Вероятностные модели информационного поиска

- векторная модель
- классич. модель BIM
- BM25
- языковая модель

① Основные св-ва естественного языка

- неограниченная семантическая мощность - способность к передаче инф. относительно любой области
- эволютивность - неограниченная способность к бесконечному развитию и модификациям
- проявление языка в виде речи
- этничность
- двойственность
- словарь и грамат. правила определяются практикой применения и не всегда могут быть формально зафикс.

② Что такое графематический анализ

- начальный этап обработки текста, в ходе которого определяются элементы грамматической структуры (слова, знаки пунктуации, числа, сокращения)

③ Что такое лемматизация

- процесс приведения словоформы к лемме - её нормальной (словарной форме).

④ Как работает словарный морф. анализ

Каждой словоформе сопоставлена основа или лемма
Для неслужебных слов:
- выделить возможные окончания длиной 0-3 симв.
- для каждого окончания определить код по табл. окончаний и номер флективного класса по словарю лемм
- проверить их согласованность по морф. таблице
- сохранить варианты, если согласуются

5 Как морф. анализаторы обрабатывают слова, отсутствующие в словаре

Предсказание – выявление аналогий со словоформами, распознаваемыми имеющимся словарём

Алгоритм предсказания:
1) предсказание префиксального образования
2) предсказание по концовке, взятой из известных словоформ

По префиксу:
- попытка найти словоформу макс. совпад. справа
- если префикс ≤ M символов, а п.ч. ≥ N симв, то слово разбирается по образцу известной словоформы

По концовке
- опред-ся известные концовки длины k
- сопоставление с ~~известными~~ морф. характеристик.
- заносится в исх. лексикон, если встречается не менее L раз и чаще конкурентов в пределах одной части речи

Всегда предусматривается разбор сущ.

6 Постморф. анализ. Основные методы
//
Предсинтаксический
- процедура, предназначенная для устранения морфологической омонимии слов
  - выбор правильной леммы
  - уточнение морф. характеристик

Основные методы:
- написание правил
- статистические методы (на основе размеченного корпуса)

7 Основные понятия инф. поиска

- IR – процесс поиска в большой коллекции нек. неструктурированного материала, удовлетворяющего информационные потребности.
- Релевантность – св-во документа содержать информацию, которую искал пользователь
- Оценка качества – экспериментальные процедуры и меры для сравнения результатов работы системы с ожиданиями пользователей
- Потребность пользователя, инф. потребность

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

По охвату:

1) Интернет-поиск
   - собирает рез-ты по общедоступному Интернету
   - проблема ранжирования рез-тов
   - большие объёмы
   - Интернет-реклама
   - хорошее качество
2) Корпоративный поиск
   - собирает инф. разных форматов из совокупности хранилищ
   - ранжирование док. разного типа
   - относительно малый объём
   - хуже кач-во поиска

По направленности:
- мед. поиск
- патентный поиск
- научный
- по законодательству
- ...

9 Особенности научного поиска
- рост публикаций
- устаревание
- проблема литературы до эпохи интернета
- цитаты, которые можно использовать как ссылки

10) Основные этапы обработки текстов в поисковой машине

1) Извлечение текстов
- Краулер идентифицирует и извлекает док-ты для поисковой машины.
- Хранилище документов - текст, метаданные. Быстрый доступ к содержанию

2) Преобразование текстов
- Анализатор обрабатывает послед-ть токенов, распознает структурные элем.
- Токенизатор распознает слова
- Обработка структуры HTML, XML
- Удаление наиболее частотных слов
- Стемминг (морф. анализ)
- Анализ ссылок
- Извлечение инф. (распозн. имён, мест, дат ...)
- Классификация (по тематике, тональности и т.д.)

3) Создание индекса
- Статистика по документам (частота слов и т.д.)
- Определение весов для индексных термов
- Инвертирование табл. документ-терм
- Распред. хранение

11) Основные этапы обработки запроса - " -.

1) Взаимодействие с пользователем
- Ввод запроса (интерфейс и парсер)
- Трансформация запроса (стемминг, подсказки, расширение)
- Выдача рез-тов

2) Ранжирование
- приписывание веса соответствия документа запросам
- Оптимизация выполнения запроса
- Распред. выполнение

3) Оценка качества

12) Булевская модель IR. + и -.

