

Tìm kiếm thông tin

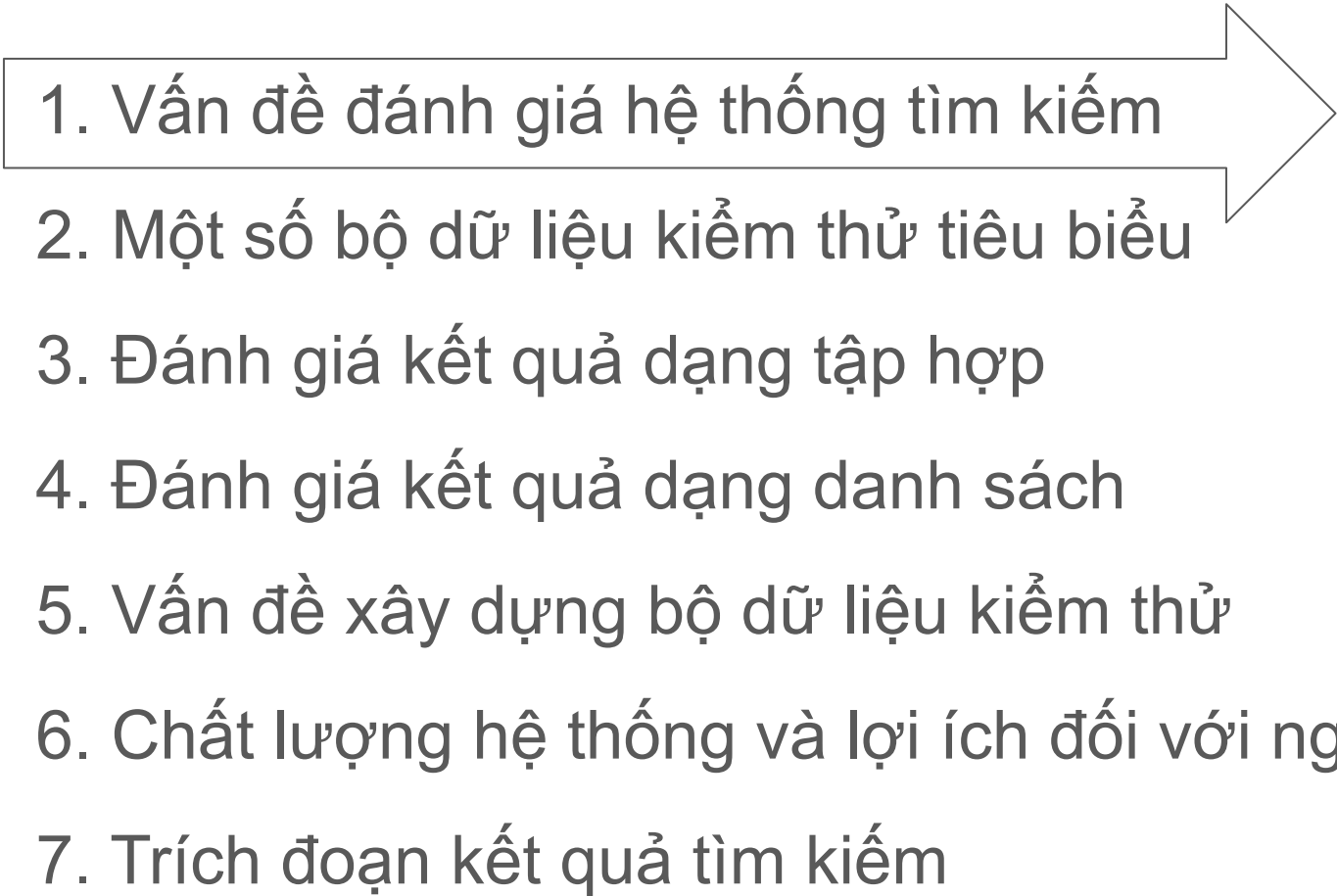
Chương 5. Đánh giá kết quả tìm kiếm thông tin

Soạn bởi: TS. Nguyễn Bá Ngọc

Nội dung

1. Vấn đề đánh giá hệ thống tìm kiếm
2. Một số bộ dữ liệu kiểm thử tiêu biểu
3. Đánh giá kết quả dạng tập hợp
4. Đánh giá kết quả dạng danh sách
5. Vấn đề xây dựng bộ dữ liệu kiểm thử
6. Chất lượng hệ thống và lợi ích đối với người dùng
7. Trích đoạn kết quả tìm kiếm

Nội dung

- 
1. Vấn đề đánh giá hệ thống tìm kiếm
 2. Một số bộ dữ liệu kiểm thử tiêu biểu
 3. Đánh giá kết quả dạng tập hợp
 4. Đánh giá kết quả dạng danh sách
 5. Vấn đề xây dựng bộ dữ liệu kiểm thử
 6. Chất lượng hệ thống và lợi ích đối với người dùng
 7. Trích đoạn kết quả tìm kiếm

Mục đích đánh giá kết quả tìm kiếm

- Đánh giá kết quả tìm kiếm cho phép:
 - So sánh các tùy chỉnh khác nhau của một mô hình;
 - So sánh các mô hình tìm kiếm khác nhau;
 - V.v.
- Các mô hình tìm kiếm chủ yếu được xây dựng dựa trên thực nghiệm, các kết quả đánh giá tạo cơ sở xác định tính hiệu quả của các giả thuyết.

Bộ dữ liệu kiểm thử

- Trong các nghiên cứu tìm kiếm thông tin kết quả tìm kiếm thường được đánh giá dựa trên các bộ dữ liệu kiểm thử.
- Bộ dữ liệu kiểm thử thường bao gồm 3 thành phần:
 - Tập văn bản kiểm thử
 - Tập truy vấn kiểm thử
 - Tập đánh giá tính phù hợp của các văn bản với truy vấn
 - Được thực hiện bởi người;
 - Mô hình phù hợp thường được sử dụng là mô hình nhị phân, kết quả đánh giá có thể nhận giá trị 0 - (nếu không phù hợp) hoặc 1 - (nếu phù hợp).
 - Trong mô hình phù hợp đa mức, kết quả đánh giá có thể nhận nhiều giá trị khác nhau. Mô hình này ít phổ biến hơn.

Vấn đề đánh giá kết quả tìm kiếm

- Câu truy vấn được gửi về hệ thống về bản chất là mô tả của nhu cầu thông tin.
 - Nhu cầu thông tin vốn trừu tượng, người dùng có thể không mô tả được chính xác nhu cầu thông tin.
- Hệ thống tính toán độ phù hợp của văn bản với câu truy vấn, còn người dùng đánh giá kết quả tìm kiếm dựa trên khả năng đáp ứng nhu cầu thông tin.
- Hệ quả là: Văn bản được đánh giá cao nhất vẫn có khả năng không đáp ứng được nhu cầu thông tin.

Ví dụ 5.1. Tính phù hợp vs. Độ tương đồng

- Nhu cầu thông tin: *Muốn biết hoa gạo có mấy cánh*
- Câu truy vấn: *hoa gạo*
- Ví dụ văn bản phù hợp: *Hoa gạo nở đỏ rực một góc trời như xua đi cái lạnh của mùa đông. Khi rụng xuống hoa vẫn giữ nguyên 5 cánh như lúc chớm nở.*
- Ví dụ văn bản không phù hợp: *Cây hoa gạo là một loài cây thân gỗ.*
- Văn bản chứa từ có thể là văn bản phù hợp hoặc không, đồng thời văn bản không chứa từ cũng có thể là văn bản phù hợp (như một tấm ảnh chụp bông hoa gạo cho ví dụ này).

Nội dung

1. Vấn đề đánh giá hệ thống tìm kiếm
2. Một số bộ dữ liệu kiểm thử tiêu biểu
3. Đánh giá kết quả dạng tập hợp
4. Đánh giá kết quả dạng danh sách
5. Vấn đề xây dựng bộ dữ liệu kiểm thử
6. Chất lượng hệ thống và lợi ích đối với người dùng
7. Trích đoạn kết quả tìm kiếm

Một số bộ dữ liệu nhỏ

Cranfield là một trong những bộ dữ liệu đầu tiên, có kích thước nhỏ nhưng đầy đủ. Bên cạnh đó còn có một số bộ dữ liệu khác với kích thước tương tự, các bộ dữ liệu văn bản được liệt kê đều là tiếng Anh:

Tên bộ dữ liệu	Số lượng văn bản	Số lượng truy vấn	Kích thước (Mb)
CACM	3204	64	1.5
CISI	1460	112	1.3
CRAN	1400	225	1.6
MED	1033	30	1.1

Ví dụ 5.2. Một văn bản trong Cranfield

.I 1

.T

experimental investigation of the aerodynamics of a wing in a slipstream .

.A

brenckman,m.

.B

j. ae. scs. 25, 1958, 324.

.W

experimental investigation of the aerodynamics of a wing in a slipstream . an experimental study of a wing in a propeller slipstream was made in order to determine the spanwise distribution of the lift increase due to slipstream at different angles of attack of the wing and at different free stream to slipstream velocity ratios . the results were intended in part as an evaluation basis for different theoretical treatments of this problem .

the comparative span loading curves, together with supporting evidence, showed that a substantial part of the lift increment produced by the slipstream was due to a /destalling/ or boundary-layer-control effect . the integrated remaining lift increment, after subtracting this destalling lift, was found to agree well with a potential flow theory .

an empirical evaluation of the destalling effects was made for the specific configuration of the experiment .

Các hội nghị TREC

- TREC - Text REtrieval Conferences
- Bắt đầu từ 1992
- Mục đích: Đánh giá các mô hình tìm kiếm bằng các bộ dữ liệu kiểm thử lớn
- Được tổ chức dưới hình thức hội thảo, các thành viên sau khi đăng ký có thể:
 - Nhận dữ liệu kiểm thử, thực hiện tìm kiếm, và gửi kết quả tìm được theo mô hình.
 - Các kết quả được đánh giá để so sánh tính hiệu quả của các mô hình.
 - Các kết quả thực nghiệm được chia sẻ ở hội nghị.

