

Некоторые вероятностные модели в информационном поиске

Антон Алексеев
anton.m.alexeyev@math.spbu.ru

Computer Science Center

21 марта 2013 г.

Дано: документы D , запрос $q \in Q$
Понять: насколько релевантен $d \in D$ запросу q
И выдать самые (все) релевантные.

Мы уже знакомы с

- ггер
- булевым поиском
- векторным поиском

В программе:

- 1 Необходимые понятия из теории вероятностей
- 2 PRP + BIM + relevance feedback
- 3 Прочее

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = P(A) \left(\frac{P(B|A)}{\sum_{X \in \{A, \bar{A}\}} P(B|X)P(X)} \right)$$

Философия:

$P(A)$ — предполагаемая оценка вероятности события A

$P(A|B)$ — уточнённая свидетельством B оценка $P(A)$

Он не получка, не аванс

Отношение шансов (odds):

$$O(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)}$$

Упражнение

Свойство, которое нам пригодится. Легко проверить, что

$$P(A), P(B) \in (0, 1) \Rightarrow (P(A) > P(B) \Leftrightarrow O(A) > O(B))$$



Гипотеза: волк режет овец!

false positive = ложная тревога: овцы целы = ошибка I рода

false negative = не опознали волка: овцы съедены = ошибка II рода

Запрос q , коллекция D .

Для каждого q есть своё множество релевантных документов $Rel_q \subset D$, то есть можно ввести функцию

$$\tilde{R}_q(d) = \begin{cases} 1 & \text{if } d \in Rel_q \\ 0 & \text{otherwise} \end{cases}$$

Но, увы, мы ничего не знаем :(

Вместо \tilde{R}_q — случайная величина R_q .

Probability Ranking Principle (van Rijsbergen, 1979)

План:

- упорядоченный список документов — выводим top_k из списка документов, отсортированного по убыванию $P(R = 1|d, q)$ или
- множество всех релевантных — выводим всякий $d : P(R = 1|d, q) > P(R = 0|d, q)$.

Функция потерь штрафует одним очком как за нерелевантный документ, так и за непоявление релевантного документа в выдаче.

Теорема (Riply, 1996)

Правило PRP минимизирует байесовский риск.

C_0 — стоимость неизвлечения релевантного документа (FN)

C_1 — стоимость извлечения нерелевантного документа (FP)

Тогда d попадает в выдачу, если

$$C_0 \cdot P(R = 0|d) - C_1 \cdot P(R = 1|d) \leq C_0 \cdot P(R = 0|d') - C_1 \cdot P(R = 1|d'),$$

где d и d' ещё не показаны.

Binary Independence Model (1)

Документ — вектор весов термов.

$$w_{d,t} \in \{0, 1\}$$

Допущение 1: Naive Bayes assumption

Термы встречаются в документе независимо друг от друга.

Допущение 2

Релевантность документа не зависит от релевантности никакого другого документа.

(Умножение — мать сомнения!)

Binary Independence Model (2)

\vec{q} и \vec{x} — бинарные векторы, соответствующие запросу и документу.
Хотим уметь вычислять $P(R = 1|\vec{x}, \vec{q})$ —?

$$P(R = 1|\vec{x}, \vec{q}) = \frac{P(\vec{x}|R = 1, \vec{q})P(R = 1, \vec{q})}{P(\vec{x}, \vec{q})} = \frac{P(\vec{x}|R = 1, \vec{q})P(R = 1|\vec{q})}{P(\vec{x}|\vec{q})}$$

Аналогично для $P(R = 0|\vec{x}, \vec{q})$. Интерпретация?

Хитрые оценки, не забывать про определение вероятности.

Binary Independence Model (3)

Условные вероятности — тоже вероятности

$$\begin{aligned} O(R|\vec{x}, \vec{q}) &= \frac{P(R = 1|\vec{x}, \vec{q})}{P(R = 0|\vec{x}, \vec{q})} = \frac{P(R = 1|\vec{q})P(\vec{x}|R = 1, \vec{q})}{P(R = 0|\vec{q})P(\vec{x}|R = 0, \vec{q})} \\ &= O(R|\vec{q}) \frac{P(\vec{x}|R = 1, \vec{q})}{P(\vec{x}|R = 0, \vec{q})} \end{aligned}$$

Убили $P(\vec{x}|\vec{q})!$

Применяем Допущение 1:

$$\frac{P(\vec{x}|R = 1, \vec{q})}{P(\vec{x}|R = 0, \vec{q})} = \prod_t \frac{P(x_t|R = 1, \vec{q})}{P(x_t|R = 0, \vec{q})}$$

Binary Independence Model (4)

Теперь

$$p_t = P(x_t = 1 | R = 1, \vec{q}), \quad u_t = P(x_t = 1 | R = 0, \vec{q})$$

Допущение 3

Если $q_t = 0$ (т.е. если соответствующий терм не появляется в запросе), то присутствие термина t в релевантном и нерелевантном документах равновероятны, то есть $p_t = u_t$.