Два возможных рез-та TRUE и FALSE.
Поиск по полному совпадению
Запрос составляется с использованием булевских операторов и операторов близости.
Для каждого терма t хранится список документов, содержащих t. Докумен → docID
Создание инверт. индекса:
1) преобразование док. в поток токенов
2) модификация токенов
3) пары (токен, docID)
4) сортировка по токенам и затем docID

Для булевских выр: пересечение, объединение, дополнение списка ~~терм~~ док. для входящих в выр. токенов

⊕
- рез-ты предсказуемы, их легко объяснить
- могут быть встроены различные признаки
- эффективная обработка

⊖
- качество выдачи зависит от пользователя
- простые запросы дают слишком много рез-тов
- длинные запросы сложно составлять

13) Как измеряется кач-во булевского поиска

Рез-ты не ранжированы
Поисковая система разделяет коллекцию на мн-ва: выдано - не выдано
Эксперт разделяет на: релевантный - нерелевантный

Меры качества:

- точность P – доля релевантных в выданных
- полнота R – доля выданных среди релевантных
- F-мера $F1 = \frac{2PR}{R+P}$

14 Алгоритм сопоставления запроса с док-тами (Merge)

Пересечение списков docID

```
answer ← ()
while (p1 ≠ NIL) && (p2 ≠ NIL) do
    if docID(p1) = docID(p2) then
        ADD (answer, docID(p1))
        p1 ← next(p1)
        p2 ← next(p2)
    else if docID(p1) < docID(p2) then
        p1 ← next(p1)
    else
        p2 ← next(p2)
```

Если docID не равны, двигаться по списку, где docID меньше

Необходимо, чтобы записи в списке были отсортированы по docID.

Тогда будет $O(len(p1) + len(p2))$ операций.

15 Векторная модель инф. поиска

Документы и запрос представляются в виде векторов N-мерного пространства.

N – общее кол-во термов.

$v(i)$ – вес i-го терма в документе.

Для каждого док-та считается вектор:
для каждого терма – tf в этом документе.

Для каждого терма считается idf:

$$idf = \log_{10} \frac{N}{df_t}$$, где df_t – кол-во док-тов, содерж. t

Формализация близости векторов – косинусная мера (вес)
Док. ранжируются по мере снижения веса
Выдача первых K док-тов.

16 Смысл показателей idf и tf.idf

tf – сколько раз терм встречается в документе

df – в скольких документах встречается терм

$idf = \log_{10} \frac{N}{df}$ – увеличивается для редких термов

tf·idf – увеличивается при росте количества упоминания терма в документе и для редких термов.

17 Классическая процедура оценки качества инф. поиска

Мера кач-ва инф. поиска – удовлетворённость пользователя
Приближение такой оценки – релевантность
Класич. процедура оценки:

1) Составим список запросов и ограничим кол-во документов
2) Для каждой пары запрос-документ выставим

экспертную оценку релевантности

3) Ответ системы рассматривается как множ-во или последовательность оценок релевантности

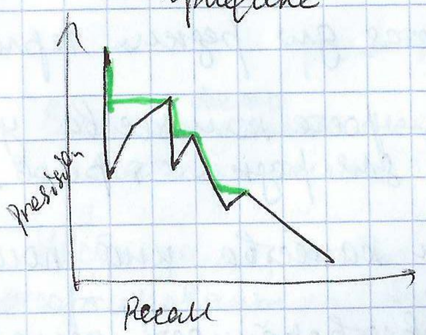
4) Построим метрики на полученных рез-тах

18 Что такое РОМИП, какие задачи решаются?

РОМИП - российский семинар по методам инф. поиска
Проведение независимой оценки методов инф. поиска.
Набор - дорожек - секций, посвящённых конкретным проектам с опред. задачей и правилами оценки (например: классиф-я, кластеризация, аннотирование текстов и т.д.)

19 Что такое кривая полнота-точность?

Кривая строится для совокупности запросов.
Изначально - набор точек, нужно интерполировать.
Идея интерполяции - если локально точность возросла с увеличением полноты, то зачем-всем её max, т.е. берём max. точности на графике



20 Что такое 11-точечный график TREC?

Рассм-ся значения полноты 0, 0.1, .., 1.0 - 11 шт,

$r_i \in \{0, 0.1, .., 1\}$, q_j - запрос. Для каждого запроса ~~вычисляется~~ и значения точности вычисляется

значение интерп. точности $p(r_i, q_j)$:

- если $r_i > recall(q_j)$, то $p(r_i, q_j) = 0$
- если $r_i \le recall(q_j)$, то $p(r_i, q_j) = \max_{n \ge pos(r_i, q_j)} (precision(n))$

где $pos(r_i, q_j)$ - min длина списка рез-тов для запроса q_j, на которой достигается r_i.