Ví dụ 5.3. Một văn bản trong TREC

```
<DOC>
<DOCNO> WSJ880406-0090 </DOCNO>
<HL> AT&T Unveils Services to Upgrade Phone Networks Under Global Plan </HL>
<AUTHOR> Janet Guyon (WSJ Staff) </AUTHOR>
<DATELINE> NEW YORK </DATELINE>
<TEXT>
    American Telephone & Telegraph Co. introduced the first of a new generation
    of phone services with broad implications for computer and communications
    equipment markets.
    AT&T said it is the first national long-distance carrier to announce prices
    for specific services under a world-wide standardization plan to upgrade phone
    networks. By announcing commercial services under the plan, which the industry
    calls the Integrated Services Digital Network, AT&T will influence evolving
    communications standards to its advantage, consultants said, just as
    International Business Machines Corp. has created de facto computer standards
    favoring its products.
    .
    .
    .
</TEXT>
</DOC>
```

Ví dụ 5.4. Một truy vấn trong TREC

```
<top>
<head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing

<desc> Description:
Document will identify a type of natural language processing technology which
is being developed or marketed in the U.S.

<narr> Narrative:
A relevant document will identify a company or institution developing or
marketing a natural language processing technology, identify the technology,
and identify one or more features of the company's product.

<con> Concept(s):
1. natural language processing
2. translation, language, dictionary, font
3. software applications

<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>
```

Đánh giá tính phù hợp ở TREC

- Các bộ dữ liệu đều quá lớn để có thể đánh giá đầy đủ
- Các kết quả thu được bởi các mô hình khác nhau trên cùng 1 bộ dữ liệu kiểm thử và 1 câu truy vấn được gom thành nhóm:
 - Ví dụ, top 100 kết quả từ các mô hình;
 - Tính phù hợp được đánh giá trong phạm vi các nhóm.
- Các nghiên cứu đã cho thấy
 - Một số văn bản phù hợp đã bị bỏ qua tuy nhiên không làm thay đổi kết quả đánh giá tính hiệu quả của các mô hình.
 - Sự khác biệt trong các đánh giá của người thẩm định không ảnh hưởng tới kết quả đánh giá các mô hình.
 - Các đánh giá sử dụng được cho các hệ thống không tham gia thực nghiệm.

Các bộ dữ liệu của TREC

- Chỉ bao gồm các tập văn bản và các câu truy vấn được biên soạn theo các chủ đề khác nhau;
- Các đánh giá được thực hiện cho các kết quả tìm kiếm;
- Các bộ dữ liệu được tạo mới và cập nhật hàng năm để sử dụng cho hội thảo.

Nội dung

1. Vấn đề đánh giá hệ thống tìm kiếm
2. Một số bộ dữ liệu kiểm thử tiêu biểu
3. Đánh giá kết quả dạng tập hợp
4. Đánh giá kết quả dạng danh sách
5. Vấn đề xây dựng bộ dữ liệu kiểm thử
6. Chất lượng hệ thống và lợi ích đối với người dùng
7. Trích đoạn kết quả tìm kiếm

Độ chính xác và độ đầy đủ

- Độ chính xác là tỉ lệ văn bản phù hợp trong số văn bản được trả về:

$$\text{Precision} = \#(\text{văn bản phù hợp trả về}) / \#(\text{văn bản trả về})$$

- Độ đầy đủ là tỉ lệ văn bản phù hợp được trả về trong số văn bản phù hợp có trong bộ dữ liệu kiểm thử

$$\text{Recall} = \#(\text{văn bản phù hợp trả về}) / \#(\text{văn bản phù hợp})$$

Ký hiệu: P - độ chính xác; R - độ đầy đủ; # - số lượng.

Ví dụ 5.5. Tính P/R

Giả sử tập văn bản phù hợp với truy vấn đang được khảo sát:

Rel = {3, 9, 10, 11, 14, 15, 20, 35}

P = ?

R = ?

Rank	Doc#	Rel?
1	5	
2	3	YES
3	10	
4	35	YES
5	4	
6	270	
7	14	YES
8	15	YES
9	11	YES
10	1	

Ví dụ 5.5. Tính $P/R_{(2)}$

Giả sử tập văn bản phù hợp với truy vấn đang được khảo sát:

$Rel = \{3, 9, 10, 11, 14, 15, 20, 35\}$

$$P = 5/10 = 0.5$$

$$R = 5/8 = 0.63$$

Rank	Doc#	Rel?
1	5	
2	3	YES
3	10	
4	35	YES
5	4	
6	270	
7	14	YES
8	15	YES
9	11	YES
10	1	

Độ chính xác tổng quát

- Độ chính xác tổng quát / Accuracy thường được sử dụng để đánh giá kết quả phân lớp.
- Có thể coi vấn đề tìm kiếm thông tin như 1 vấn đề phân lớp: Cho một truy vấn hệ thống tìm kiếm phân lớp các văn bản thành hai lớp: "Phù hợp" và "Không phù hợp"
- Độ chính xác tổng quát: Tỷ lệ văn bản được phân lớp đúng dựa trên cả 2 lớp:
 - $$\frac{(\text{\#văn bản phù hợp được trả về} + \text{\#văn bản không phù hợp không được trả về})}{\text{\#văn bản trong bộ dữ liệu kiểm thử}}$$

Có thể đánh giá hệ thống tìm kiếm theo độ chính xác tổng quát hay không?

Độ chính xác tổng quát₍₂₎

- Làm sao để xây dựng một máy tìm kiếm với độ chính xác tổng quát khoảng 99.999%?

Snoogle.com

Tìm kiếm:

Không tìm thấy kết quả nào

- Số lượng văn bản phù hợp thường chiếm một tỉ lệ rất nhỏ trong bộ dữ liệu
 - Hệ thống không trả về kết quả nào là có thể đạt độ chính xác tổng quát rất cao

Máy tìm kiếm như vậy tất nhiên là không hữu ích.

Ma trận nhầm lẫn

Thống kê số lượng văn bản trong tập văn bản theo hai cặp dấu hiệu: phù hợp/không phù hợp, trả về/không trả về

	Phù hợp	Không phù hợp
Trả về	Đúng - trả về (tp)	Sai - Trả về (fp)
Không trả về	Sai - không trả về (fn)	Đúng - Không trả về (tn)

$$P = tp / (tp + fp)$$

$$R = tp / (tp + fn)$$

$$ACC = (tp + tn) / (tp + fp + fn + tn)$$

Các giá trị thống kê trong ma trận nhầm lẫn là cơ sở để tính nhiều độ đo khác nhau

Độ chính xác và độ đầy đủ

- Có thể đạt độ đầy đủ cao bằng cách trả về nhiều kết quả
 - Khi trả về thêm các kết quả độ đầy đủ sẽ giữ nguyên hoặc tăng lên (không giảm). Độ đầy đủ luôn = 100% khi trả về tất cả văn bản trong bộ dữ liệu
- Tuy nhiên độ chính xác có thể giảm khi tăng số lượng kết quả trả về.
- Có những trường hợp độ đầy đủ có vai trò quan trọng hơn độ chính xác, ví dụ: Khi tổng hợp thông tin về một chủ đề người dùng có thể rà soát tập kết quả để gom đầy đủ thông tin, và có thể chấp nhận độ chính xác thấp.
- Nhưng cũng có những trường hợp (có thể là phần lớn trường hợp trong môi trường Web) độ chính xác quan trọng hơn, ví dụ: Khi tra cứu thông tin người dùng thường muốn có câu trả lời ngay trong những kết quả đầu tiên.

Độ đo F

Ý nghĩa toán học của độ đo F

- Độ đo F là một đại lượng kết hợp độ chính xác và độ đầy đủ (trung bình điều hòa có trọng số):

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} \quad F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

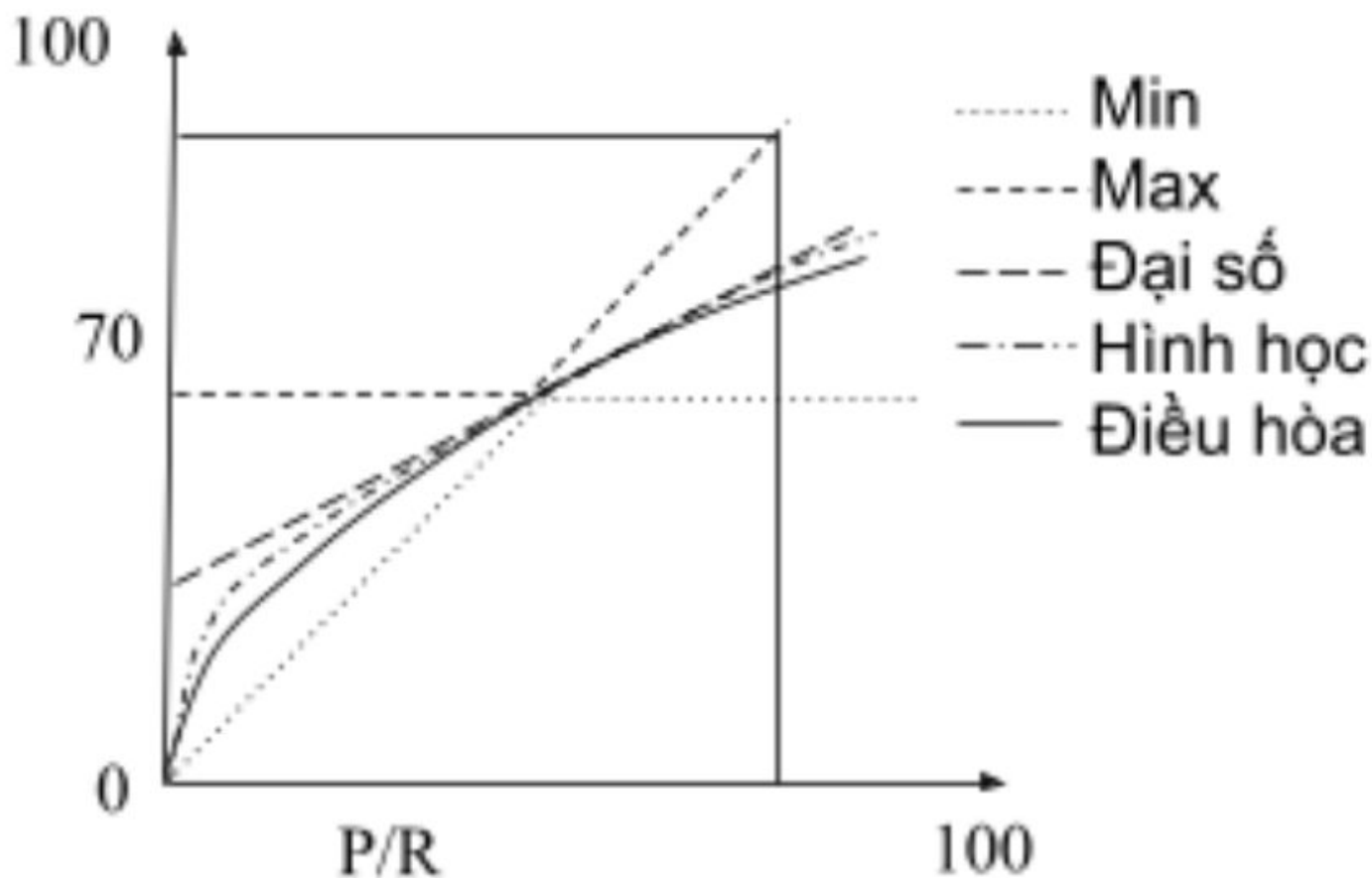
- $\alpha \in [0, 1], \beta^2 \in [0, +\infty]$
- Một trường hợp cụ thể của độ đo F thường được sử dụng là độ đo F_1 - cân bằng P và R
 - $\beta = 1$ hoặc $\alpha = \frac{1}{2}$
- Các trường hợp đặc biệt: Nếu $\beta = 0$ thì F là độ chính xác; Nếu $\beta = +\infty$ (rất lớn) thì F là độ đầy đủ.