В формуле кое-что сократилось, а кое-что можно и выбросить:

$$\begin{aligned} O(R|\vec{x}, \vec{q}) &= O(R|\vec{q}) \prod_{x_t=q_t=1} \frac{p_t}{u_t} \prod_{x_t=0, q_t=1} \frac{1-p_t}{1-u_t} = \\ &= O(R|\vec{q}) \prod_{x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \prod_{q_t=1} \frac{1-p_t}{1-u_t} \end{aligned}$$

Логарифмируем то, что осталось:

$$RSV_d = \log \prod_{x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{x_t=q_t=1} \left(\log \frac{p_t}{1-p_t} + \log \frac{1-u_t}{u_t} \right)$$

Нас интересуют только термы, встречающиеся и в документе, и в запросе.

Величина $c_t = \log \frac{p_t}{1-p_t} + \log \frac{1-u_t}{u_t}$ — основа вычислений в модели.

Напоминание

$$p_t = P(x_t = 1 | R = 1, \vec{q})$$

$$u_t = P(x_t = 1 | R = 0, \vec{q})$$

Ранжируем по RSV . Как оценить?

Пример

Для фиксированного термина t :

N — число документов

S — число релевантных документов

$s \subset S$ — число релевантных документов, содержащих t

Тогда $p_t = \frac{s}{S}$ и $u_t = \frac{df_t - s}{N - S}$

$$c_t = \log \frac{s}{S - s} \cdot \frac{N - S - df_t + s}{df_t - s}$$

Деление на ноль — это смерть!

$$c_t = \log \frac{s + 0.5}{S - s + 0.5} \cdot \frac{N - S - df_t + s + 0.5}{df_t - s + 0.5}$$

Откуда берутся s и S или p_t и u_t ?

Нерелевантных много больше, чем релевантных!

$$\Rightarrow u_t = \frac{df_t}{N}$$

Тогда

$$\log((1 - u_t)/u_t) = \log((N - df_t)/df_t) \rightarrow \log(N/df_t)$$

На что похоже?

Что делать с p_t ? Варианты

- $p_t - const$ (Croft & Harper (1979))
- Магия: $p_t = \frac{1}{3} + \frac{2}{3} \frac{df_t}{N}$ (Greiff (1998))

BIM: Probabilistic relevance feedback

В идеальном мире с идеальными пользователями...

- 1 Задать начальные значения p_t и u_t . Да хоть константы!
- 2 Показать пользователю $\{d : R_{d,q} = 1\}$
- 3 Пользователь: «хочу/не хочу»: $V = VR \cup VNR$
- 4 Пересчитываем!

Как?

$$p_t = \frac{|VR_t|}{|VR|}$$

Нет, не так.

$$p_t = \frac{|VR_t| + \frac{1}{2}}{|VR| + 1}$$

А лучше — так:

$$p_t^{(k+1)} = \frac{|VR_t| + Kp_t^{(k)}}{|VR| + K}$$

Не все знают, чего хотят, и не все любят общаться с поисковиком.

- 1 Задать начальные значения p_t и u_t . Так же.
- 2 Показать пользователю V ($top|V|$ по рангу)
- 3 Пересчитываем! Например,

$$p_t = \frac{|V_t| + \frac{1}{2}}{|V| + 1}$$

$$u_t = \frac{df_t - |V_t| + \frac{1}{2}}{N - |V| + 1}$$

- 4 Если ранжирование не устаканилось, переходим к шагу 2.

Если Вы думаете, что ДА ЭТО ЖЕ TF-IDF, Вам показалось (наиболее любопытным считать проверку этого упражнением).

Хорошо?

1. Строгая мат.модель!
2. Лучше булевой

Плохо?

1. Холодный старт
2. Сильные предположения
3. Никак не учитываются длина документа и частоты термов

- Некогда очень популярная в промышленности схема
- Идея: будем учитывать длины документов и частоты термов!
- Формулы RSV , учитывающие эти характеристики:



$$idf * something(length, tfs)$$

За введением — в IIR [1] или Wikipedia.

За подробностями — статьи.

- Очень, очень сильные предположения.
Может, делать не так топорно?
- Термы как множество случайных элементов с редкими зависимостями
- Естественное представление зависимостей — орграф
(а лучше дерево)
- Байесовские сети доверия (понятие о применении даётся в [2]):
сеть документов, сеть запросов. Переменные: документы, термы, «понятия». Логико-вероятностный вывод.
(1991 год. Хорошие результаты на TREC. Продавалось.)

Спасибо за внимание!

-  Introduction to Information Retrieval (2008) by Christopher D. Manning, Prabhakar Raghavan, Hinrich Schtze.
-  Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. Modern Information Retrieval. Boston, MA, USA. Addison-Wesley Longman Publishing Co., Inc. 1999.

Прочие ссылки в скобках — найдутся в [1]

Некоторые вероятностные модели в информационном поиске

Антон Алексеев
anton.m.alexeyev@math.spbu.ru

Computer Science Center

21 марта 2013 г.