21 Что такое пулинг в IR? Сложности

Для каждого запроса:
- собрать рез-ты систем участников глубины A
- выбрать B первых рез-тов
- удалить дубликаты
- проставить оценки релевантности
- не оценённые док. считаются нерелевантными
- оценить весь ответ системы (глубины A)

Сложности:
- взаимное усиление систем
- переоценка систем, не участвовавших в оценке
- полученная оценка - оценка снизу

22 Оценка кач-ва в поисковых сист. Интернет

Особенности:
- невозможно измерить полноту
- для оценки выбираются k первых документов
- нужно учитывать при оценке, что релевантные документы должны ~~[illegible]~~ показываться раньше

23 Шкалы оценок. Мера NDCG

TREC:
раньше - бинарная
сейчас - высоко релевантный, релевантный, нерелевантный

РОМИП: соответствует, скорее соответствует, возможно соответствует, не соответствует, не м.б. оценен

NDCG: g_i - оценка i-го документа

$$DCG_\lambda = g_1 + \sum_{i=2}^{\lambda} \frac{g_i}{\log i}$$

IDCG - наилучшее упорядочение по запросу

$$NDCG = \frac{DCG}{IDCG}$$

24) Что такое информационно-поисковые тезаурусы? Зачем нужны и где применяются?

Инф.-поисковый тезаурус - нормативный словарь терминов предметной области, создаваемый для улучшения качества инф. поиска в данной предметной области.

Цели: • перевод языка автора на нормативный язык, используемый для индексации и поиска
• обеспечение последовательности в присвоении индексных терминов
• обозначение отношений между терминами
• облегчение инф. поиска.

Использование: • В настоящее время - рост интереса к предметно-ориентированному и корпоративному поиску
• индексирование и расшир. запроса

25) Что такое рубрикаторы. Чем отличаются от инф.-поисковых тезаурусов?

Рубрикатор - таблица иерархической классификации, содержащая полный перечень включённый в систему классов и предназначенная для систематизации инф. фондов, массивов и изданий и поиска в них.

Термины тезауруса явл-ся фундаментально языковыми, в то время как рубрики соответствуют концептуальным категориям.

26) Методы расширения запросов пользователей при IR.

- глобальные
  - ручные тезаурусы
  - автоматич. поддерж. тезаурус
- локальные
  - обратная связь по
    - релевантност
    - нерелевантно

27) Что означает термин relevance feedback? Осн. принципы

Пользователь оценивает документы в поисковой выдаче:
- задаёт простой запрос
- размечает часть результатов как релевантные/нерелевантные
- система улучшает соответствие документов запросу на основе разметки пользователя
- процедура м.б. итеративной

Основная идея: сформулировать хороший запрос трудно, если пользователь не знаком с коллекцией, поэтому запрос строится итеративно

28) Алгоритм Рокио для relevance feedback

Документы - точки в многомерном пр-ве
Центроид - ЦМ этих точек: $\mu(C) = \frac{1}{|C|} \sum_{d_j \in C} d_j$, C - мн-во док.

C_r - релевантные, C_{nr} - нерелевантные
Вместо идеал запроса $q_{opt} = \arg\max(\cos(q, \mu(C_r)) - \cos(q, \mu(C_{nr}))$

$$q_{opt} = \frac{1}{|C_r|} \sum_{d_j \in C_r} d_j - \frac{1}{|C_{nr}|} \sum_{d_j \notin C_r} d_j \quad q$$

На практике: $q_m = \alpha q_0 + \beta \cdot \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j$

D_r и D_{nr} - мн-во известных рел. и нерел. векторов.
α, β, γ - веса

29 Проблемы расширения запроса при помощи обратной связи по релевантности

- длинные ~~запросы~~ запросы неэффективны:
  - большое ожидание для польз.
  - высокая стоимость для системы
- пользователи часто не хотят размечать документы
- трудно понять, почему данный документ был выдан после relevance feedback

30 Основные методы автоматической рубрикации текстов

- ручное рубрицирование
- автоматическое
  - инженерный подход
  - методы машинного обучения
- полуавтоматическое

31 Что такое инженерный метод классиф. текстов? +и −

Рубрикатор создаётся осмысленно, содержание рубрики можно выразить ограниченным кол-вом понятий в виде формулы

Эксперты описывают смысл рубрики в виде булевских выражений, правил продукции

⊕ • правила обладают хорошими возможностями для масштабирования

• эксперты могут создать набор правил, превосходящий автоматически генерир. классификаторы