Trong kịch bản tìm kiếm có độ đầy đủ được ưu tiên cao hơn thì nên thiết lập β trong miền giá trị nào?

Vì sao nên sử dụng độ đo F?

- Trung bình điều hòa có trọng số của P và R: Đại lượng nghịch đảo của trung bình có trọng số của đại lượng nghịch đảo của P và đại lượng nghịch đảo của R
 - Chúng ta lấy nghịch đảo 2 lần trong độ đo F
- Vì sao không kết hợp theo cách khác: Ví dụ, lấy trung bình đại số?
- Chúng ta muốn: Những hệ thống có độ chính xác hoặc độ đầy đủ rất thấp thì kết quả đánh giá cũng phải rất thấp (dù đại lượng còn lại có thể rất cao).
 - Nghịch đảo của đại lượng vô cùng nhỏ là đại lượng vô cùng lớn và nghịch đảo của đại lượng vô cùng lớn là đại lượng vô cùng nhỏ

So sánh độ đo F và các đại lượng khác



Độ đo F có tính đến sự bù trừ của hai đại lượng. Nếu R rất cao và P thấp (như trường hợp trả về tất cả các văn bản) hoặc ngược lại thì F vẫn thấp.

Ví dụ 5.6. Tính F_1

Giả sử tập văn bản phù hợp với truy vấn đang được khảo sát:

$Rel = \{3, 9, 10, 11, 14, 15, 20, 35\}$

$$P = 5/10 = 0.5$$

$$R = 5/8 = 0.63$$

$$F_1 = ?$$

Rank	Doc#	Rel?
1	5	
2	3	YES
3	10	
4	35	YES
5	4	
6	270	
7	14	YES
8	15	YES
9	11	YES
10	1	

Ví dụ 5.6. Tính $F_1(2)$

Giả sử tập văn bản phù hợp với truy vấn đang được khảo sát:

$Rel = \{3, 9, 10, 11, 14, 15, 20, 35\}$


$$P = 5/10 = 0.5$$

$$R = 5/8 = 0.63$$

$$F_1 = 2 * 0.5 * 0.63 / (0.5 + 0.63) \\ = 0.56$$

Rank	Doc#	Rel?
1	5	
2	3	YES
3	10	
4	35	YES
5	4	
6	270	
7	14	YES
8	15	YES
9	11	YES
10	1	

Nội dung

1. Vấn đề đánh giá hệ thống tìm kiếm
 2. Một số bộ dữ liệu kiểm thử tiêu biểu
 3. Đánh giá kết quả dạng tập hợp
 4. Đánh giá kết quả dạng danh sách
 5. Vấn đề xây dựng bộ dữ liệu kiểm thử
 6. Chất lượng hệ thống và lợi ích đối với người dùng
 7. Trích đoạn kết quả tìm kiếm
- 

Đánh giá kết quả dạng danh sách

- Trong tìm kiếm có xếp hạng các văn bản được trả về dưới dạng danh sách.
- Hệ thống có thể trả về số lượng kết quả bất kỳ, danh sách kết quả có thể được chia thành nhiều trang.
- Người dùng thường ưa thích những kết quả ở đầu danh sách.

Đường cong P/R là một công cụ đánh giá phổ biến cho các danh sách kết quả

Đường cong P/R

Vẽ đường cong P/R

- Ký hiệu $P@i$ và $R@i$ lần lượt là độ chính xác và độ đầy đủ trong giới hạn i kết quả tìm kiếm đầu tiên
- $P@i = \#(\text{văn bản phù hợp trong } i \text{ kết quả đầu tiên})/i$
- $R@i = \#(\text{văn bản phù hợp trong } i \text{ kết quả đầu tiên})/\#(\text{văn bản phù hợp trong bộ dữ liệu})$
- Ví dụ:
 - Giả sử, kết quả tìm kiếm là: $d1^*$, $d2$, $d3^*$, $d4$, $d5^*$
 - ... và có 5 văn bản phù hợp trong bộ dữ liệu.
 - $P@3 = 2/3$ $R@3 = 2/5$
 - $P@4 = 2/4$ $R@4 = 2/5$
 - $P@5 = 3/5$ $R@5 = 3/5$

Vẽ đường cong P/R ₍₂₎

- Cho i biến thiên từ 1 đến hết danh sách kết quả tìm kiếm (phạm vi đánh giá);
- Đo $P@i$ và $R@i$ tại các vị trí i của danh sách kết quả;
- Nối các điểm $(R@i, P@i)$ trên mặt phẳng ta thu được đường cong P/R .

Đường cong P/R thể hiện sự biến thiên của P và R khi mở rộng danh sách kết quả (R không giảm), trên đường cong P/R chúng ta có thể tra cứu được giá trị độ chính xác ở từng mức độ đầy đủ cụ thể.

Ví dụ 5.7. Vẽ đường cong P/R

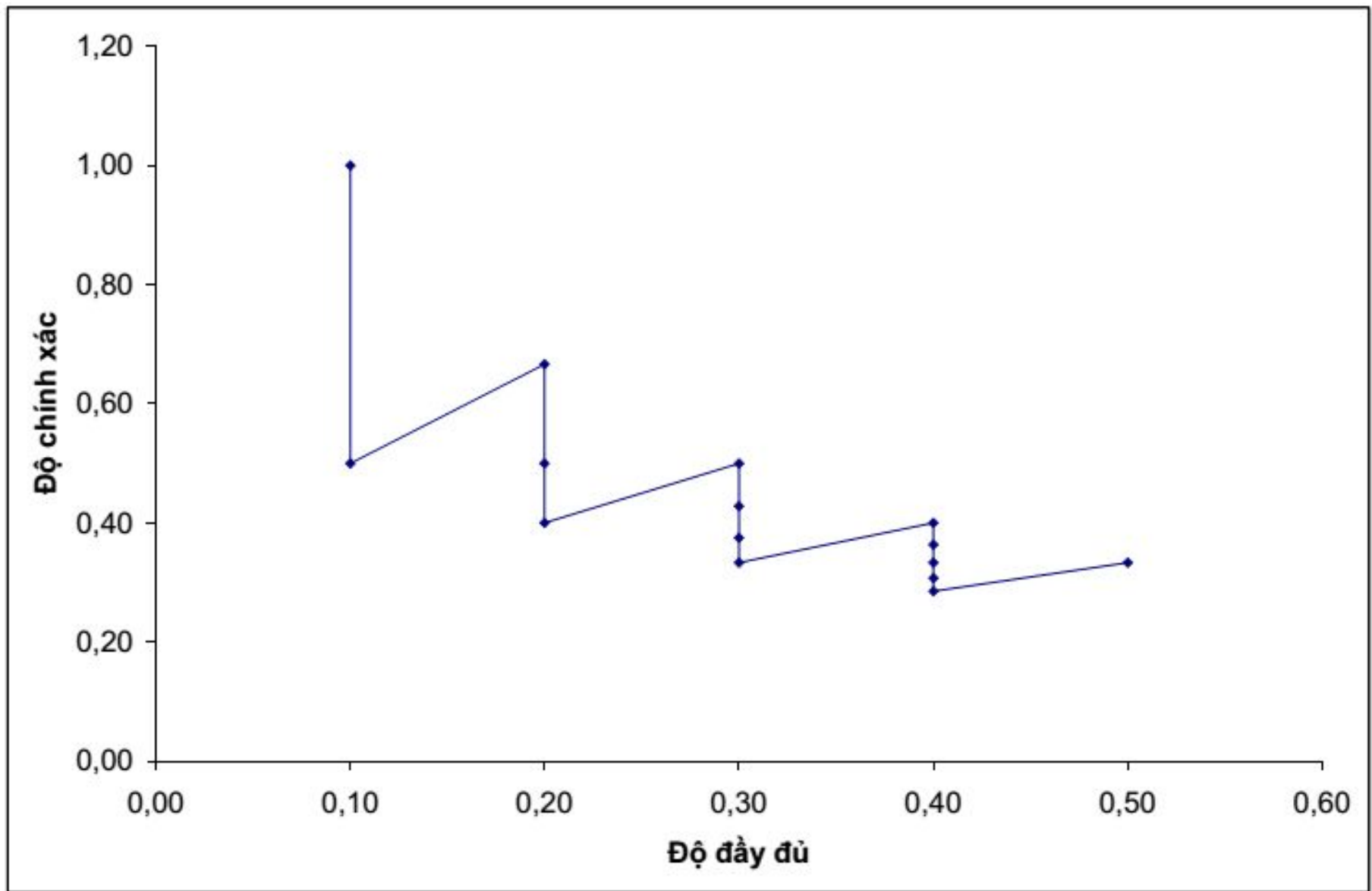
Cho biết tập kết quả phù hợp với truy vấn q chứa 10 văn bản:

$$R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{46}, d_{56}, d_{71}, d_{89}, d_{123}\}.$$

