Đánh giá truy hồi thông tin

Nguyễn Mạnh Hiển
hiennm@tlu.edu.vn

Vì sao phải đánh giá?

- Có nhiều mô hình/thuật toán/hệ thống truy hồi thông tin (THTT), vậy cái nào tốt nhất?
- Cái nào tốt nhất cho:
  - Hàm phân hạng (tích vô hướng, côsin, …)
  - Chọn từ khóa (loại bỏ từ dừng, tách gốc từ, …)
  - Định trọng số từ (tf, tf-idf, …)
- Người dùng phải nhìn xuống tới đâu trong danh sách phân hạng trả về để tìm được các văn bản phù hợp?

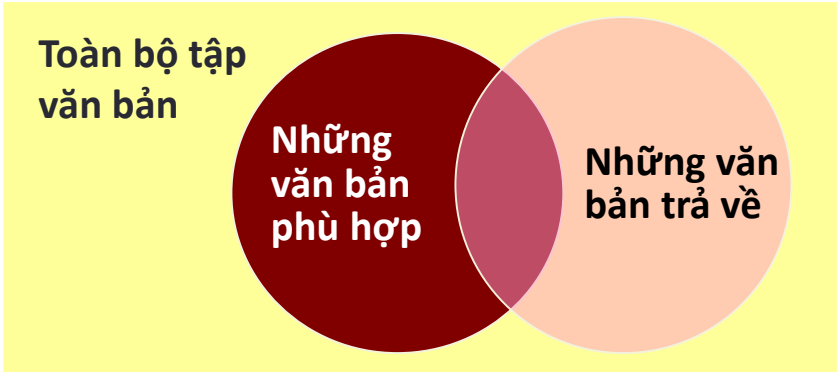
Khó khăn khi đánh giá các hệ THTT

- Tính hiệu quả liên quan đến độ phù hợp của các văn bản trả về.
- Độ phù hợp thường không phải chỉ là đúng/sai (nhị phân) mà là một giá trị liên tục.
- Kể cả khi độ phù hợp chỉ là đúng/sai, khó mà đánh giá được, vì:
  - Đúng/sai phụ thuộc vào từng người dùng.
  - Kể cả với cùng một người dùng, lúc này đúng lúc khác lại không đúng.

Gắn nhãn văn bản

- Bắt đầu với một tập văn bản.
- Lập ra một tập câu truy vấn cho tập văn bản này.
- Nhờ một hoặc nhiều người gắn nhãn tất cả những văn bản phù hợp với mỗi câu truy vấn.
- Thường thì chỉ gắn nhãn đúng/sai.
- Tốn nhiều công sức với những tập văn bản và tập truy vấn cỡ lớn.

Độ chính xác (precision) và độ thu hồi (recall)



không phù hợp	trả về & không phù hợp	không trả về & không phù hợp
phù hợp	trả về & phù hợp	không trả về nhưng phù hợp
	trả về	**không trả về**

$$\text{Độ chính xác (precision)} = \frac{\text{Số văn bản trả về và phù hợp}}{\text{Tổng số văn bản trả về}}$$

$$\text{Độ thu hồi (recall)} = \frac{\text{Số văn bản trả về và phù hợp}}{\text{Tổng số văn bản phù hợp}}$$

Độ chính xác và độ thu hồi (tiếp)

- Độ chính xác:
  - Khả năng trả về những văn bản hầu như sẽ phù hợp.
- Độ thu hồi:
  - Khả năng tìm ra ***tất cả*** những văn bản phù hợp đang có trong tập văn bản đã cho.

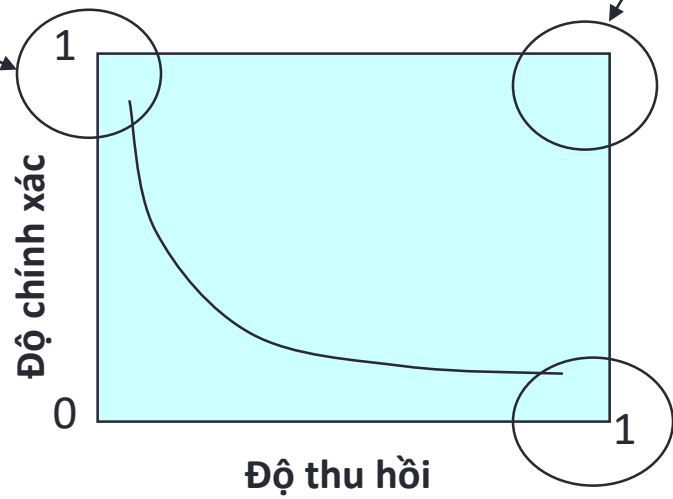
Tính độ thu hồi khó!

