Ранжирование (часть 1)

Свербягин Никита

Факультет компьютерных наук, ПМИ
НИУ ВШЭ

18 января 2019

Оглавление

Постановка задачи

Постановка задачи

Дано:

X - множество объектов

$X_l = \{x_1, \ldots, x_l\}$ - обучающая выборка

Задан порядок на парах $(x_i, x_j) \in X_l^2$

Найти:

Ранжирующую модель $a : X \to \mathbb{R}$, такую что

$$x_i < x_j \Rightarrow a(x_i) < a(x_j)$$

Пример: линейная модель ранжирования

$$a(x \mid w) = \langle x, w \rangle$$

Примеры

Поисковая выдача

D - коллекция текстовых документов

Q - множество запросов

$D_q \subseteq D$ - множество документов, найденных по запросу q

$X = Q \times D$ - объекты - это пары «запрос, документ»

Y - упорядоченное множество рейтингов

$y : X \to Y$ - асессорские оценки

Чем выше оценка, тем релевантнее документ.

Правильный порядок определен среди документов по одному запросу.

$$(q, d) < (q, d') \Leftrightarrow y(q, d) < y(q, d')$$

Рекомендательная система

U - пользователи

I - предметы (товары, фильмы, книги и т.п.)

$X = U \times I$ - объекты - это пары

«пользователь, предмет»

Правильный порядок определен среди предметов, относящихся к одному пользователю:

$$(u, i) < (u, i') \Leftrightarrow y(u, i) < y(u, i')$$

В роли признаков объекта (u, i) могут выступать $y(u', i)$ - рейтинги, поставленные предмету другими пользователями.

Признаки для ранжирования поисковой выдачи

Типы признаков

Признаки могут являться функцией:

- только документа d
- только запроса q
- запроса и документа (q, d)

Признаки можно разделить на:

- текстовые
  - кол-во вхождений слов из q в d
  - слова из q есть в заголовках или выделены в d
- ссылочные
  - кол-во ссылок на документ d
  - полезность ссылок, содержащихся в d
- кликовые
  - кол-во кликов на d
  - кол-во кликов на d по запросу q

TF-IDF

n_{dw} (term frequency) - число вхождений слова w в текст d

N_w (document frequency) - число документов, содержащих w

N - число документов в коллекции. $N = |D|$

N_w/N - оценка вероятности встретить слово w в документе

$(N_w/N)^{n_{dw}}$ - оценка вероятности встретить его n_{dw} раз

$P = \prod_{w \in q} (N_w/N)^{n_{dw}}$ - оценка вероятности встретить в документе d слова запроса $q = \{w_1, \ldots, w_k\}$ случайным образом.

TF-IDF

Оценка релевантности документа d запросу q:

$$-logP = \sum_{w \in q} \underbrace{n_{dw}}_{TF(w,d)} \underbrace{log(N/N_w)}_{IDF(w)}$$

$TF(w,d) = n_{dw}$ - term frequency

$IDF(w) = log(N/N_w)$ - inverted document frequency

PageRank

Важность документа d определяется:

- кол-вом документов c, ссылающихся на d
- важностью документов c, ссылающихся на d
- кол-вом других ссылок в документах c

Вероятность попасть на страницу d, если кликать случайно:

$$PR(d) = \frac{1-\delta}{N} + \delta \sum_{c \in D_d^{in}} \frac{PR(c)}{|D_c^{out}|}$$