⊖ • сложное создание и сопровождение правил

• трудно найти хороших экспертов

32 + и − ручного рубрицирования

⊕ высокая точность

⊖ низкая скорость

33 Метод Байеса для автоматической классиф. текстов

C - набор классов, d - документ

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}, \quad c \in C$$

$$d = (x_1, .., x_n)$$

$$c = \arg\max_{c_j \in C} P(c_j | x_1, .., x_n) = \arg\max_{c_j \in C} \frac{P(x_1, .., x_n | c_j) \cdot P(c_j)}{\underbrace{P(x_1, .., x_n)}_{\text{одинак. для классов}}}$$

$$= \arg\max_{c_j \in C} P(x_1, .., x_n | c_j) \cdot P(c_j)$$

34 Метод Роккио — " —

D_c - кол-во документов ∈ классу c.

v(d) - вектор tf·idf для документа d.

Центроид: $\mu(c) = \frac{1}{|D_c|} \sum_{d \in D_c} v(d)$ вычисл-ся при обучении

Классификация основывается на близости по соз. мере

35 Метод Knn — " —

Чтобы классифицировать документ d в класс c нужно:

- определить k-окружение из k ближайших соседей d
- подсчитать число документов в N, кот. относятся к c.
- оценить P(c|d) как i/k
- выбрать класс $\arg\max_{c_j} P(c_j|d)$

36 Основной [redacted] принцип метода SVM — »—.

Метод опорных векторов:

- точки имеют вид $\{(x_1, c_1), .., (x_n, c_n)\}$, где $c_i = 1$ или -1 в завис-ти от принадлежности к классу; x_i – p-мерный вещ. вектор.
- разделяющая гиперплоскость имеет вид $wx - b = 0$, w – $1-p$ к разделяющей гиперплоскости
- нужны 2 || гиперплоскости, они м.б. описаны равенствами $\begin{cases} wx - b = 1 \\ wx - b = -1 \end{cases}$
- нужно максимизировать расстояние между плоскостями, $d = \frac{2}{|w|} \Rightarrow |w| \to min$, при условиях:
$\begin{cases} wx - b >> 1, & c_i = 1 \\ wx - b << -1, & c_i = -1 \end{cases}$

Результирующая ф-ла:

$f(x) = \sum \alpha_i c_i \langle x_i, x \rangle + b$

x_i – опорные вектора

x – вектор текущего документа

α_i – вес опорных векторов

$c_i \in \{-1, 1\}$

37 + и – методов машинного обучения для рубрикации текстов

⊕ • эффективно при наличии качественно размеченной обучающей коллекции

⊖ • низкая эффективность при большом числе рубрик
• трудно интерпретируемые результаты

38 Особенности применения методов машинного обучения для рубрикации текстов в завис-ти от размера обучающей коллекции
Кол-во записей

- нет данных для обучения – правила пишутся вручную
- мало данных – классификация "с учителем" + нужно получить размеченные записи по возможности быстрее
- достаточное количество данных – подходит для применения SVM, можно использовать булевские модели
- большое кол-во данных – хорошо подходит в теории, но SVM и kNN непрактичны; можно использовать Naïve Bayes

[redacted]

39 Что такое кластеризация текстов? Чем отличается от классификации?

Кластеризация – разделение совокупности документов на однородные внутри, но чётко отличающиеся друг от друга подмножества (кластеры)
Цель классификации (разновидность обучения с учителем) – разделить данные по категориям, установленным экспертом
Кластеризация – без учителя, разделение на основе метрики

40 Метод k-means для кластеризации текстов.

Документы - вектора над $\mathbb{R}$
Кластеры базируются на понятии центроида точек в кластере c: $\mu(c) = \frac{1}{|c|} \sum_{x \in c} x$

Присваивание документов к кластеру базируется на сх-ве с текущими центроидами

Алгоритм:
1) выберем K случайных векторов $\{s_1, .., s_k\}$ как исходное мн-во (seeds) - центроиды
2) до выполнения критерия остановки:
   для каждого d_i
   $c_i = \max_{s_j \in seeds} similarity(x_i, s_j)$.
3) обновляем мн-во $\{s_1, .., s_k\}$, заменяя на центроиды текущих кластеров:
   $\forall c_j$
   $s_j = \mu(c_j)$

Условия остановки:
- опред. число итераций
- не меняется разделение документов
- не меняются позиции центроидов

41 Агломеративная кластеризация - основной принцип и подходы

- Начинает с рассмотрения документов как отдельных кластеров
- Итеративно объединяет ближайшую пару кластеров, пока не останется один
- История объединения образует иерархию.