Và danh sách kết quả được trả về bởi hệ thống như sau:

- | | | |
|----------------|----------------|---------------|
| 1. d_{123} * | 6. d_9 * | 11. d_{38} |
| 2. d_{84} | 7. d_{515} | 12. d_{48} |
| 3. d_{56} * | 8. d_{129} | 13. d_{250} |
| 4. d_6 | 9. d_{187} | 14. d_{113} |
| 5. d_8 | 10. d_{25} * | 15. d_3 * |

Thử vẽ đường cong P/R với các dữ liệu được cho



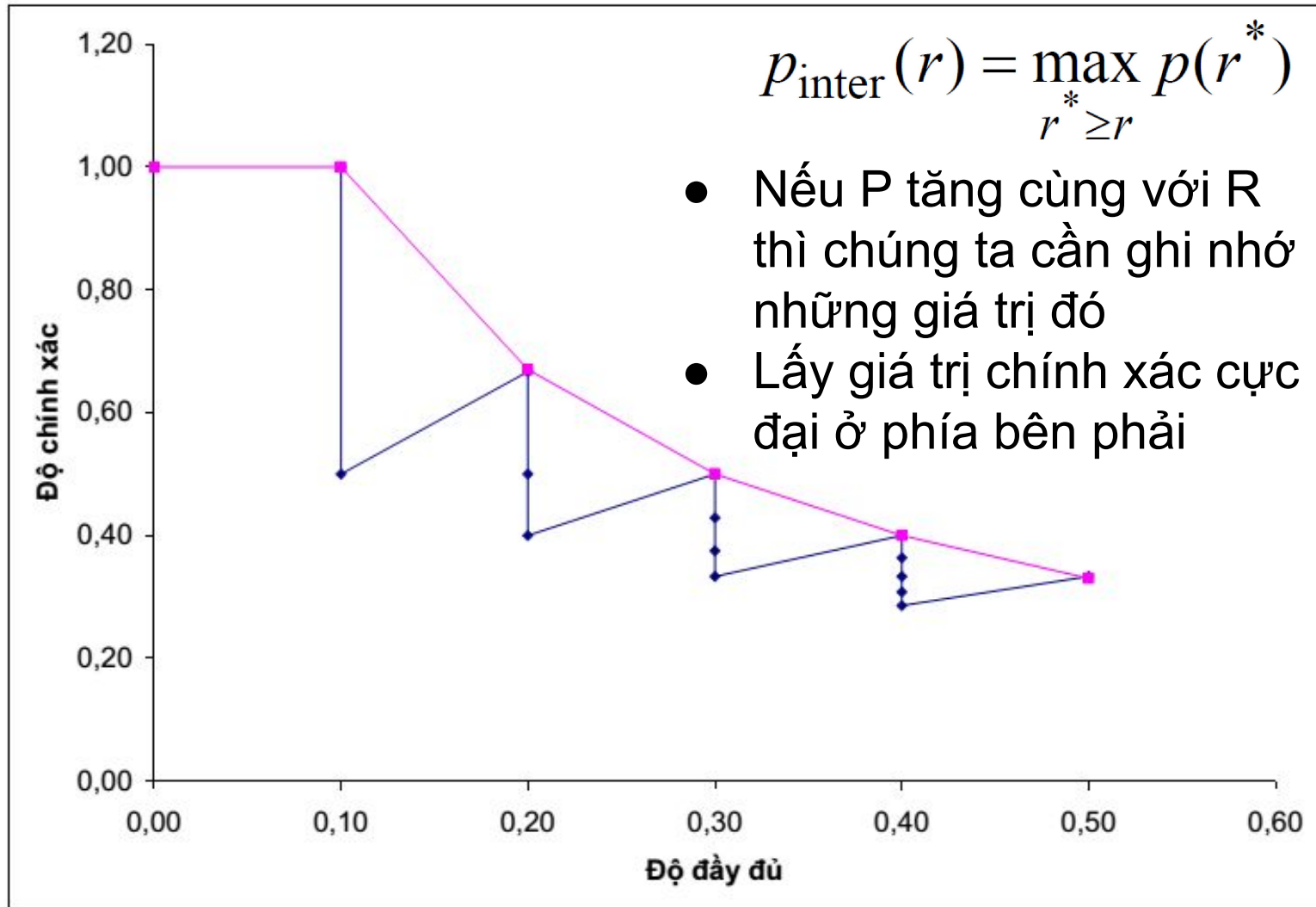
Mỗi điểm trên đồ thị tương ứng với $P@k$ và $R@k$ của một giá trị k , ($k = 1, 2, 3, 4, \dots$).

Lấy trung bình trên nhiều truy vấn

- Đường cong độ chính xác/độ đầy đủ cho một truy vấn không đủ tin cậy để so sánh các mô hình.
- Chúng ta cần lấy trung bình nhiều kết quả trên một tập truy vấn
- Tuy nhiên các truy vấn khác nhau có thể có số lượng kết quả phù hợp khác nhau

Làm sao để xác định giá trị trung bình trên 1 tập truy vấn?

Giá trị nội suy của độ chính xác

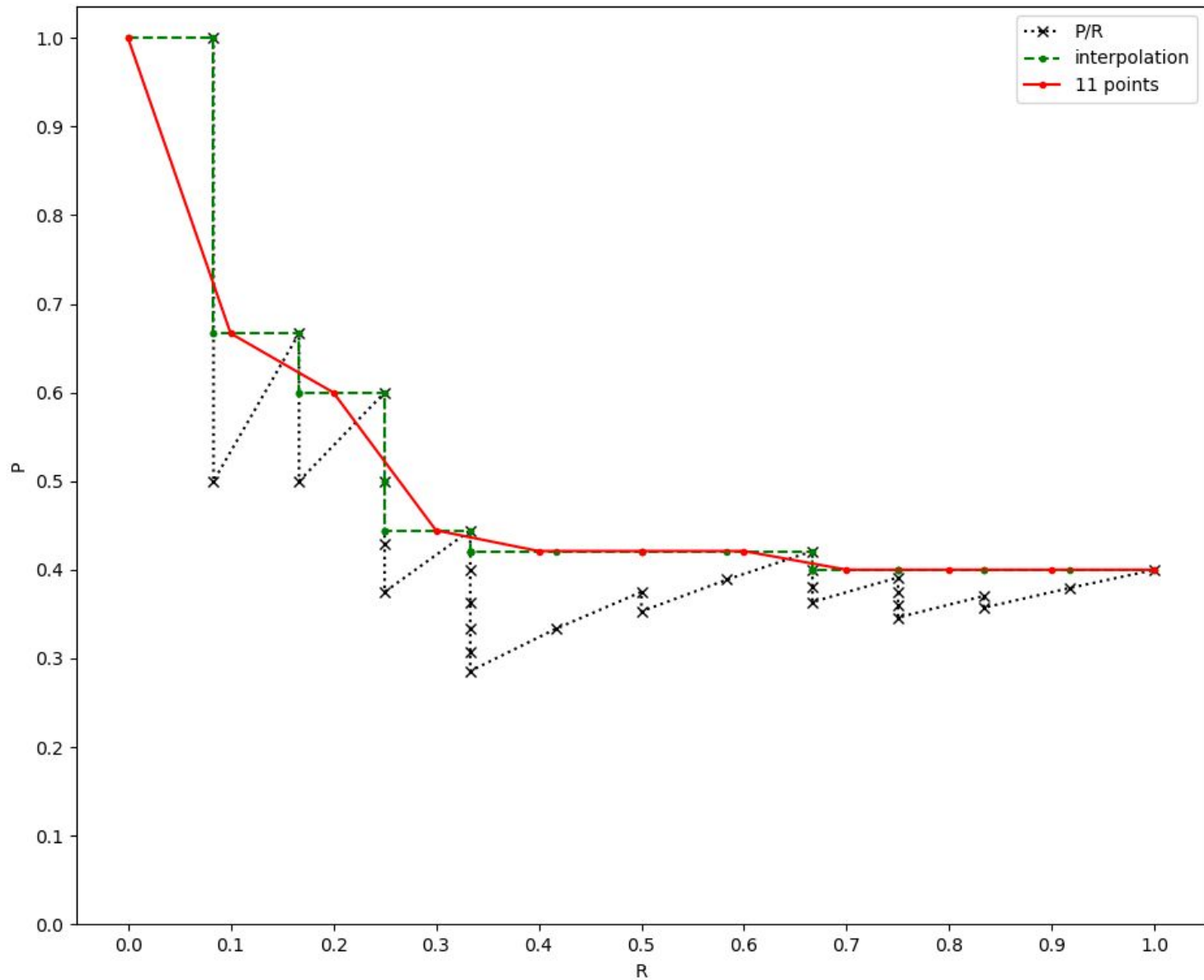


Giả thuyết: Người dùng sẵn sàng rà soát thêm kết quả tìm kiếm nếu phần sau danh sách có chứa kết quả phù hợp.

