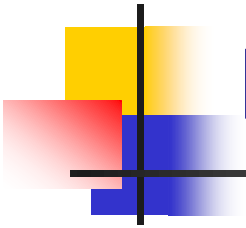
IT4853
Tìm kiếm và trình diễn thông tin

Bài 8. Đánh giá kết quả tìm kiếm (2)

IIR.C8. Evaluation in information retrieval

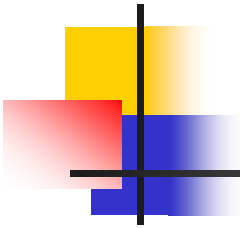
TS. Nguyễn Bá Ngọc, *Bộ môn Hệ thống thông tin, Viện CNTT & TT*
ngocnb@soict.hust.edu.vn

Hà Nội, 2016



Nội dung chính

- 1. MRR
- 2. NDCG
- 3. Xây dựng bộ dữ liệu kiểm thử



MRR

- MRR đánh giá cao kết quả phù hợp ở đầu danh sách.
- MRR thường được sử dụng để đánh giá kết quả tìm kiếm khi chỉ có một văn bản phù hợp:
  - Tìm kiếm trang chủ của một tổ chức, vấn tin về một sự kiện v.v.;
  - Kết quả phù hợp càng xa vị trí đầu danh sách người dùng càng tốn nhiều thời gian tiếp cận văn bản đó;

Trung bình hạng nghịch đảo: MRR: **M**ean **R**eciprocal **R**ank

MRR (2)

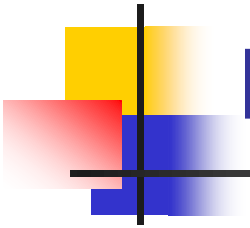
- Gọi K là vị trí của kết quả đầu tiên phù hợp với q

$$RR(q) = \frac{1}{K}$$

- Gọi Q là tập truy vấn mẫu:

$$MRR(Q) = \frac{1}{|Q|} \cdot \sum_{q \in Q} RR(q)$$

$$MRR(Q) = \frac{1}{|Q|} \cdot \sum_{q \in Q} \frac{1}{K_q}$$

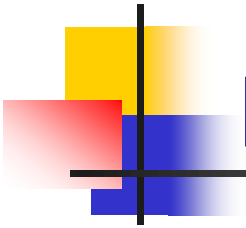


Nội dung chính

- 1. MRR
- 2. NDCG
- 3. Xây dựng bộ dữ liệu

Phù hợp đa mức

- Đánh giá sự phù hợp của văn bản và truy vấn theo nhiều mức khác nhau:
  - Ký hiệu rel_i là mức phù hợp của văn bản d_i;
  - rel = 0 là không phù hợp; $rel_i > rel_j$, thể hiện văn bản d_i phù hợp hơn so với văn bản d_j.

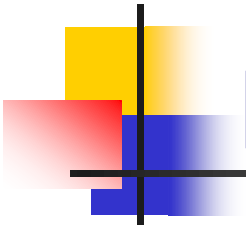


NDCG

- NDCG:
  - Được đo trên bộ dữ liệu kiểm thử phù hợp đa mức;
  - Ngày càng được sử dụng rộng rãi hơn để đánh giá kết quả tìm kiếm trên Web và đánh giá các phương pháp học xếp hạng;
  - Khái niệm cơ bản của NDCG là khái niệm lợi ich.

Thuật ngữ:

N: Normalized: Chuẩn hóa; D: Discounted: cắt giảm;
C: Cumulative: Tổng hợp; G: Gain: Lợi ích;
NDCG: **N**ormalized **D**iscounted **C**umulative **G**ain.

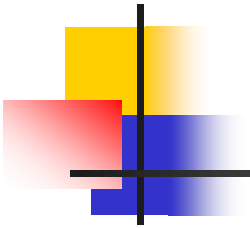


Lợi ích

- Lợi ích của một kết quả tìm kiếm tỉ lệ thuận với mức phù hợp của kết quả: Kết quả càng phù hợp thì càng hữu ích với người dùng, và càng đóng góp nhiều vào lợi ích của tập kết quả.

Thuật ngữ:

Lợi ích: G: Gain



Tổng lợi ích

- CG của n kết quả tìm kiếm đầu tiên
  - CG = $r_1+r_2+...+ r_n$
  - Với r_1, r_2, ...r_n là mức phù hợp của các văn bản

Thuật ngữ:

Tổng lợi ích: CG: Cumulative Gain

Tổng lợi ích thuyên giảm

- Kết quả càng xa vị trí đầu danh sách càng kém hữu ích (lợi ích bị thuyên giảm);
- DCG tại vị trí n
  - DCG = $rel_1 + rel_2/\log_2 2 + \ldots rel_n/\log_2 n$
- $DCG_p = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{log_2 i}$
- Có thể sử dụng hệ cơ số bất kỳ cho hàm log

Thuật ngữ:

Tổng lợi ích thuyên giảm:

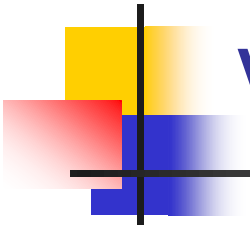
DCG: Discounted Cumulative Gain

Tổng lợi ích thuyên giảm (2)

- Công thức khấu trừ giá trị lợi ích khác:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  - Nhấn mạnh những văn bản có độ phù hợp cao



Ví dụ

- 10 văn bản đã xếp hạng được đánh giá theo thang điểm phù hợp 0-3:

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0

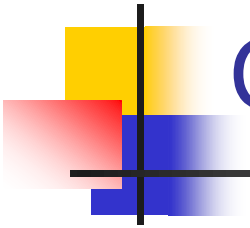
- DG:

  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0

- DCG:

  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61



Chuẩn hóa

- NDCG: là giá trị chuẩn hóa bằng cách chia DCG của tập kết quả cho DCG của xếp hạng mẫu.
  - Xếp hạng mẫu là thứ tự giảm dần mức phù hợp của văn bản;
  - Giá trị chuẩn hóa thích hợp để so sánh những kết quả có số lượng văn bản phù hợp khác nhau.

