Теория информационного поиска

Лекция 8. Оценка качества поиска

Дмитрий Грановский

СПбГУ

05.05.2025

Релевантность

- задает ранжирование
- безразмерная величина, важно только сравнивать документы между собой
- считается для пары (запрос, документ)
  - точнее: для тройки (запрос, документ, **пользовательский контекст**)
- бывает разная:
  - бинарная и произвольная
  - пользовательская и вычисленная

Общий подход

- оценка = сравнение с идеальной системой
- идеальная система:
  - показывает только релевантные документы
  - самые релевантные ранжирует выше
  - может иметь дополнительные требования: свежесть, разнообразие и т.д.
  - не существует :-(
- оценка — численная, состоит в расчете **метрики**

Общий подход конкретнее

- оцениваем качество по корзине запросов и усредняем
  - составить хорошую корзину — отдельная задача
- как оценить качество по одному запросу:
  - попросим экспертов (*асессоров*) оценить какое-то множество документов на релевантность запросу
    - желательно много асессоров и > 1 асессора на документ
  - эти оценки используем для расчета метрик (подробности далее)

Оценка качества булева (двоичного) поиска

- далее везде считаем, что смотрим на один запрос q
- $Rel(D_i, q) \in \{0, 1\}$
- рассматриваем ситуацию:
  - асессоры: релевантны документы $d_{i_1} \dots d_{i_n}$
  - система: нашла документы $d_{j_1} \dots d_{j_k}$
  - ранжирования нет
- возможные состояния документа:
  - релевантный, нашелся
  - нерелеватный, нашелся
  - релевантный, не нашелся
  - нерелевантный, не нашелся (их большинство)

Оценка

	Rel	**NRel**
нашли	true positive (tp)	false positive (fp)
не нашли	false negative (fn)	true negative (tn)
Σ	R	$N - R$

Оценка

	Rel	**NRel**
нашли	true positive (tp)	false positive (fp)
не нашли	false negative (fn)	true negative (tn)
Σ	R	$N - R$

- мы не можем оценить всю коллекцию
- поэтому обычно оцениваем верх фактической выдачи
- количество **tn** нам безразлично

Типы ошибок

	Rel	**NRel**
нашли	true positive (tp)	false positive (fp)
не нашли	false negative (fn)	true negative (tn)

Типы ошибок

	Rel	**NRel**
нашли	true positive (tp)	false positive (fp)
не нашли	false negative (fn)	true negative (tn)

- булев поиск можно рассмотреть как задачу *классификации* на 2 класса (по релевантности)
- false positive = ошибка I рода
- false negative = ошибка II рода
- таблица называется матрицей несоответствий/ошибок (*confusion matrix*)

Метрики классификации

	Rel	**NRel**
нашли	true positive (tp)	false positive (fp)
не нашли	false negative (fn)	true negative (tn)

Метрики классификации

	Rel	**NRel**
нашли	true positive (tp)	false positive (fp)
не нашли	false negative (fn)	true negative (tn)

Две фундаментальные метрики:

- точность (*precision*): $P = \frac{tp}{tp+fp}$
- полнота (*recall*): $R = \frac{tp}{tp+fn}$
- ещё бывает *accuracy*: $Acc = \frac{tp+tn}{tp+tn+fp+fn}$

Точность и полнота

- идеальная классификация недостижима
- почти всегда можем сдвигать качество либо к точности с потерей полноты, либо наоборот (tradeoff)
- в некоторых задачах важнее точность, в некоторых полнота
- если мы понимаем, в какой пропорции важны точность и полнота, можно использовать F*-меру* (*f-score*)
- в частности, F_1-меру: $F_1(P, R) = \frac{2 \times P \times R}{P + R}$
- пример: $F_1(0.3, 0.1) = \frac{2 \times 0.3 \times 0.1}{0.3 + 0.1} = 0.15$

Оценка качества поиска с ранжированием

- $Rel(d_i, q) \in [0, 1]$
- рассматриваем ситуацию:
  - асессоры: релевантность документа $d_i = Rel_{d_i}$
  - система: нашла документы $d_{j_1} \ldots d_{j_k}$ в этом порядке
- набор дискретных оценок релевантности, например:
  - Vital (в статье $Rel = 0.4$)
  - Useful
  - Rel+
  - Rel-
  - Stupid