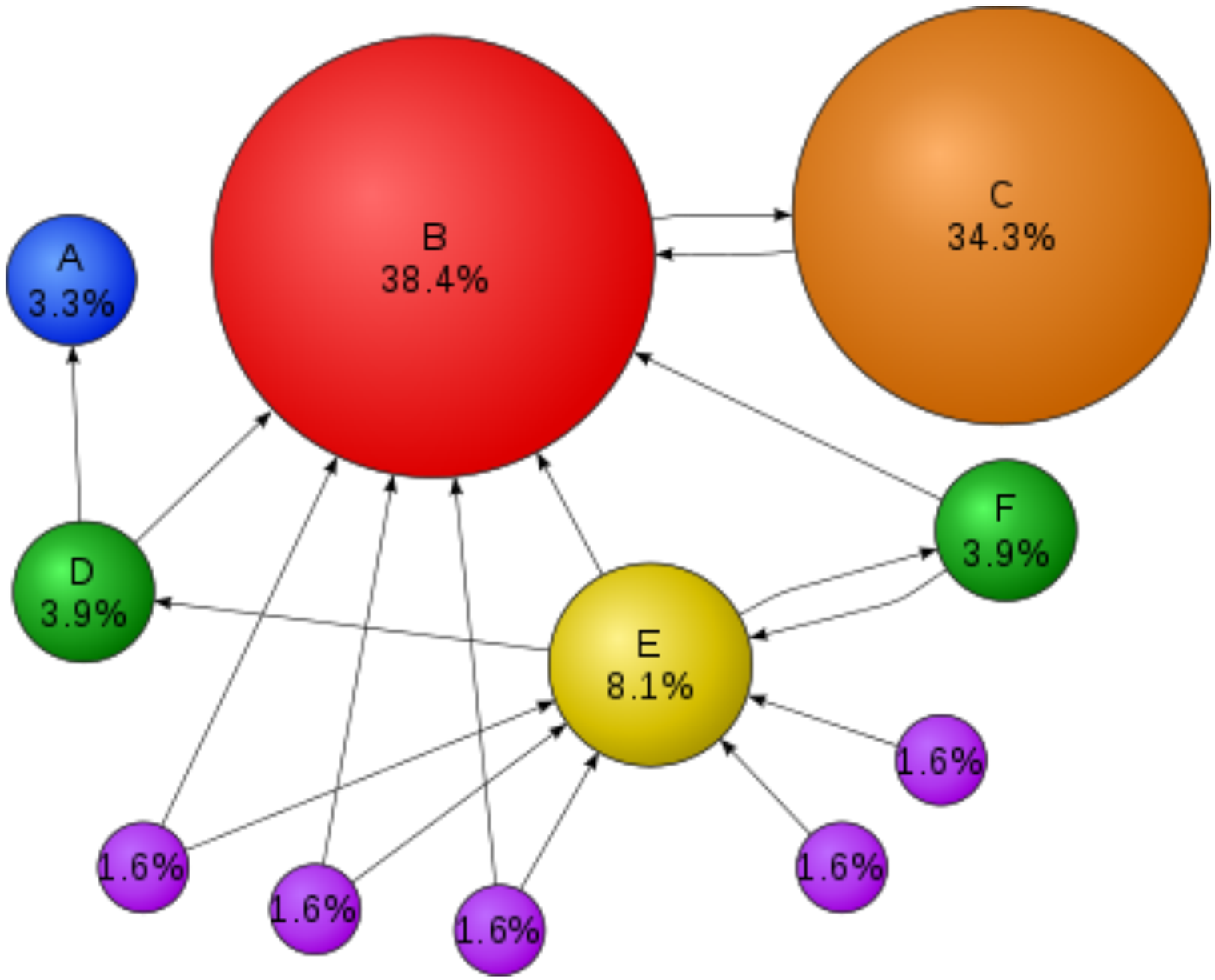
$D_d^{in} \subset D$ - мн-во документов, ссылающихся на d

$D_c^{out} \subset D$ - мн-во документов, на которые ссылается c

$\delta = 0.85$ - вероятность продолжать клики (damping factor)

N - кол-во документов в коллекции.

