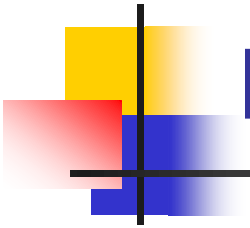
IT4853
Tìm kiếm và trình diễn thông tin

Bài 7. Đánh giá kết quả tìm kiếm

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

- Vấn đề đánh giá kết quả tìm kiếm
- Độ chính xác, độ đầy đủ
- Độ đo F
- Đường cong P/R
  - Vẽ đường cong P/R
  - Chuẩn hóa đường cong P/R
  - So sánh đường cong P/R và ROC
- Các giá trị trung bình

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 bằng con đường thực nghiệm, vì vậy đánh giá kết quả là một khâu rất quan trọng trong nghiên cứu các mô hình tìm kiếm.

Tiêu trí chất lượng

- Khả năng đáp ứng nhu cầu thông tin của người dùng là tiêu trí chất lượng cơ bản của công cụ tìm kiếm :
  - Người dùng sẽ hài lòng nếu công cụ tìm kiếm trả về văn bản phù hợp, đáp ứng được nhu cầu thông tin;

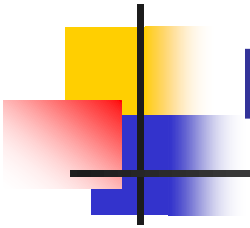
Tín hiệu hài lòng của người dùng

- Công cụ tìm kiếm trên Web
  - Người dùng hài lòng nếu tìm thấy thông tin cần thiết. Đo: Tỉ lệ quay trở lại công cụ tìm kiếm.
  - Nhà quảng cáo hài lòng nếu người tìm kiếm mở quảng cáo. Đo: Tỉ lệ mở quảng cáo.
- Thương mại điện tử
  - Khách hàng được cho là hài lòng nếu mua sản phẩm. Đo: Tỉ lệ người mua hàng;
  - Người bán hài lòng nếu bán được sản phẩm. Đo: Lợi nhuận trên sản phẩm bán được.
- Công ty
  - CEO hài lòng nếu nhân viên làm việc năng suất hơn nhờ áp dụng công cụ tìm kiếm. Đo: Mức tăng lợi nhuận của công ty.

Phụ thuộc vào từng tình huống sử dụng cụ thể

Truy vấn vs. nhu cầu thông tin

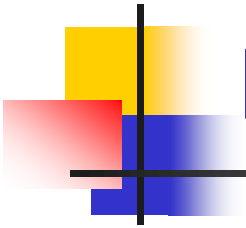
- Phù hợp với truy vấn chưa chắc đã đáp ứng được nhu cầu thông tin.
- Ví dụ, nhu cầu thông tin *i* : "Liệu rượu vang có tác dụng làm giảm nguy cơ mắc bệnh tim hay không? Nếu có thì vang đỏ có tốt hơn vang trắng không?"
- Truy vấn *q*: [vang đỏ vang trắng tim]
- Xét văn bản *d*: *Bài diễn thuyết từ trái tim của anh ấy là một đòn tấn công trực diện hướng vào những công ty sản xuất rượu vang nhằm làm giảm ảnh hưởng của vang trắng và đỏ đến vấn nạn lái xe trong tình trạng say xỉn*.
- *d* rất khớp với truy vấn *q* . . . nhưng không phù hợp với nhu cầu thông tin *i* .



Nhu cầu thông tin vs. truy vấn

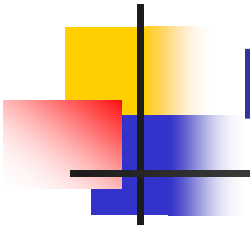
- Con người đánh giá sự phù hợp với nhu cầu thông tin.
- Giải thuật tìm kiếm đánh giá sự phù hợp với truy vấn.

Có thể sử dụng đánh giá của con người làm chuẩn mực để đánh giá giải thuật tìm kiếm.



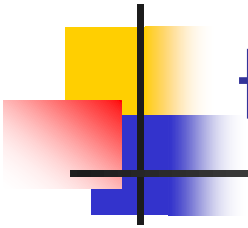
Dữ liệu kiểm thử

- Dữ liệu để đánh giá kết quả tìm kiếm gồm:
  - Bộ văn bản được lựa chọn kỹ lưỡng;
  - Tập truy vấn mẫu;
  - Đánh giá phù hợp cho tất cả các cặp truy vấn – văn bản.