- Đôi khi không biết có tất cả bao nhiêu văn bản phù hợp:
  - Lấy mẫu một tập con của tập văn bản đã cho và đánh giá trên tập con đó.
  - Áp nhiều các thuật toán truy hồi khác nhau với cùng một câu truy vấn trên cùng một tập văn bản. Kết hợp các tập văn bản phù hợp được trả về thành tập văn bản phù hợp tổng thể.

Thỏa hiệp giữa độ chính xác và độ thu hồi

Trả về những văn bản phù hợp, nhưng bỏ qua nhiều văn bản khác cũng phù hợp

Vùng lý tưởng



Trả về hầu hết các văn bản phù hợp, nhưng cũng bao gồm nhiều rác

Tính các điểm độ thu hồi/độ chính xác

- Xét một câu truy vấn, tạo một danh sách phân hạng của các văn bản trả về.
- Chỉnh ngưỡng trên danh sách phân hạng để tạo ra các tập văn bản trả về khác nhau, từ đó tính được các cặp độ thu hồi/độ chính xác khác nhau.
- Đánh dấu mỗi văn bản trên danh sách phân hạng có phù hợp hay không theo các nhãn đã gắn trước đây.
- Tính một cặp độ thu hồi/độ chính xác cho mỗi vị trí được đánh dấu là phù hợp trên danh sách phân hạng.

Tính các điểm độ thu hồi/độ chính xác: Ví dụ 1

n	mã vb	phù hợp
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Giả sử tổng số văn bản phù hợp = 6
Xét mỗi điểm đánh dấu:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; P=5/13=0.38

Thiếu một văn bản phù hợp; không đạt tới độ thu hồi 100% được

Tính các điểm độ thu hồi/độ chính xác: Ví dụ 2

n	mã vb	phù hợp
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

Giả sử tổng số văn bản phù hợp = 6
Xét mỗi điểm đánh dấu:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/3=0.667