Precision @k

k	Rel?	Всего Rel	P@k
1	R	1	$\frac{1}{1} = 1.0$
2	NR	1	$\frac{1}{2} = 0.5$
3	NR	1	$\frac{1}{3} \approx 0.33$
4	R	2	$\frac{2}{4} = 0.5$

Precision @k, пример #2

k	Rel?	Всего Rel	P@k
1	R	1	$\frac{1}{1} = 1.0$
2	R	2	$\frac{2}{2} = 1.0$
3	NR	2	$\frac{2}{3} \approx 0.67$
4	NR	2	$\frac{2}{4} = 0.5$

Precision @k, пример #2

k	Rel?	Всего Rel	P@k
1	R	1	$\frac{1}{1} = 1.0$
2	R	2	$\frac{2}{2} = 1.0$
3	NR	2	$\frac{2}{3} \approx 0.67$
4	NR	2	$\frac{2}{4} = 0.5$

Проблема: эта выдача лучше предыдущей, а P@4 такой же

Average precision

k	Rel?	AP@k
1	R	$\frac{1\times1/1}{1} = 1.0$
2	NR	$\frac{(1\times1/1)+(0\times1/2)}{2} = 0.5$
3	NR	$\frac{(1\times1/1)+(0\times1/2)+(0\times1/3)}{3} \approx 0.33$
4	R	$\frac{(1\times1/1)+(0\times1/2)+(0\times1/3)+(1\times2/4)}{4} = 0.375$

Average precision @k, пример #2

k	Rel?	AP@k
1	R	$\frac{1\times1/1}{1} = 1.0$
2	R	$\frac{(1\times1/1)+(1\times2/2)}{2} = 1.0$
3	NR	$\frac{(1\times1/1)+(1\times2/2)+(0\times2/3)}{3} \approx 0.67$
4	NR	$\frac{(1\times1/1)+(1\times2/2)+(0\times2/3)+(0\times2/4)}{4} = 0.5$

Average precision @k, пример #2

k	Rel?	AP@k
1	R	$\frac{1\times1/1}{1} = 1.0$
2	R	$\frac{(1\times1/1)+(1\times2/2)}{2} = 1.0$
3	NR	$\frac{(1\times1/1)+(1\times2/2)+(0\times2/3)}{3} \approx 0.67$
4	NR	$\frac{(1\times1/1)+(1\times2/2)+(0\times2/3)+(0\times2/4)}{4} = 0.5$

mAP@k (mean AP@k) — усредненное по корзине запросов

Оценка ранжированной выдачи: DCG

- CG@K (Cumulative Gain at k)

$$CG_k = \sum_{i=1}^{k} rel_i$$

- DCG@K (Discounted Cumulative Gain at k)

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$$ или

- $$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{\log_2(i+1)}$$

Оценка ранжированной выдачи: DCG

NDCG@k (Normalized DCG@k)

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

, где $IDCG$ — Ideal DCG:

$$IDCG_k = \sum_{i=1}^{|REL_k|} \frac{2^{rel_i}-1}{\log_2(i+1)}$$

, где REL_k — все релевантные документы, отсортированные по убыванию релевантности

Оценка ранжированной выдачи: pFound

$$pFound_k = \sum_{i=1}^{k} pLook_i \times pRel_i$$, где:

- $pLook_i$ — вероятность просмотреть i-й документ
- $pRel_i$ — вероятность, что i-й документ релевантный

Оценка ранжированной выдачи: pFound

$$pLook_i = pLook_{i-1} \times (1 - pRel_{i-1}) \times (1 - pBreak)$$

Идея этой формулы:

- пользователь просматривает результаты сверху вниз
- он прекращает просмотр, найдя релевантный документ
- ни у какого документа нет $pRel_i = 1$
  - в статье единственное ненулевое значение 0.4
- с вероятностью $pBreak$ он прекращает просмотр просто так («надоело» ©)
  - в статье $pBreak = 0.15$

Другие метрики, которые не влезли

- ROC-AUC (площадь под кривой ошибок)
- MRR (mean reciprocal rate)

Онлайн

- все метрики выше считаются **оффлайн**, по оценкам релевантности
- но иногда лучше/проще смотреть на поведение пользователей онлайн
  - соответствующие *онлайн-метрики*, например:
    - CTR,
    - время до клика,
    - средняя позиция первого клика,
    - …
  - часто в виде *A/B тестирования*