Nội dung chính

- Vấn đề đánh giá kết quả tìm kiếm
- Độ chính xác, độ đầy đủ
- Độ đo F
- Đường cong P/R
  - Vẽ đường cong P/R
  - Chuẩn hóa đường cong P/R
  - So sánh đường cong P/R và ROC
- Các giá trị trung bình



Độ chính xác và độ đầy đủ

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

  Precision = #(văn bản phù hợp trả về)/#(văn bản trả về)

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

  Recall = #(văn bản phù hợp trả về)/#(văn bản phù hợp)

Ký hiệu P: độ chính xác; R: độ đầy đủ.

Ví dụ P/R

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

P = ?

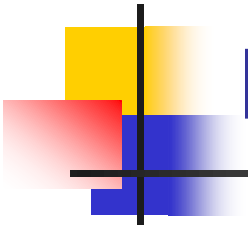
R = ?

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	

Kết hợp độ chính xác và độ đầy đủ

- Đánh giá chỉ dựa trên độ chính xác hoặc chỉ dựa trên độ đầy đủ bộc lộ nhiều điểm hạn chế:
  - Có thể tăng độ đầy đủ bằng cách trả về nhiều văn bản hơn, độ đầy đủ luôn đạt 100% nếu trả về tất cả văn bản;
  - Ngược lại, thường dễ đạt được độ chính xác cao khi trả về ít văn bản (độ đầy đủ thấp).
- Xét đến những đối tượng người dùng khác nhau
  - Một người tìm kiếm trên Web thường chỉ xem khoảng 20 văn bản đầu tiên => độ chính xác quan trọng hơn;
  - Một nhà nghiên cứu lại muốn nhận được tất cả văn bản liên quan đến chủ để được quan tâm => độ đầy đủ quan trọng hơn.

Cần sử dụng đồng thời độ chính xác và độ đầy đủ để đánh giá kết quả tìm kiếm.



Nội dung chính

- Vấn đề đánh giá kết quả tìm kiếm
- Độ chính xác, độ đầy đủ
- Độ đo F
- Đường cong P/R
  - Vẽ đường cong P/R
  - Chuẩn hóa đường cong P/R
  - So sánh đường cong P/R và ROC
- Các giá trị trung bình

Độ đo F

- Độ đo F kết hợp độ chính xác và độ đầy đủ thành một đại lượng duy nhất:

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

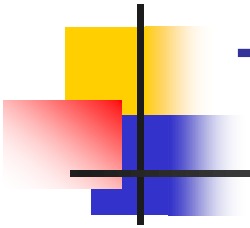
Trong đó $\beta^2 = \frac{1-\alpha}{\alpha}$

- $\alpha \in [0, 1]$, $\beta^2 \in [0, \infty]$
- Miền giá trị nào của β đề cao độ đầy đủ hơn độ chính xác?

Độ đo F

- Nếu $\beta = 1$ hoặc $\alpha = 0.5$, thì F là trung bình điều hòa của *P* và *R*;
- Nếu $\beta = 0$, F là độ chính xác;
- Nếu $\beta = \text{Inf}$, F là độ đầy đủ.