Đánh giá dựa trên P/R

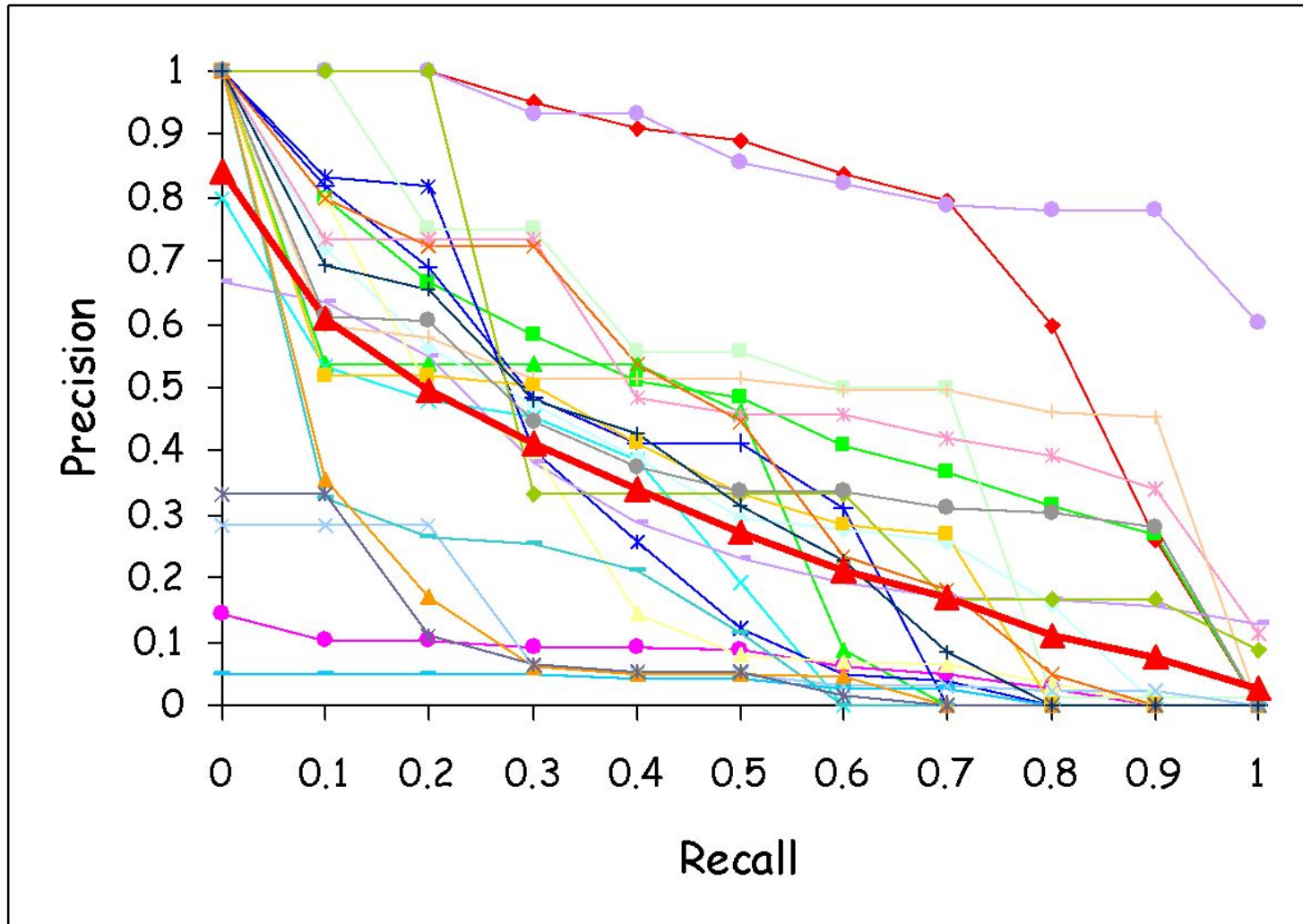
- Đồ thị có tính trực quan, tuy nhiên để có thể so sánh chúng ta cần các đại lượng có tính tổng hợp
 - Ví dụ, độ chính xác ở một ngưỡng cố định
 - $P@k$, Trong môi trường Web, người dùng thường muốn có những kết quả chính xác ở đầu danh sách.
- Độ đo tiêu chuẩn trong các cuộc thi TREC
 - Lấy trung bình giá trị nội suy của độ chính xác ở 11 điểm đầy đủ tiêu chuẩn: $R = 0, 0.1, 0.2, \dots, 1.0$
 - Đánh giá ở tất cả các mức đầy đủ

Ví dụ 5.8. P/R và các giá trị nội suy



Ví dụ 5.9. Các đường cong P/R nội suy

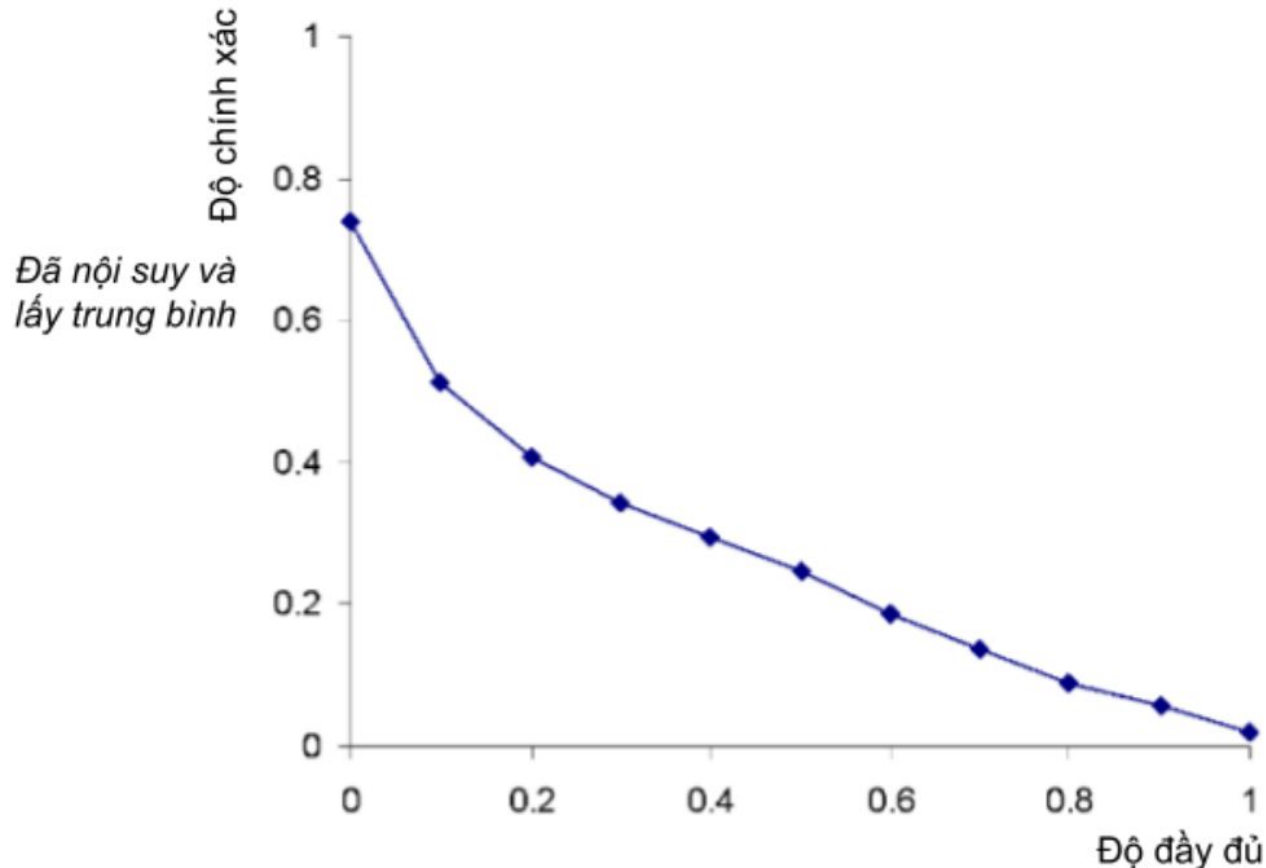
Sau khi lấy nội suy P ở 11 điểm R tiêu chuẩn. Để có thể so sánh chúng ta cần tiếp tục lấy trung bình các giá trị.



[Ellen Voorhees]

Ví dụ 5.10. Các giá trị nội suy tiêu chuẩn

Một kết quả được coi là tốt



- Với mỗi truy vấn: Tính giá trị nội suy cho độ chính xác ở tất cả các mức đầy đủ 0.0, 0.1, ..., 1.0
- Sau đó lấy trung bình cho tất cả các truy vấn

So sánh ROC và P/R

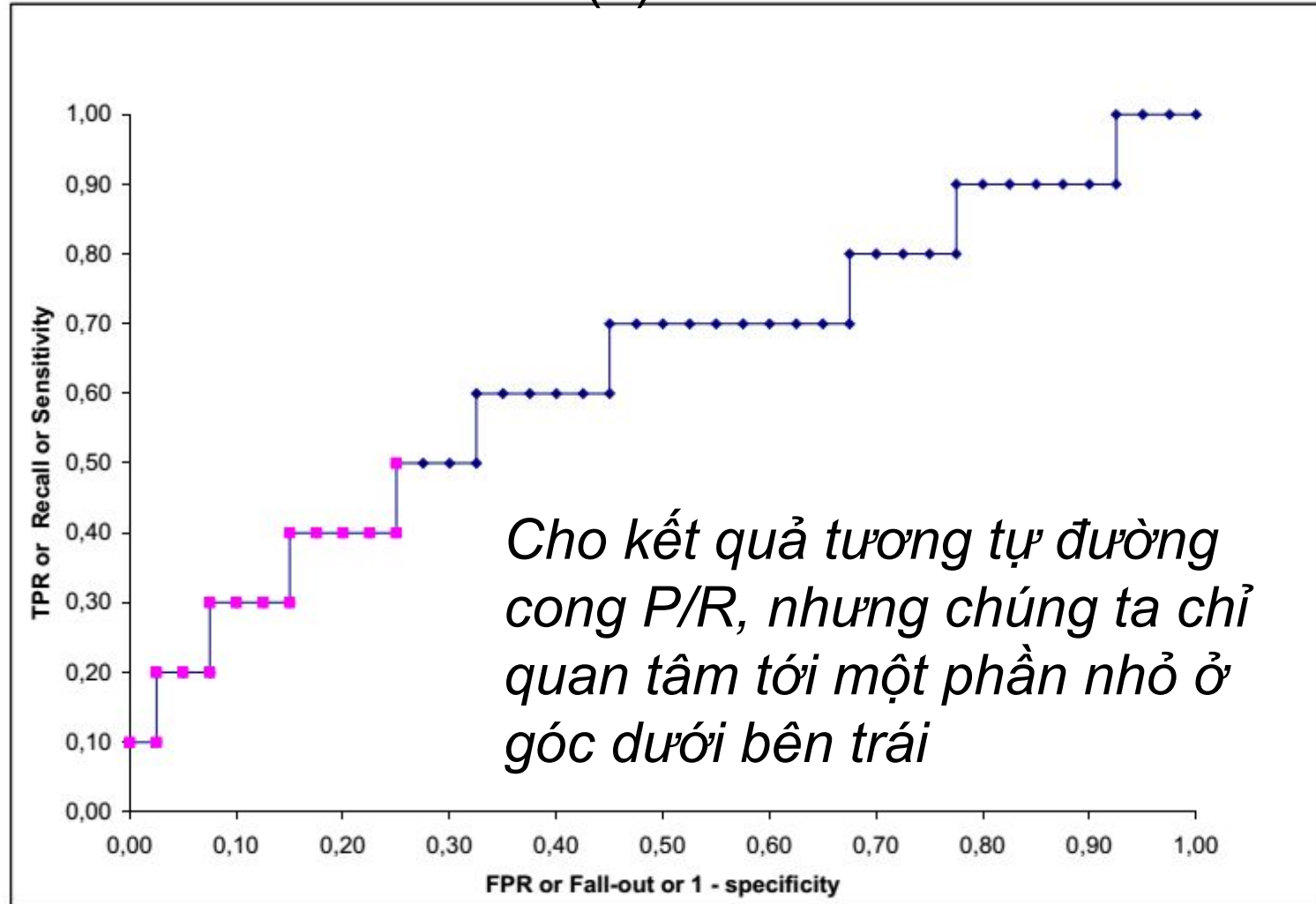
Đường cong ROC

- ROC - Receiver Operatin Characteristics
- ROC: TPR/FPR, là một công cụ khác thường được sử dụng để đánh giá các kết quả phân lớp.