NDCG: Normalized Discounted Cumulative Gain

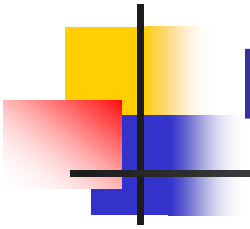
Ví dụ

4 văn bản: d_1, d_2, d_3, d_4

i	Giá trị mẫu		Hàm xếp hạng$_1$		Hàm xếp hạng$_2$	
	Thứ tự văn bản	r_i	Thứ tự văn bản	r_i	Thứ tự văn bản	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309 \quad DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619 \quad MaxDCG = DCG_{GT} = 4.6309$$



Nội dung chính

- 1. MRR
- 2. NDCG
- 3. Xây dựng bộ dữ liệu kiểm thử

Đánh giá tính phù hợp

- Khó khăn: Sự phù hợp là rất trừu tượng
  - Người dùng thường kết luận văn bản có phù hợp hay không sau khi đọc;
  - Những người dùng khác nhau có thể có đánh giá khác nhau về tính phù hợp của văn bản.
- Hướng khắc phục: Cần sử dụng chung một định nghĩa tường minh thế nào là văn bản phù hợp cho cả nhóm xây dựng tập kết quả mẫu.

Ví dụ một truy vấn trong TREC

<top>

<num> Number: 351

<title> Falkland petroleum exploration

<desc> Description:

What information is available on petroleum exploration in the South Atlantic near the Falkland islands?

<narr> Narrative:

Any document discussing petroleum exploration in the South Atlantic near the Falkland Islands is considered relevant. Documents discussing petroleum exploration in continental South America are not relevant.

</top>

Định nghĩa sự phù hợp

- TREC định nghĩa sự phù hợp như sau:

If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant. Only binary judgments ("relevant" or "not relevant") are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document).

Giả sử nếu bạn đang viết một báo cáo về chủ đề đang xét và bạn muốn sử dụng thông tin chứa trong một văn bản cụ thể trong báo cáo của mình thì văn bản đó được coi là phù hợp. Chỉ thực hiện đánh giá nhị phân ("phù hợp" hoặc "không phù hợp"), và một văn bản được coi là phù hợp nếu một phần bất kỳ của nó là phù hợp (không quan tâm phần đó nhỏ tới mức nào nếu so sánh với phần còn lại của văn bản).

Kiểm định đánh giá phù hợp

- Kết quả thu được bởi các thành viên có thể được sử dụng để đánh giá kết quả tìm kiếm nếu đảm bảo tính thống nhất trên một ngưỡng xác định
- Đo sự thống nhất bằng cách nào?

Mức độ thống nhất giữa các bộ kết quả thường được đo bằng hệ số Kappa

Hệ số Kappa

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- $P(E)$ = giá trị mong đợi của tỉ lệ thống nhất ngẫu nhiên,
- $P(A)$ = tỉ lệ thống nhất giữa những đánh giá
- Thường chấp nhận κ trong khoảng [2/3, 1.0].
- Cần điều chỉnh phương pháp đánh giá phù hợp đang sử dụng nếu κ quá nhỏ.

Ví dụ tính chỉ số kappa

		Đánh giá 2		
		Yes	No	Total
Đánh giá 1	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

Theo dõi tỉ lệ số lần thống nhất của kết quả

$P(A) = (300 + 70)/400 = 370/400 = 0.925$

Giá trị biên tổng hợp

$P(\textit{không phù hợp}) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(\textit{phù hợp}) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

Giá trị xác suất của sự thống nhất ngẫu nhiên $P(E)$ =

$P(\textit{không phù hợp})^2 + P(\text{phù hợp})^2 = 0.21252^2 + 0.78782^2 = 0.665$

Chỉ số kappa $\kappa = (P(A) - P(E))/(1 - P(E)) =$

$(0.925 - 0.665)/(1 - 0.665) = 0.776$ (trong khoảng được chấp nhận)

Bài tập 8.1

	GT1	GT2
q_1	NRNNN	NNNNR
q_2	NNRNN	RNNNN

So sánh hai giải thuật theo tham số MRR

Bài tập 8.2

Giả sử hệ thống tìm kiếm trả về tập kết quả là {4, 5, 6, 7, 8}:

a) Tính kappa giữa hai danh sách kết quả đánh giá;

b) Tính P, R và F1 trong trường hợp văn bản được coi là phù hợp nếu cả hai cùng đánh giá là phù hợp;

c) Tính P, R và F1 trong trường hợp văn bản được coi là phù hợp nếu một trong hai đánh giá là phù hợp.

d) Thiết lập hai danh sách kết quả bất kỳ để:

d1) kappa = -1; d2) kappa = 1;

docID	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1