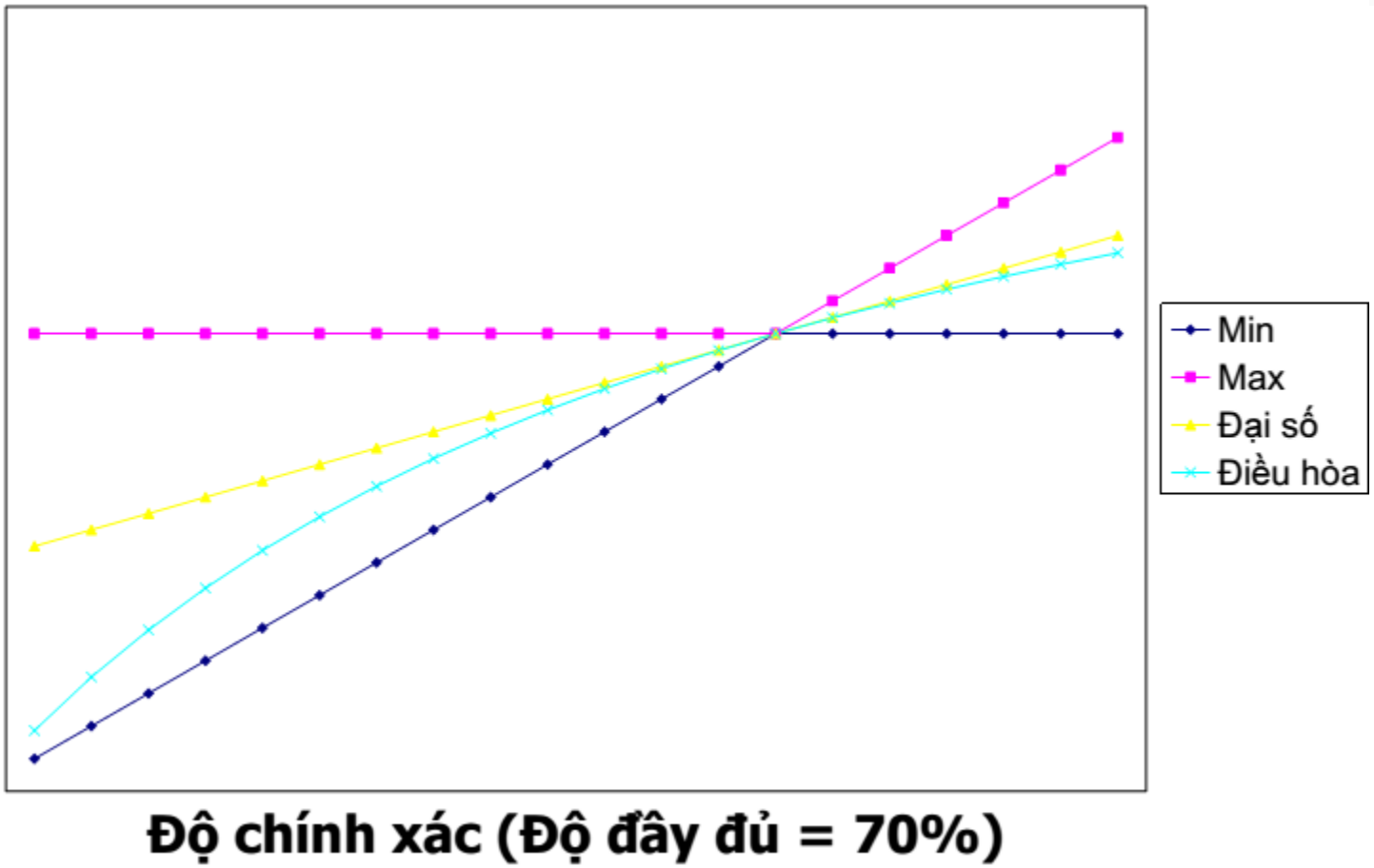
Ký hiệu độ đo F với $\beta = 1$ là F_1



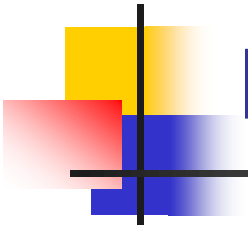
Trung bình điều hòa

- Độ đo F là trung bình điều hòa của P và R
- Vì sao không tổng hợp P và R theo cách khác?
  - Ví dụ, trung bình đại số?
- Mong muốn: Phạt những kết quả có độ chính xác hoặc độ đầy đủ thấp.
  - Lấy giá trị cực tiểu giúp ta đạt được mục đích này.
  - Vì sao không sử dụng giá trị cực tiểu?
- V.V.

So sánh các phương pháp tổng hợp P và R

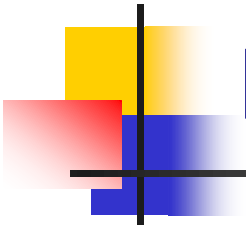


F là trung bình điều hòa của P và R.



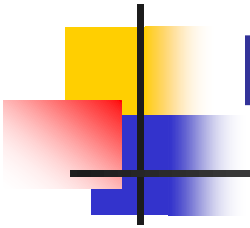
Nội dung chính

- Vấn đề đánh giá kết quả tìm kiếm
- Độ chính xác, độ đầy đủ
- Độ đo F
- Đường cong P/R
  - Vẽ đường cong P/R
  - Chuẩn hóa đường cong P/R
  - So sánh đường cong P/R và ROC
- Các giá trị trung bình



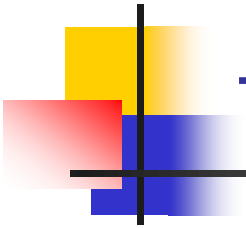
P@i và R@i

- Độ chính xác, độ đầy đủ là những độ đo được thiết kế cho tìm kiếm không xếp hạng.
- Tuy nhiên chúng ta có thể mở rộng những độ đo này cho danh sách xếp hạng.