Способы определения, что такое наиболее сходная пара:
- single-link - сходство по наиболее похожим документам
- complete-link - по наиболее непохожим
- центроид - по наиболее похожим центроидам
- average-link - средний сос между парами элем. двух кластеров

? 42 Методы тестирования автоматической кластеризации

Критерии
- внутренний:
  - внутри кластера высокое сх-во
  - между классами - низкое
- внешние:
  - измерение способности кластеризации обнаруживать скрытые классы объектов в эталонных данных

Внешняя оценка кач-ва - чистота. Нужны размеченные данные: документы с C правильными кластерами. Тестируемая система порождает $w_1, .., w_k$ с n_i элем.

$Purity(w_i) = \frac{1}{n_i} \max_j (n_{ij})$, $j \in C$

Rand-index: $\frac{A+D}{A+B+C+D}$

43 Особенности кластеризации потока новостей в реальном времени
- корпус документов постоянно пополняется
- временное окно
- разные размеры

- наличие дубликатов ⇒ определение первоисточника
- ошибки при сборе новостных сообщений
- спамерские технологии источников

44 Анализ тональности как задача классификации: методы, признаки, меры качества

Методы:
- агрегация оценочных выражений + правила их комбинирования
- машинное обучение

Признаки для классификации машинным обучением:
- слова с весами (унитермы)
- произвольные биграммы
- знаки препинания
- смайлики

Оценка кач-ва:
- точность – полнота – F-мера
- асимметрия (правильность) – отношение правильно классиф. ко всем остальным

45 Вероятностная модель IR; осн. идея и отличия от векторной модели

Дан запрос q и идеальное мн-во R релевантных документов. Задача – определить св-ва этого мн-ва.
Вероятностная модель пытается оценить вероятность, что пользователь оценит документ d_j как релев. посредством отношения:

$$\frac{P(d_j \text{ relevant to } q)}{P(d_j \text{ nonrelevant to } q)}$$

Различия от векторной модели: сходство между запросом и документом считается по ф-ле из теории вероятности, а не как cos. мера.

$$sim(d_j, q) = \frac{P(R|d_j)}{P(\bar{R}|d_j)} = \frac{P(d_j|R)\cdot P(R)}{P(d_j|\bar{R})\cdot P(\bar{R})} \approx$$

$$\approx \log \frac{P(d_j|R)}{P(d_j|\bar{R})} + \log \frac{P(R)}{P(\bar{R})}$$

∃ k_i - слово, $q = (d_1, .., d_t)$

$$P(d_j|R) = \prod_{i=1}^{t} P(k_i = d_i|R)$$

$$P(d_j|\bar{R}) = \prod_{i=1}^{t} P(k_i = d_i|\bar{R})$$

$$\Rightarrow sim(d_j, q) \approx \frac{\prod_{g_i(d_j)=1} P(k_i|R) \cdot \prod_{g_i(d_j)=0} P(\bar{k}_i|R)}{\prod_{g_i(d_j)=1} P(k_i|\bar{R}) \cdot \prod_{g_i(d_j)=0} P(\bar{k}_i|\bar{R})} \approx$$

$$\approx \sum_{i=1}^{t} \left(\log \frac{P(k_i|R)}{1-P(k_i|R)} + \log \frac{1-P(k_i|\bar{R})}{P(k_i|\bar{R})} \right).$$

46 Что такое языковые статистические модели?

Языковая модель – матем. модель которая вычисляет вероятность последовательности слов или усл. вероятность следования слова в контексте.
Стат. модель порождения текста определяет вер-ть строк для данного языка

$$P(w_n|w_1, .., w_{n-1}) = \frac{P(w_1, .., w_n)}{P(w_1, .., w_{n-1})}$$

(47) Языковая модель IR

Каждый запрос трактуется как случайный процесс.

Подход:

- Высчитываем языковые модели для каждого документа
- Выводим вероятность порождения запроса на основе модели каждого док.
- ранжируем док. по вероятностям

Ранжирующая ф-ла: $P(Q, d) = p(d)\,p(Q|d) = p(d)\,p(Q|M_d)$

$$p(Q|M_d) = \prod_{t \in Q} \frac{tf_{t,d}}{dl_d}$$

M_d - яз. модель док. d
dl_d - число слов в d

(48) Вопросно-ответные системы: осн. компоненты, тестирование

Этапы работы:

1) анализ вопроса (опред. типа вопроса и формирование запроса к поисковой системе)
2) поиск релев. док. или абзацев (упор. список, n первых)
3) анализ полученных док. (содержит ли тип ответа)
4) извлечение ответа заданного типа