Một số đại lượng được tính dựa trên ma trận nhầm lẫn:

- $TPR = Recall = Sensitivity = TP / (TP + FN) = p(\text{trả về} | \text{phù hợp})$
- $FPR = Fall-out = FP / (FP + TN) = p(\text{trả về} | \text{không phù hợp})$
- $Specificity = TN / (FP + TN) = p(\text{không trả về} | \text{không phù hợp})$
- $FPR = 1 - Specificity$

Đường cong ROC₍₂₎



Quan sát thuận tiện hơn trên đường cong P/R

Các giá trị trung bình

Độ chính xác trung bình - AP

- AP - Average Precision
- Với 1 truy vấn q , chúng ta ký hiệu vị trí của các kết quả phù hợp trong danh sách kết quả là: K_1, K_2, \dots, K_R , trong đó R là số lượng kết quả phù hợp có trong bộ dữ liệu kiểm thử.
 - Với những văn bản phù hợp nhưng không được trả về chúng ta quy ước thành phần $P@K_i$ tương ứng bằng 0.
- Chúng ta có độ chính xác trung bình:

$$AP = \frac{1}{R} \sum P@K_i$$

Ví dụ 5.10. Tính AP

- Số lượng văn bản phù hợp với q trong bộ dữ liệu kiểm thử là 5. Yêu cầu tính AP cho danh sách kết quả? Trong 2 trường hợp:

Trường hợp 1: $d1^* d2 d3^* d4 d5^*$

Trường hợp 2: $d1^* d2 d3^* d4 d6$

Chúng ta có :

Trường hợp 1: $AP = 1/5 * (1/1 + 2/3 + 3/5 + 0 + 0) \approx 0.45$

Trường hợp 2: $AP = 1/5 * (1/1 + 2/3 + 0 + 0 + 0) \approx 0.33$

Ví dụ 5.11. Tính $AP_{(2)}$

- Số lượng văn bản phù hợp với q trong bộ dữ liệu kiểm thử là 5. Yêu cầu tính AP cho danh sách kết quả? Trong 2 trường hợp:

Trường hợp 1: $d1^* d2 d3^* d4 d5^*$

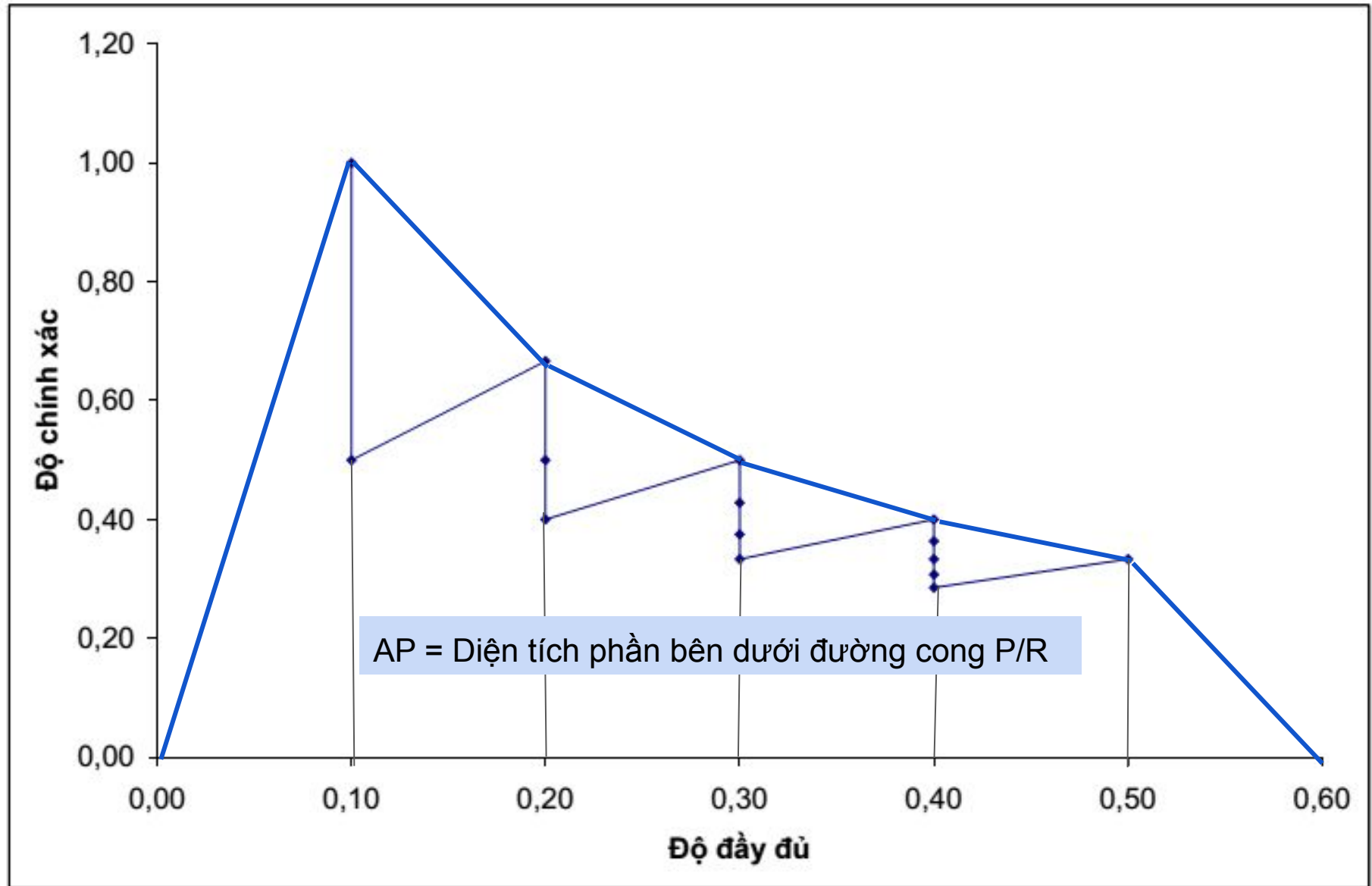
Trường hợp 2: $d1^* d2 d3^* d4 d6$

Chúng ta có :

Trường hợp 1: $AP = 1/5 * (1/1 + 2/3 + 3/5 + 0 + 0) \approx 0.45$

Trường hợp 2: $AP = 1/5 * (1/1 + 2/3 + 0 + 0 + 0) \approx 0.33$

Ý nghĩa hình học của AP



Giá trị trung bình của AP

- MAP - Mean Average Precision
 - Tính AP cho từng câu truy vấn trong bộ dữ liệu kiểm thử
 - Sau đó lấy trung bình các giá trị thu được

$$MAP = \frac{1}{|Q|} \cdot \sum \left(\frac{1}{R_q} \cdot \sum P@K_i \right)$$

Ví dụ 5.12. Tính MAP

- Được biết có 10 văn bản phù hợp với câu truy vấn q_1 , và danh sách kết quả trả về có đặc điểm như sau:

RNRNNRNNRR

- Có 8 văn bản phù hợp với câu truy vấn q_2 , và danh sách kết quả trả về có đặc điểm như sau:

NRNNRNRNNN

Yêu cầu: Tính MAP

Ví dụ 5.12. Tính $MAP_{(2)}$

- Được biết có 10 văn bản phù hợp với câu truy vấn q_1 , và danh sách kết quả trả về có đặc điểm như sau:

RNRNNRNNRR

- Có 8 văn bản phù hợp với câu truy vấn q_2 , và danh sách kết quả trả về có đặc điểm như sau:

NRNNRNRNNN

Yêu cầu: Tính MAP

Chúng ta có:

$$AP_1 = (1 + 2/3 + 3/6 + 4/9 + 5/10)/10 = 0.31$$

$$AP_2 = (1/2 + 2/5 + 3/7)/8 = 0.17$$

$$MAP = (AP_1 + AP_2)/2 = (0.31 + 0.17)/2 = 0.24$$

Một số độ đo khác

Độ chính xác-R

- R-Precision
- Đặt Rel là số lượng văn bản phù hợp với câu truy vấn
 - Chúng ta có: $P@Rel = R@Rel = F_1@Rel$
- Độ chính xác trong giới hạn Rel kết quả đầu tiên trong danh sách ($P@Rel$) còn được gọi là Độ chính xác-R
 - Hệ thống hoàn hảo có (Độ chính xác-R) $P@Rel = 1$.

