квадратный из суммы квадратов всех оценок авторитетности.
- Повторение со второго шага по мере необходимости.

36. Особенности использования кликов пользователя в качестве фидбека от пользователя

Смотреть 39

37. Классификация запросов по цели. Зачем нужна. Особенности обработки разных типов запросов

Классификация запросов — определение разных видов запросов в группы по свойственным только этим группам признакам.

Поисковый запрос — последовательность символов, чаще всего представляющая из себя слово или словосочетание, описывающая какой-либо объект или явление, вводимая пользователем в поисковую строку поисковой системы с целью получить информацию об этом объекте или явлении.

Классификация запросов по целям пользователей

В зависимости от того, какую цель пользователь преследует при создании запроса, можно выделить:

- **Информационные запросы**, при которых пользователю важен лишь конечный результат в виде необходимой ему информации, и не имеет значения, где она будет найдена.

- **Навигационные запросы**, создаваемые для поиска конкретного адреса веб-ресурса с необходимой информацией.
- **Транзакционные запросы**, создаваемые для приобретения конкретных товаров, услуг и т.п.

Классификация запросов по однозначности

По тому, насколько четко и конкретно задаются формулировки запросов, их можно разделить на:

- **Четкие**, в которых объект поиска задан достаточно четко и недвусмысленно (Пример: *купить квартиру в москве*).
- **Нечеткие**, трактовка формулировки которых может иметь несколько вариантов (Пример: *квартиры*).

Классификация по длине запроса

Объект поиска информации и ее объем могут быть найдены с помощью разных по количеству слов запросов. По этому признаку запросы делятся на:

- **Однословные**.
- **Многословные** (двухсловные, трехсловные и т.д.)

По географии поиска

- **С географией**, когда объект поиска привязан к определенному месту (город, страна, регион и т.п.)
- **Без географии**. В этом случае пользователю не важно местонахождение объекта, или объект не может по определению занимать место в пространстве (например, запрос — с целью скачать программное обеспечение)
- **С географией по умолчанию**. Местоположение объекта поиска в этом случае пользователем прямо не задается, но подразумевается.

Классификация запросов по частотности

Один из наиболее важных видов классификации для вебмастеров, занимающихся продвижением сайтов. В зависимости от популярности того или иного запроса в поисковых системах, все запросы делятся на:

- **Высокочастотные** (ВЧ) — наиболее часто вводимые в поисковую строку слова, словосочетания и их морфологические формы по определенной тематике. Ввиду популярности продвижение по ВЧ является наиболее сложным, трудоемким и финансово затратным.
- **Низкочастотные** (НЧ) — соответственно, наименее часто используемые пользователями запросы определенной тематики. Являются одним из наиболее эффективных инструментов в продвижении.
- **Среднечастотные** (СЧ) — запросы, популярность которых относительно высока. Достаточно часто используются оптимизаторами при продвижении ресурсов в сети интернет.

Точных количественных характеристик, определяющих, к какому виду по частоте относится запрос, не существует. При составлении семантического ядра, алгоритмы специального программного обеспечения используют эмпирический подход в определении частотности: составляется массив всех возможных запросов по определенной тематике, среди них выбираются лидирующие с отрывом по популярности и относятся к ВЧ, наименее популярные с отрывом — к НЧ, остальные к СЧ.

Классификация запросов в зависимости от их конкурентности

Если разделение запросов по их частотности определяется на основе статистики пользовательского поведения, то конкурентность запросов — понятие, определяемое вебмастерами с помощью анализа их использования конкурентами по той же тематике.

Можно считать запросы высококонкурентными, если поисковая выдача по ним показывает ссылки на первые страницы сайтов. Но это очень относительный критерий

конкурентности. Вообще же, для оценки конкурентности целевых запросов учитываются ссылки (с целевыми запросами в анкорах) на ТОПовые страницы ресурсов конкурентов. Вес ссылочной массы на страницы конкурентов принято рассчитывать по формуле:

где Pri — PageRank или вИЦ i-й страницы, на которой находится внешняя ссылка, ni — общее число ссылок на i-й странице, N — число внешних ссылок (в которых присутствует ключевой запрос) на страницу конкурента.

Чем выше вес ссылочной массы, рассчитанной для определенных запросов, на ТОПовые страницы конкурентов, тем выше и конкурентность таких запросов.

Таким образом, различают запросы:

- **Высококонкурентные** (ВК). Продвижение по ним, соответственно, наиболее затруднительное и ресурсоемкое.
- **Низкоконкурентные** (НК).
- **Конкурентные** (СК).

Виды поисковой выдачи, формирующейся под давлением оптимизаторов

Основная статья: ***Поисковая выдача***

В некоторых случаях определенные тенденции создают условия для применения оптимизаторами одних и тех же запросов для продвижения своих ресурсов по схожей тематике. Тогда такие запросы могут влиять на поисковую выдачу.

К примеру, коммерциализированность некоторых направлений приводит к тому, что по определенным высококонкурентным запросам первые страницы выдачи содержат только продающие сайты. Информация же по объекту запроса может быть расположена далеко ниже ТОПов. Например, по запросу «ноутбуки», какую-то общую, полезную и интересную информацию можно найти не ранее, чем на 50-й странице выдачи Яндекса.

В других случаях воздействие факторов оптимизации может быть меньшим или отсутствовать. Это в большей степени относится к запросам, не содержащим объект, который может быть принят за популярный товар. К примеру, по запросу «цветы», выдача будет перемежаться коммерческими и информационно-познавательными сайтами, а по запросу «соцветия» коммерческих сайтов на первых страницах предложено не будет вовсе.