R=3/6=0.5; P=3/5=0.6

R=4/6=0.667; P=4/8=0.5

R=5/6=0.833; P=5/9=0.556

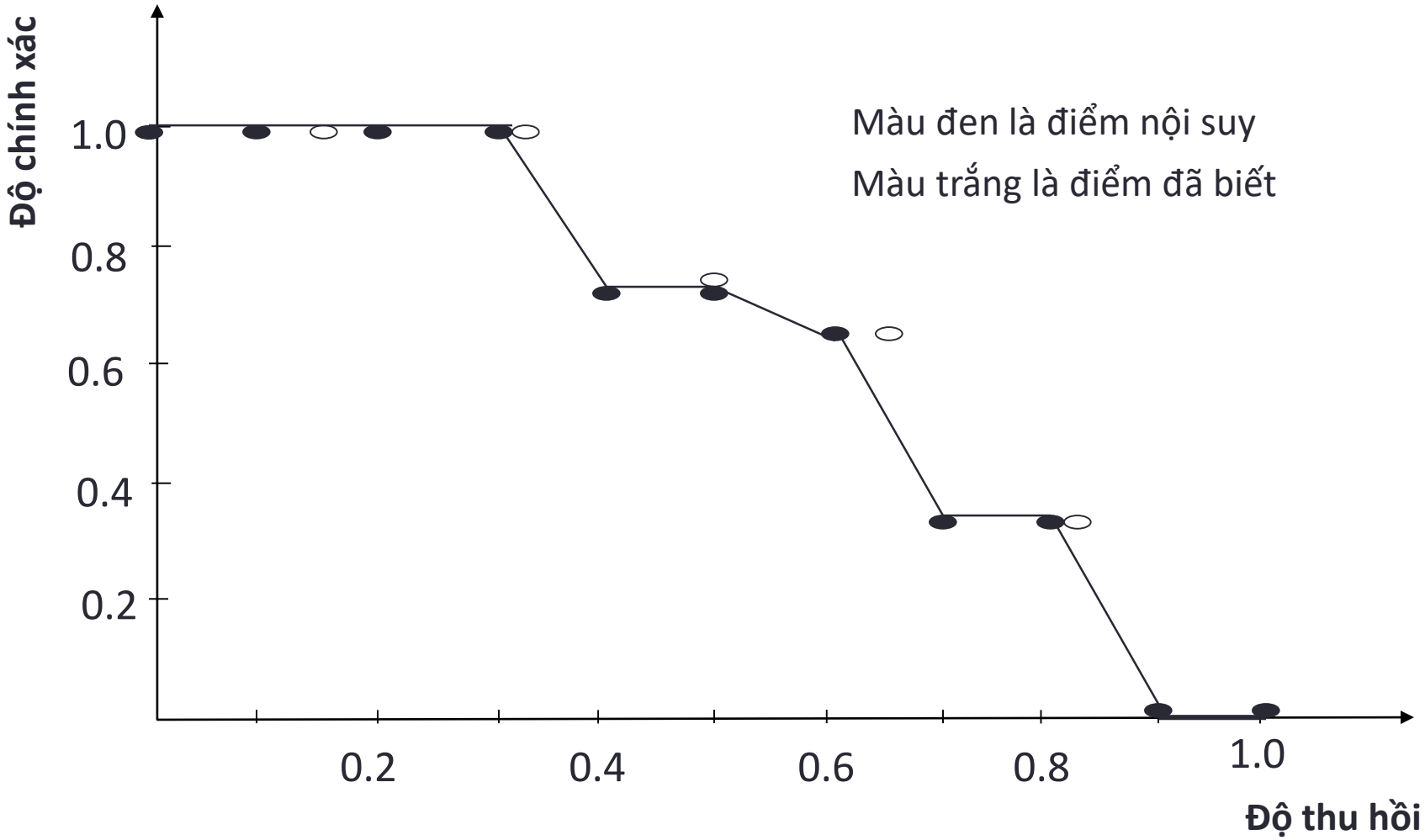
R=6/6=1.0; P=6/14=0.429

Nội suy đường cong R/P

- Với mỗi mức *R* chuẩn, nội suy một giá trị *P*:
  - $R_j \in \{ 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 \}$
  - 11 mức: $R_0 = 0.0$, $R_1 = 0.1$, …, $R_{10}=1.0$
- Giá trị *P* nội suy được ở mức R_j bằng giá trị *P* lớn nhất đã biết giữa các mức R_j và R_{j+1}:

$$P(R_j) = \max_{R_j \le R \le R_{j+1}} P(R)$$

Nội suy đường cong R/P: Ví dụ

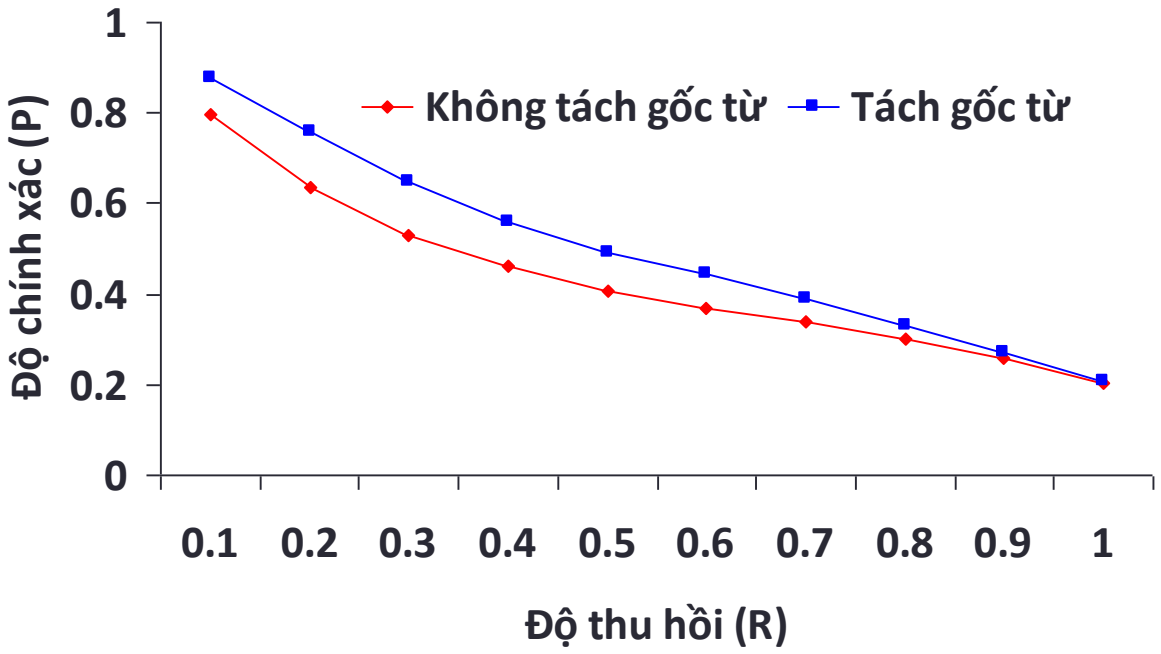


Đường cong R/P trung bình

- Mỗi đường cong R/P ứng với một câu truy vấn cụ thể.
- Tính đường cong R/P cho tất cả các câu truy vấn trong một tập câu truy vấn đang xét.
- Tính và vẽ đường cong R/P trung bình để đánh giá hiệu suất tổng thể của một hệ truy hồi trên một tập văn bản/câu truy vấn đã cho.

So sánh hai hoặc nhiều hệ truy hồi

- Đường cong gần với góc phải-trên hơn của đồ thị phản ánh hiệu suất tốt hơn



R-độ chính xác (*R*-precision)

- Độ chính xác ở vị trí thứ *R* trong danh sách phân hạng cho một câu truy vấn có *R* văn bản phù hợp.

n	mã vb	phù hợp
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R = số văn bản phù hợp = 6

R-độ chính xác = 4/6 = 0.67

Độ đo *F*

- Kết hợp độ thu hồi *R* và độ chính xác *P*:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- So với trung bình cộng, cả hai *P* và *R* đều phải cao để *F* cũng cao.

Phù hợp liên tục

- Hiếm khi các văn bản hoàn toàn phù hợp hay không phù hợp với một câu truy vấn.
- Có nhiều nguồn dẫn tới điểm số phù hợp liên tục (thay vì chỉ đúng/sai):
  - Đánh giá phù hợp trên dải 5 điểm (ví dụ, 1–5 sao).
  - Nhiều người cùng đánh giá.
  - Đánh giá dựa vào số lần click chuột.

Thu hoạch tích lũy (Cumulative Gain – CG)

- Thu hoạch tích lũy tại hạng *n* (CG_n) được tính bằng tổng độ phù hợp (thu hoạch) tại tất cả các hạng từ 1 đến *n*:

$$CG_n = \sum_{i=1}^{n} rel_i$$

rel_i là độ phù hợp của văn bản ở hạng $n = i$.

n	mã vb	phù hợp	CG_n
1	588	1.0	1.0
2	589	0.6	1.6
3	576	0.0	1.6
4	590	0.8	2.4
5	986	0.0	2.4
6	592	1.0	3.4
7	984	0.0	3.4
8	988	0.0	3.4
9	578	0.0	3.4
10	985	0.0	3.4
11	103	0.0	3.4
12	591	0.0	3.4
13	772	0.2	3.6
14	990	0.0	3.6

Thu hoạch tích lũy có khấu trừ (Discounted CG – DCG)

- Người dùng quan tâm nhiều hơn đến các văn bản có hạng (*rank*) cao hơn, vì vậy ta ***khấu trừ*** kết quả bằng một lượng $1/\log_2(rank)$.
- Thu hoạch tích lũy có khấu trừ:

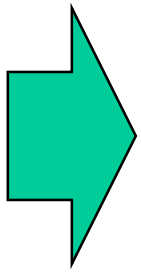
$$DCG_n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}$$

n	mã vb	phù hợp	CG_n	$\log_n$	DCG_n
1	588	1.0	1.0	-	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44

Thu hoạch tích lũy có khấu trừ được chuẩn hóa (Normalized DCG – NDCG)

- Để so sánh các DCG (ví dụ, của các câu truy vấn khác nhau), ta chuẩn hóa chúng về khoảng [0, 1] dựa trên danh sách phân hạng lý tưởng tương ứng.
- Phân hạng lý tưởng (phân hạng theo đúng thứ tự độ phù hợp đã biết):

n	mã vb	phù hợp	CG_n	$\log_n$	DCG_n
1	588	1.0	1.0	0.00	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44



n	mã vb	phù hợp	CG_n	$\log_n$	$IDCG_n$
1	588	1.0	1.0	0.00	1.00
2	592	1.0	2.0	1.00	2.00
3	590	0.8	2.8	1.58	2.50
4	589	0.6	3.4	2.00	2.80
5	772	0.2	3.6	2.32	2.89
6	576	0.0	3.6	2.58	2.89
7	986	0.0	3.6	2.81	2.89
8	984	0.0	3.6	3.00	2.89
9	988	0.0	3.6	3.17	2.89
10	578	0.0	3.6	3.32	2.89
11	985	0.0	3.6	3.46	2.89
12	103	0.0	3.6	3.58	2.89
13	591	0.0	3.6	3.70	2.89
14	990	0.0	3.6	3.81	2.89

Thu hoạch tích lũy có khấu trừ được chuẩn hóa (Normalized DCG – NDCG)

- Chuẩn hóa bằng cách chia cho DCG của danh sách phân hạng lý tưởng tương ứng (Ideal DCG – IDCG):

$$NDCG_n = \frac{DCG_n}{IDCG_n}$$

- Chú ý rằng $NDCG \leq 1$ ở tất cả các hạng.

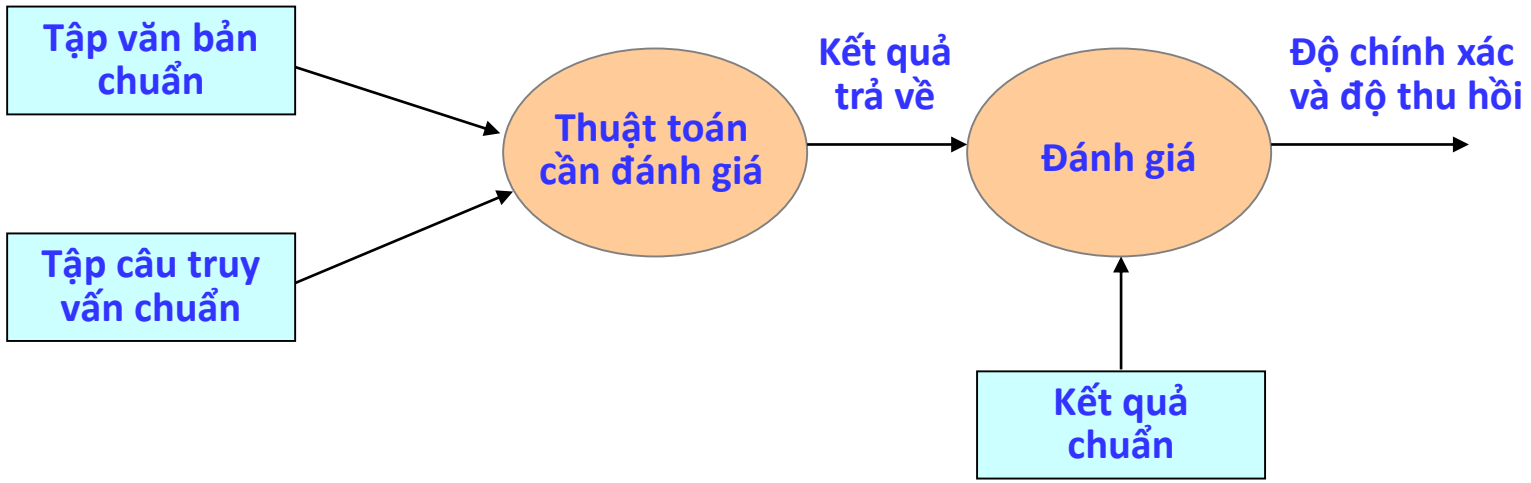
n	vb #	phù hợp	DCG_n	$IDCG_n$	$NDCG_n$
1	588	1.0	1.00	1.00	**1.00**
2	589	0.6	1.60	2.00	**0.80**
3	576	0.0	1.60	2.50	**0.64**
4	590	0.8	2.00	2.80	**0.71**
5	986	0.0	2.00	2.89	**0.69**
6	592	1.0	2.39	2.89	**0.83**
7	984	0.0	2.39	2.89	**0.83**
8	988	0.0	2.39	2.89	**0.83**
9	578	0.0	2.39	2.89	**0.83**
10	985	0.0	2.39	2.89	**0.83**
11	103	0.0	2.39	2.89	**0.83**
12	591	0.0	2.39	2.89	**0.83**
13	772	0.2	2.44	2.89	**0.84**
14	990	0.0	2.44	2.89	**0.84**