MRR

- Trong một số kịch bản tìm kiếm chúng ta mong muốn nhận được 1 kết quả phù hợp ở vị trí đầu danh sách
 - Ví dụ: Tra cứu các dữ kiện: Ai, Cái gì, Ở đâu, v.v..
- Trong những tình huống đó, MRR - **Mean Reciprocal Rank** (Trung bình của giá trị nghịch đảo vị trí của kết quả phù hợp đầu tiên) có thể là một độ đo phù hợp.
- Đặt K là vị trí của kết quả đầu tiên phù hợp với q. Chúng ta có: $RR(q) = 1/K$
- MRR là trung bình các giá trị RR trên tập truy vấn Q

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

Tính ổn định của các đại lượng

- Với một bộ dữ liệu kiểm thử, thường xảy ra tình huống một hệ thống cho kết quả kém trên một số truy vấn (ví dụ, $AP = 0.1$) nhưng cho kết quả rất tốt trên các truy vấn khác (ví dụ $AP = 0.8$)
- Trong thực tế biên độ dao động giá trị của độ đo cho một hệ thống trên nhiều truy vấn khác nhau thường lớn hơn nhiều biên độ dao động của các hệ thống khác nhau trên cùng câu truy vấn. Hoặc có thể diễn đạt theo cách khác: Các câu truy vấn có độ khó khác nhau đáng kể.

Cần lấy giá trị trung bình của độ đo trên nhiều truy vấn để đảm bảo tính ổn định của kết quả so sánh.

Giá trị phù hợp đa mức

Các mức phù hợp

- Tính phù hợp của văn bản đối với truy vấn được đánh giá theo nhiều mức:
 - Ký hiệu rel_i là mức phù hợp của văn bản d_i ;
 - $rel_i = 0$ nghĩa là d_i không phù hợp; Với 2 văn bản d_i và d_j , nếu $rel_i > rel_j$ thì văn bản d_i (được coi là) phù hợp hơn so với văn bản d_j .
 - Mô phỏng tình huống người dùng ưa thích một kết quả tìm kiếm hơn một kết quả tìm kiếm khác.
- Các đánh giá phù hợp đa mức được sử dụng trong một số nghiên cứu tìm kiếm thông tin
 - Nhưng vẫn ít phổ biến hơn so với đánh giá phù hợp nhị phân

Các độ đo dựa trên giá trị phù hợp đa mức?

NDCG

- Được đo dựa trên các giá trị phù hợp đa mức;
- Được sử dụng trong một số nghiên cứu để đánh giá kết quả tìm kiếm trên Web và đánh giá các phương pháp học xếp hạng;

Giải mã các ký hiệu:

- **NDCG - Normalized Discounted Cumulative Gain.**
 - N - Normalized/Chuẩn hóa;
 - D - Discounted/Cắt giảm;
 - C - Cumulative/Tổng hợp;
 - G - Gain/Lợi ích;

Lợi ích là khái niệm cơ bản của NDCG.

Lợi ích tích lũy

- Nhận xét 1: Lợi ích của một kết quả tìm kiếm tỉ lệ thuận với mức phù hợp của nó
 - Kết quả càng phù hợp thì càng hữu ích với người dùng
 - (G - Gain)
- Nhận xét 2: Lợi ích của danh sách kết quả bằng tổng lợi ích của tất cả các kết quả có trong danh sách; Đặt n là số lượng kết quả trong danh sách. Chúng ta có:

CG (Cumulative Gain) = $r_1 + r_2 + \dots + r_n$, trong đó r_1, r_2, \dots, r_n là mức phù hợp của các văn bản trong danh sách.

Quy luật cắt giảm

- Nhận xét 3: Kết quả càng cách xa vị trí đầu danh sách thì tính hữu ích của nó càng giảm xuống. Chúng ta có DCG của danh sách n kết quả đầu tiên:

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

$$DCG = rel_1 + rel_2/\log_2 2 + \dots rel_n/\log_2 n$$

Có thể có nhiều công thức cắt giảm khác nhau (nhưng vẫn đảm bảo tỉ lệ nghịch với khoảng cách tới đầu danh sách). Ví dụ một công thức cắt giảm khác (tham khảo):

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

Ví dụ 5.13. Tính DCG

- Cho danh sách kết quả gồm 10 văn bản với các giá trị phù hợp được đánh giá theo thang điểm 0-3:

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

Yêu cầu: Tính DCG

Ví dụ 5.13. Tính $DCG_{(2)}$

- Cho danh sách kết quả gồm 10 văn bản với các giá trị phù hợp được đánh giá theo thang điểm 0-3:

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

Yêu cầu: Tính DCG

- $DCG = 3 + (2/\log_2 2 + 3/\log_2 3 + 1/\log_2 6 + 2/\log_2 7 + 2/\log_2 8 + 3/\log_2 9)$
 $= 3 + 2 + 1.89 + 0.39 + 0.71 + 0.67 + 0.96 = 9.61$

(Chúng ta có các giá trị DG:

3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
 $= 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0)$

Chuẩn hóa

- NDCG: là giá trị chuẩn hóa bằng cách chia DCG của danh sách kết quả cho DCG của xếp hạng mẫu với các văn bản như trong danh sách.
 - Xếp hạng mẫu đơn giản là danh sách xếp hạng các kết quả theo thứ tự giảm dần giá trị phù hợp của văn bản;
 - Giá trị chuẩn hóa chỉ phù hợp để so sánh các phương án xếp hạng khác nhau của cùng một tập văn bản;
 - Thường được sử dụng để đánh giá các phương pháp xếp hạng lại một danh sách kết quả được trả về.

**Lưu ý: Giá trị chuẩn hóa có thể cho kết quả bất thường khi so sánh các danh sách xếp hạng của các mô hình khác nhau.*

Ví dụ 5.14. Tính NDCG

Cho 4 văn bản: d_1, d_2, d_3, d_4 với các giá trị phù hợp theo thứ tự là 0, 1, 2, 3, và các kết quả xếp hạng:

i	Xếp hạng 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						

Yêu cầu: Tính NDCG cho các phương án xếp hạng.

Ví dụ 5.14. Tính NDCG₍₂₎

Cho 4 văn bản: d_1, d_2, d_3, d_4 với các giá trị phù hợp theo thứ tự là 0, 1, 2, 3, và các kết quả xếp hạng:

i	Xếp hạng 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	NDCG _{GT} =1.00		NDCG _{RF1} =1.00		NDCG _{RF2} =0.9203	

Yêu cầu: Tính NDCG cho các phương án xếp hạng.

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

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

Nội dung

1. Vấn đề đánh giá hệ thống tìm kiếm
2. Một số bộ dữ liệu kiểm thử tiêu biểu
3. Đánh giá kết quả dạng tập hợp
4. Đánh giá kết quả dạng danh sách
5. Vấn đề xây dựng bộ dữ liệu kiểm thử
6. Chất lượng hệ thống và lợi ích đối với người dùng
7. Trích đoạn kết quả tìm kiếm

Vấn đề đánh giá tính phù hợp

- Như đã biết số lượng cặp văn bản-truy vấn cần đánh giá có thể rất lớn.
- Đồng thời tính phù hợp cũng rất trừu tượng:
 - Nhiều người có thể có nhiều đánh giá khác nhau về tính phù hợp cho cùng 1 cặp văn bản và truy vấn.
- Để hạn chế sự khác biệt trong các kết quả đánh giá được thực hiện bởi nhiều người, các thành viên tham gia đánh giá có thể sử dụng chung 1 định nghĩa cho khái niệm kết quả phù hợp.

Định nghĩa kết quả phù hợp

TREC định nghĩa kết quả 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 bạn đang viết báo cáo về chủ đề như trong câu truy vấn và bạn muốn sử dụng thông tin có 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 so với phần còn lại của văn bản).

Tính thống nhất của các đánh giá

- Kết quả thu được bởi các thành viên có thể được sử dụng để đánh giá kết quả tìm kiếm nếu đảm bảo mức thống nhất lớn hơn một giá trị ngưỡng
- Mức thống nhất giữa các danh sách kết quả đánh giá thường được đo bằng 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
- Kết quả đánh giá được cho là tin cậy (và có thể sử dụng cho kiểm thử) nếu κ trong khoảng $[2/3, 1.0]$, nếu ngược lại (κ quá nhỏ) thì cần kiểm tra lại các đánh giá.

Ví dụ 5.15. Tính hệ số Kappa

Đánh giá 2

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

Thống kê trên các kết quả đánh giá

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

Các giá trị xác suất tổng hợp:

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

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

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

$$P(\text{không phù hợp})^2 + P(\text{phù hợp})^2 = 0.2125^2 + 0.7878^2 = 0.665$$

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

$$(0.925 - 0.665)/(1 - 0.665) = 0.776 \text{ (trong khoảng được chấp nhận)}$$

Hệ quả của sự khác biệt trong đánh giá

- Những kết quả đánh giá được thực hiện bởi nhiều người có thể khác nhau. Như vậy kết quả kiểm thử dựa trên đó có phải là vô nghĩa?
 - Sự khác biệt trong đánh giá có thể ảnh hưởng đáng kể đến các giá trị cụ thể của các độ đo
 - Nhưng có thể không ảnh hưởng tới kết quả so sánh các mô hình (mục đích kiểm thử).
 - => Có thể không ảnh hưởng tới kết quả kiểm thử

Các đánh giá vẫn có thể cho những kết quả kiểm thử tin cậy nếu sự khác biệt không quá lớn.

Đánh giá hệ thống tìm kiếm quy mô lớn

- Rất khó (hoặc có thể nói là không thể) đo độ đầy đủ ở quy mô lớn như Web.
 - => Cần sử dụng đại lượng khác.
- Bên cạnh các độ đo dựa trên tính phù hợp, máy tìm kiếm còn sử dụng các đại lượng khác:
 - Ví dụ 1: Tỷ lệ mở kết quả đầu tiên (Đánh tin cậy với dữ liệu thống kê lớn)
- Ngoài ra phương pháp kiểm thử A/B cũng là phương pháp được sử dụng phổ biến trong các máy tìm kiếm.

Kiểm thử A/B

- Mục đích: Kiểm tra tính hiệu quả của 1 cập nhật.
- Điều kiện thực hiện: Đã có hệ thống tìm kiếm đang hoạt động ở quy mô lớn.
- Cách thực hiện:
 - Điều hướng một tỉ lệ nhỏ lưu lượng (ví dụ, 1%) tới hệ thống mới đã có phần cập nhật.
 - Sau đó thu thập các chỉ số thống kê trên hệ thống hiện có và hệ thống mới.
 - Từ đó có đi tới kết luận liệu cập nhật có đem lại lợi ích hay không?

Có thể là phương pháp kiểm thử được sử dụng phổ biến nhất trong các máy tìm kiếm.