Таким образом, выдача бывает:

- **Монотонная**. Показывается по нечетким запросам, но в результате проведенной оптимизации, один вид ответов вытесняет все остальные («ноутбуки»).
- **Энциклопедическая**. Также показывается по нечетким вопросам, но представляет собой все варианты ответов, возможные по тематике запроса («цветы»).
- **Дорвейная**. Формируется в ответ на НЧ запросы на нетематических сайтах. Не является полезной и информативной.
- **Естественная**. Формируется ПС в ответ на НК запросы естественным путем, без всякого давления оптимизаторов («соцветия»).

Классификация запросов по форме их построения

- **Естественный запрос**. Строится в форме вопросительного предложения (пример: «где скачать фильмы про вампиров»).
- **Телеграфный запрос**. Строится без соблюдения структуры — связи между словами, предлогов и тому подобного (пример: «квартиры москва недорого»).
- **Запрос с операторами**. Поисковые системы предлагают более точный поиск с использованием специальных операторов (+, *, ~~ и т.д. — у разных ПС они отличаются), позволяющие уточнить запрос (пример: «Кузькина ~ мать»). Памятка с операторами запросов для ПС Яндекс.
- **Запрос-цитата** (пример: «Чтобы сварить кашу, возьмите»).

Классификация запросов по языку

- Запрос, созданный с применением единственного языка.
- С использованием нескольких языков.
- Запрос на определенном языке, но с грамматическими ошибками.
- Запрос, сформированный с помощью слов, имеющих одинаковое значение в разных языках.

38. Поисковые сессии в интернет. Как выделять, зачем.

- Совокупность запросов, которая задается пользователем в течение некоторого интервала времени.
- Граница интервала – пауза в запросах
- Типичная сессия
    - Два запроса
    - Из двух слов
    - Две страницы выдачи
    - Два клика на страницу

Методы разделения на сессии

- Разделение по времени
    - Проблема: многозадачная сессия
- Разделение по сходству (content-based)
    - Близкие по смыслу запросы могут быть не похожи по словам (vocabulary mismatch)
- Комплексный подход
    - Включает дополнительные ресурсы (например, Википедию) (Semantic-based)

39. Особенности использования кликов пользователя в качестве фидбека от пользователя. Каскадная модель при обработке кликов.

Интерпретация кликов как relevance feedback

- Клики – это хорошо…
    - Одинаково ли хороши?
- Отсутствие кликов может объясняться:
    - Не релевантно
    - Не видел

Неравноценность позиций относительно кликов

- Более высокие позиции получают больше кликов пользователя, чем более низкие позиции (eye fixation).
- Это справедливо, даже есть выдачу переставить наоборот
- “Клики информативны, но смещены (biased)”.

Гипотеза о «наблюдении»

- Документ должен быть прочитан перед кликом.
- Условная вероятность клика после прочтения зависит от релевантности документа

Гипотеза о «наблюдении»

- Вероятность клика делится на две части
  - Глобальный компонент: вероятность увидеть – зависит от позиции документа
  - Локальный компонент: зависит от пары (запрос, документ)
- Это основа любой современной модели

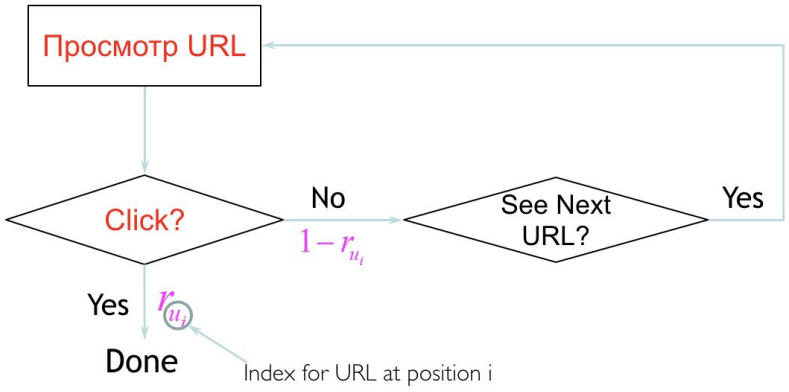
Каскадная модель

- Первый документ всегда просматривается
- Дальше модель Маркова
  - Просмотр на позиции i+1 зависит от просмотра и клика на позиции i
- Просмотр идет линейно

- Объединяем две гипотезы:

- Формальная спецификация модели:
  - $P(C_i=1|E_i=0) = 0$, $P(C_i=1|E_i=1) = r_{ui}$
  - $P(E_1=1) = 1$, $P(E_{i+1}=1|E_i=0) = 0$
  - $P(E_{i+1}=1|E_i=1, C_i=0)=1$

- Блок схема поведения пользователя:



Задачи на следующие темы:

1) Векторная модель информационного поиска

2) Языковая модель информационного поиска

3) Вычисление точности, полноты, F-меры, средней точности

Точность: $P = \frac{tp}{tp+fp}$ (доля релевантных в выдаче)

Полнота: $R = \frac{tp}{tp+fn}$ (доля выданных среди релевантных)

F-мера: $F1 = \frac{2}{1/R + 1/P} = \frac{2PR}{P+R}$

$AP = \frac{\sum_{rel} P_{rel_k}}{N_{rel}}$, где P_{rel_k} - точность в момент k, когда в выдаче релевантный документ. N_{rel} - число релевантных документов.

4) Оценка качества вопросно-ответной системы: мера MRR

5) Мера упорядочения: NDCG

$nDCG_p = \frac{DCG_p}{IDCG_p}$

$DCG_p = g_1 + \sum_{i=2}^{p} \frac{g_i}{log_i}$, g_i - значение релевантности для i-ого документа в выдаче

$IDCG_p$ - для наилучшего расположения документов по релевантности

6) Вычисление PageRank