Kiểm thử A/B trong một hệ đã triển khai

- Khai thác cơ sở người dùng hiện có để lấy những phản hồi hữu ích.
- Định hướng ngẫu nhiên một tỉ lệ nhỏ người dùng (1–10%) đến một biến thể của hệ thống bao gồm một thay đổi nào đó.
- Đánh giá tính hiệu quả bằng cách đo sự thay đổi trong hành vi click chuột của người dùng:
  - Phần trăm người dùng click trên kết quả đầu tiên (hoặc bất cứ kết quả nào trên trang đầu tiên).

Đánh giá trên tập văn bản chuẩn

- Đánh giá ở đây ***chỉ có hiệu lực*** trên bộ dữ liệu ***chuẩn*** đã cho:
  - Tập văn bản;
  - Tập câu truy vấn;
  - Sự phù hợp đã biết giữa mỗi văn bản và mỗi câu truy vấn.



Ví dụ về tập văn bản chuẩn

Tên tập văn bản	Số văn bản	Số câu truy vấn
CACM	3204	64
CISI	1460	112
CRAN	1400	225
MED	1033	30
TIME	425	83

Bài tập

1. Một hệ THTT trả về 8 văn bản phù hợp và 10 văn bản không phù hợp. Có tổng cộng 20 văn bản phù hợp trong tập văn bản. Tính độ chính xác *P*, độ thu hồi *R* và độ đo *F* của hệ thống trong trường hợp này.

2. Xét một nhu cầu thông tin có 4 văn bản phù hợp trong tập văn bản. Có hai hệ THTT chạy trên tập văn bản này; và 10 kết quả phân hạng cao nhất của chúng như sau (trái cùng là hạng cao nhất, R là phù hợp, N là không phù hợp):

Hệ 1: R N R N N N N N R R

Hệ 2: N R N N R R R N N N

Tính *R*-độ chính xác cho mỗi hệ thống.

Bài tập

3. Xét một nhu cầu thông tin có 8 văn bản phù hợp trong tập văn bản. Một hệ THTT chạy trên tập văn bản này và trả về 20 kết quả xếp hạng cao nhất như sau (trái cùng là hạng cao nhất, R là phù hợp, N là không phù hợp):

R R N N N N N N R N R N N N R N N N N R

Tính độ chính xác *P*, độ thu hồi *R* và độ đo *F* của hệ thống trên 20 kết quả hạng cao nhất này.

Bài tập

4. Bảng bên cạnh cho biết kết quả đánh giá của hai người khác nhau cho một tập 12 văn bản (0 = không phù hợp, 1 = phù hợp). Một hệ THTT trả về các văn bản { 4, 5, 6, 7, 8 }. Tính độ chính xác *P*, độ thu hồi *R* và độ đo *F* của hệ thống trong các trường hợp sau:

a. Một văn bản được xem là phù hợp khi cả hai người đánh giá đều đồng ý như vậy.

b. Một văn bản được xem là phù hợp khi ít nhất một trong hai người đánh giá đồng ý như vậy.

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