Nội dung

1. Vấn đề đánh giá hệ thống tìm kiếm
2. Một số bộ dữ liệu kiểm thử tiêu biểu
3. Đánh giá kết quả dạng tập hợp
4. Đánh giá kết quả dạng danh sách
5. Vấn đề xây dựng bộ dữ liệu kiểm thử
6. Chất lượng hệ thống và lợi ích đối với người dùng
7. Trích đoạn kết quả tìm kiếm

Các đặc điểm của máy tìm kiếm

- Tốc độ tạo chỉ mục
 - Số lượng văn bản/giờ (kích thước trung bình của văn bản?)
 - Mb/s
- Tốc độ xử lý truy vấn
 - Thời gian người dùng chờ đến lúc nhận được phản hồi
 - Sự phụ thuộc của độ trễ vào kích thước chỉ mục
- Khả năng diễn tả của ngôn ngữ truy vấn
 - Khả năng diễn đạt các nhu cầu thông tin phức tạp
 - Tốc độ xử lý các truy vấn phức tạp
- Tốc độ phản hồi của giao diện
- Miễn phí?

Có thể lượng hóa được các đặc điểm kỹ thuật

Các đặc điểm của máy tìm kiếm₍₂₎

- Đặc điểm chính: Khả năng đáp ứng nhu cầu thông tin của người dùng/Sự hài lòng của người dùng
 - Sự hài lòng là gì?
 - Khả năng đáp ứng nhu cầu thông tin khác với tốc độ phản hồi truy vấn: Nhanh nhưng sai không làm người dùng hài lòng.

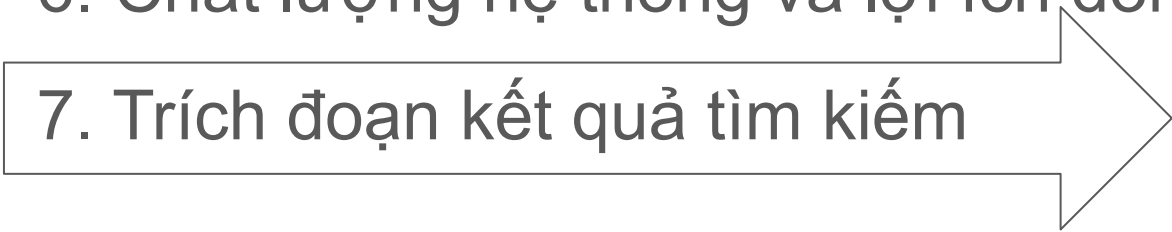
Định lượng sự hài lòng của người dùng bằng cách nào?

Sự hài lòng của người dùng

- Điều gì khiến người dùng hài lòng?
 - Câu trả lời có thể phụ thuộc vào tình huống
- Máy tìm kiếm Web:
 - Người dùng tìm thấy thông tin cần tìm và nhu cầu thông tin được đáp ứng
 - Nếu người dùng hài lòng thì họ sẽ tiếp tục sử dụng máy tìm kiếm
 - => Có thể đo tỉ lệ quay lại của người dùng
- Tìm kiếm trong các trang thương mại điện tử:
 - Người dùng tìm thấy sản phẩm và thực hiện giao dịch
 - Nếu người dùng tìm thấy sản phẩm ưng ý thì họ có thể thực hiện giao dịch.
 - => Có thể đo thời gian giao dịch, hoặc tỉ lệ người tìm kiếm trở thành khách hàng.
- Trong phạm vi công ty
 - Người sử dụng hài lòng khi giảm được thời gian TKTT để hoàn thành công việc.

Tính phù hợp của kết quả tìm kiếm là yếu tố cơ bản khiến người dùng hài lòng.

Nội dung

1. Vấn đề đánh giá hệ thống tìm kiếm
 2. Một số bộ dữ liệu kiểm thử tiêu biểu
 3. Đánh giá kết quả dạng tập hợp
 4. Đánh giá kết quả dạng danh sách
 5. Vấn đề xây dựng bộ dữ liệu kiểm thử
 6. Chất lượng hệ thống và lợi ích đối với người dùng
 7. Trích đoạn kết quả tìm kiếm
- 

Định dạng danh sách kết quả

- Định dạng thông thường: Danh sách "10 liên kết xanh"
 - Nên mô tả mỗi văn bản trong danh sách như thế nào?
 - Các mô tả là rất quan trọng để người dùng xác định kết quả có phù hợp hay không?
- Người dùng thường đánh giá tính phù hợp của kết quả dựa trên mô tả được cung cấp
 - Trước khi quyết định mở xem nội dung văn bản

Cấu trúc kết quả tìm kiếm

- Các thông tin thường được hiển thị cho 1 kết quả:
 - Tiêu đề văn bản,
 - Đường dẫn (url),
 - một số siêu dữ liệu
 - ...và mô tả ngắn gọn còn được gọi là trích đoạn

Vai trò của trích đoạn và các tính trích đoạn văn bản?

Các yêu cầu đối với trích đoạn

- Diện tích trang kết quả tìm kiếm là giới hạn! => Các trích đoạn phải ngắn gọn...
- ... nhưng đồng thời cũng phải đủ nghĩa.
 - Các trích đoạn phải cung cấp đủ thông tin để đánh giá văn bản có thể đáp ứng nhu cầu thông tin hay không
- Lý tưởng:
 - Các trích đoạn hợp lý về mặt ngôn ngữ (và dễ đọc)
 - Cung cấp đủ thông tin sao cho người dùng không cần phải đọc văn bản

Các trích đoạn chiếm tỉ lệ lớn trong trải nghiệm người dùng.

Các trích đoạn

- Hai dạng cơ bản: (i) tĩnh (ii) động
- Một trích đoạn tĩnh của văn bản luôn giống nhau, không phụ thuộc vào truy vấn được cung cấp bởi người dùng
- Trích đoạn động phụ thuộc vào câu truy vấn. Nhằm mục đích giải thích vì sao văn bản được trả về cho câu truy vấn đã cung cấp

Trích đoạn tĩnh

- Trong các hệ thống tìm kiếm thông thường trích đoạn tĩnh là một phần của văn bản
 - Cách đơn giản nhất: Lấy 50 từ đầu tiên của văn bản
- Trong một số trường hợp khác, trích đoạn tĩnh được tạo thành từ một tập các câu chính của văn bản:
 - Tính điểm từng câu,
 - Trích đoạn được tạo thành từ các câu có điểm cao nhất,
 - Cách tiếp cận dựa trên học máy.
- Tham khảo thêm các phương pháp NLP để tổng hợp/sinh trích đoạn văn bản.

Các phương pháp tổng hợp văn bản phức tạp ít được sử dụng trong các hệ thống TKTT

Trích đoạn động

- Sử dụng một hoặc nhiều "cửa sổ" hoặc trích đoạn trong văn bản có chứa các từ truy vấn
 - Ưu tiên các trích đoạn trong đó các từ truy vấn xuất hiện gần nhau (tương tự truy vấn nguyên câu).
 - Ưu tiên các trích đoạn trong đó từ truy vấn xuất hiện trong một cửa sổ nhỏ.
- Theo cách này trích đoạn chứa toàn bộ nội dung của cửa sổ bao gồm cả từ truy vấn và từ không có trong truy vấn.

Ví dụ 5.16. Trích đoạn động

Truy vấn: “new guinea economic development”

Trích đoạn (in đậm) được trích xuất từ một văn bản: . . . **In recent years, Papua New Guinea has faced severe economic difficulties and** economic growth has slowed, partly as a result of weak governance and civil war, and partly as a result of external factors such as the Bougainville civil war which led to the closure in 1989 of the Panguna mine (at that time the most important foreign exchange earner and contributor to Government finances), the Asian financial crisis, a decline in the prices of gold and copper, and a fall in the production of oil. **PNG’s economic development record over the past few years is evidence that** governance issues underly many of the country’s problems. Good governance, which may be defined as the transparent and accountable management of human, natural, economic and financial resources for the purposes of equitable and sustainable development, flows from proper public sector management, efficient fiscal and accounting mechanisms, and a willingness. . . .

Các nội dung văn bản

- Chúng ta cần lưu các nội dung văn bản để có thể sinh trích đoạn động
 - Không thể tạo trích đoạn động từ chỉ mục ngược có vị trí (vì tốn rất nhiều thời gian)
- Các nội dung văn bản có thể thay đổi theo thời gian
 - Cần cập nhật thường xuyên.
- Các văn bản rất dài có thể chiếm nhiều dung lượng và tốn nhiều thời gian xử lý
 - Thường chỉ lưu một đoạn đầu tiên trong giới hạn kích thước phù hợp.

Bài tập 5.1. Đánh giá tập kết quả

Tính độ chính xác, độ đầy đủ, và F_1 cho tập kết quả có các giá trị thống kê theo bảng nhầm lẫn như sau:

	phù hợp	không phù hợp
trả về	10	20
không trả về	80	1,000,000,000

Bài tập 5.2. Giá trị nội suy

Giải thích ý nghĩa của giá trị nội suy của độ chính xác ở mức độ đầy đủ = 0

Bài tập 5.3. Điểm cân bằng

Trong danh sách kết quả trả về, chúng ta gọi điểm cân bằng là điểm có P và R bằng nhau (nếu $P@i = R@i$ thì i được gọi là điểm cân bằng).

Có luôn tồn tại điểm cân bằng hay không? Hãy chứng minh vì sao luôn tồn tại hoặc lấy ví dụ một danh sách kết quả tìm kiếm không tồn tại điểm cân bằng.

Bài tập 5.4. MRR

Được biết các danh sách kết quả tìm kiếm cho các truy vấn và giải thuật có đặc điểm phù hợp như sau:

	GT1	GT2
q_1	NRNNN	NNNNR
q_2	NNRNN	RNNNN

Hãy so sánh các giải thuật theo MRR

Bài tập 5.5. Đánh giá tính phù hợp

Được biết hệ thống tìm kiếm trả về tập kết quả là $\{4, 5, 6, 7, 8\}$, và các kết quả đánh giá tính phù hợp trong dữ liệu kiểm thử như trong hình vẽ:

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

b) Tính P , R và F_1 trong trường hợp văn bản được coi là phù hợp nếu cả hai cùng đánh giá là phù hợp;

c) Tính P , R và F_1 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:

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

d1) $\text{kappa} = -1$; d2) $\text{kappa} = 1$;