PageRank



Еще несколько примеров

- Факторы домена: возраст домена; срок регистрации; доменная история; подозрительный владелец; наличие национального домена своей страны (.ru) …

- Факторы страницы: ключевое слово в теге «title», «description» или «H1»; ключевое слово часто встречается в контенте; длина контента; скорость загрузки страницы; давность и частота обновления контента; авторитетность хостинга; ключевое слово в url; приоритет страницы на карте сайта…

- Факторы сайта: уникальность содержимого; кол-во контактной информации; кол-во страниц; наличие карты сайта; uptime;

Еще несколько примеров

- Специальные правила алгоритмов:
  - QDF (query deserves freshness)
  - QDD (query deserves diversity) (для запросов с различной интерпретацией)
  - История посещенных сайтов
  - История поисковых запросов
  - Таргетинг по местоположению
- Социальные сигналы: кол-во лайков, репостов и т.п. постов в социальных сетях

Метрики качества

MAP

Пусть $\mathbb{Y} = \{0,1\}$, $y(q,d)$ - релевантность документа. $d_q^{(i)}$ - i-й документ в отсортированном с помощью ранжирующей модели $a(q,d)$ списке документов по убыванию.

precision at K:

$$p@K(q) = \frac{1}{K}\sum_{i=1}^{K} y(q, d_q^{(i)})$$

Недостаток данной метрики: не учитывается порядок документов среди первых K.

MAP

Данную проблему нивелирует метрика **average precision at K**: сумма $p@i$ только для релевантных документов среди первых і документов.

$$ap@K(q) = \frac{1}{K}\sum_{i=1}^{K} y(q, d_q^{(i)}) \cdot p@i(q)$$

$p@K$ и $ap@K$ считаются для конкретного запроса.

mean average precision at K:

$$map@K = \frac{1}{|Q|}\sum_{q \in Q} ap@K(q)$$

Доля «дефектных» пар

Пусть $\mathbb{Y} = \mathbb{R}$, остальное аналогично.

Доля инверсий среди первых K документов:

$$DP@K(q) = \frac{2}{K(K-1)} \sum_{i<j}^{K} \left[y(q, d_q^{(i)}) < y(q, d_q^{(j)}) \right]$$

Заметим, что эта метрика тесно связана с AUC-ROC в задачах бинарной классификации:

$$AUC@K(q) = \frac{1}{l\_l_+} \sum_{i,j=1}^{K} [y_i > y_j][a(x_i) < a(x_j)] = \frac{K(K-1)}{2l\_l_+} DP@K(q)$$

nDCG

Теперь $\mathbb{Y} = \mathbb{R}$. Остальное аналогично.

Cumulative gain at K:

$$CG@K(q) = \sum_{i=1}^{K} y(q, d_q^{(i)})$$

Недостатки: данная метрика не нормализована и не учитывает позицию релевантных документов.

Discounted cumulative gain at K:

$$DCG@K(q) = \sum_{i=1}^{K} \frac{2^{y(q,d_q^{(i)})} - 1}{log_2(i+1)}$$

nDCG

метрика DCG@K решает проблему учета позиций, но остается ненормированной.

normalized DCG at K:

$$nDCG@K(q) = \frac{DCG@K(q)}{IDCG@K(q)}$$

IDCG@K - ideal DCG@K, отличие от обычного DCG в том, что документы отсортированы по $y(q, d)$

По аналогии с map@K можно посчитать nDCG@K, усредненный по всем запросам.

pFound

Пусть $\mathbb{Y} = [0,1]$, $y(q,d)$ - оценка вероятности найти ответ на запрос q в документе d.

Оценка вероятности найти ответ в первых K документах:

$$pFound@K(q) = \sum_{i=1}^{K} P_i \cdot y(q, d_q^{(i)}),$$

где P_i - вероятность дойти до i-ого документа:

$$P_1 = 1,$$

$$P_i = P_{i-1} \cdot (1 - y(q, d_q^{(i-1)})) \cdot (1 - P_{out}),$$

P_{out} - вероятность прекратить поиск без ответа

pFound

Стандартные значения при использовании pFound:

$$P_{out} = 0.15$$

Оценка асессора	y(q, d)
Vital	0.61
Useful	0.41
Relevant+	0.14
Relevant-	0.07
Not Relevant	0.00

Источники

- видеолекция «Методы обучения ранжированию», К.В Воронцов, URL
- Статья «Метрики качества ранжирования» на habr.com, URL
- Курс «Прикладные задачи анализа данных», урок «Задача ранжирования» на coursera.org, URL