P@i và R@i

- P@i = #(văn bản phù hợp trong i kết quả đầu tiên)/i
- R@i = #(văn bản phù hợp trong i kết quả đầu tiên)/
  #(số 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



Đường cong P/R

- Cách vẽ đường cong P/R:
  - Cho i biến thiên từ 1 đến hết danh sách kết quả tìm kiếm;
  - 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 mối liên hệ phụ thuộc giữa độ chính xác và độ đầy đủ.

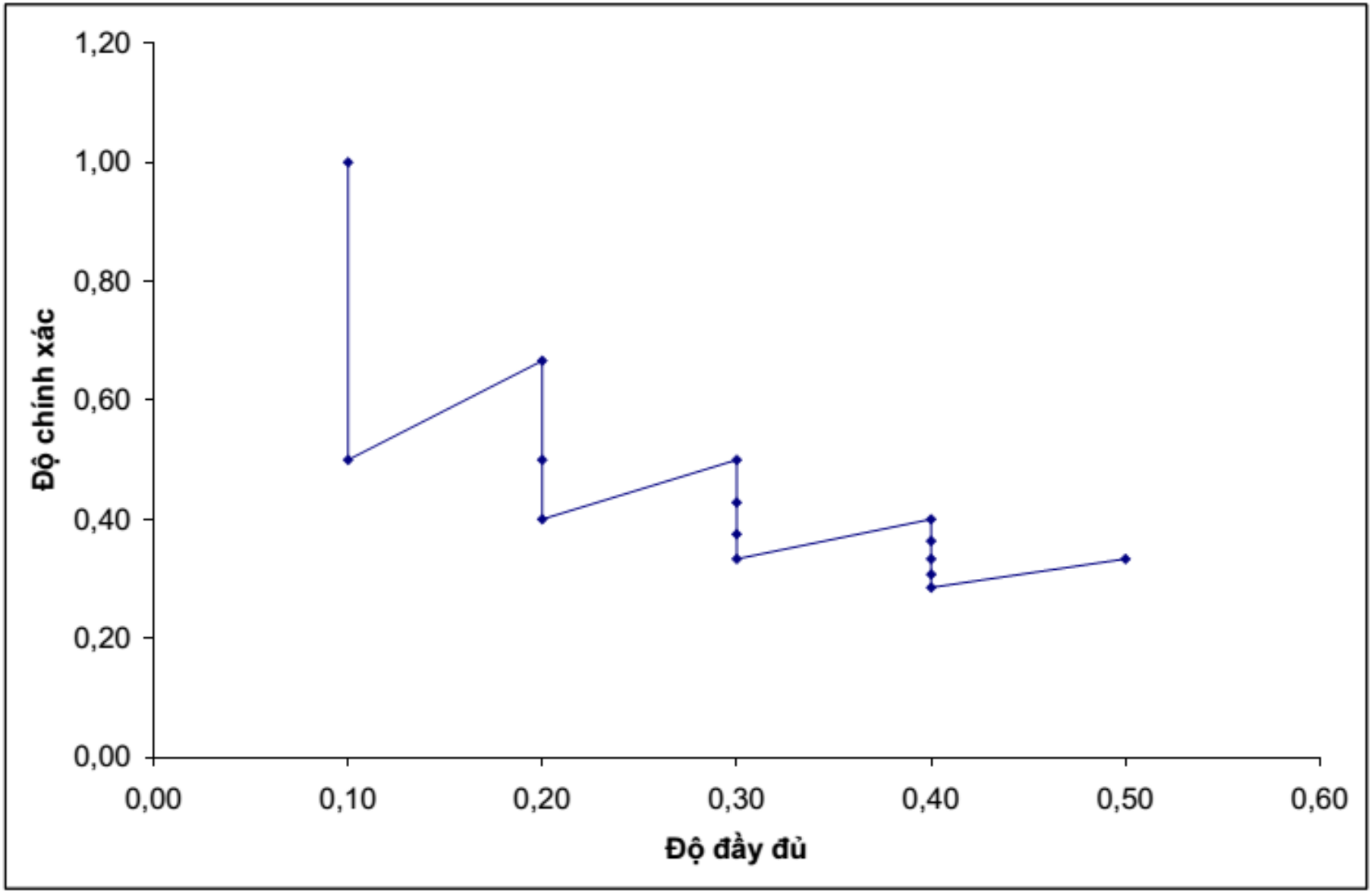
Ví dụ vẽ đường cong P/R

Tập kết quả phù hợp: 10 văn bản

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

Tập kết quả

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

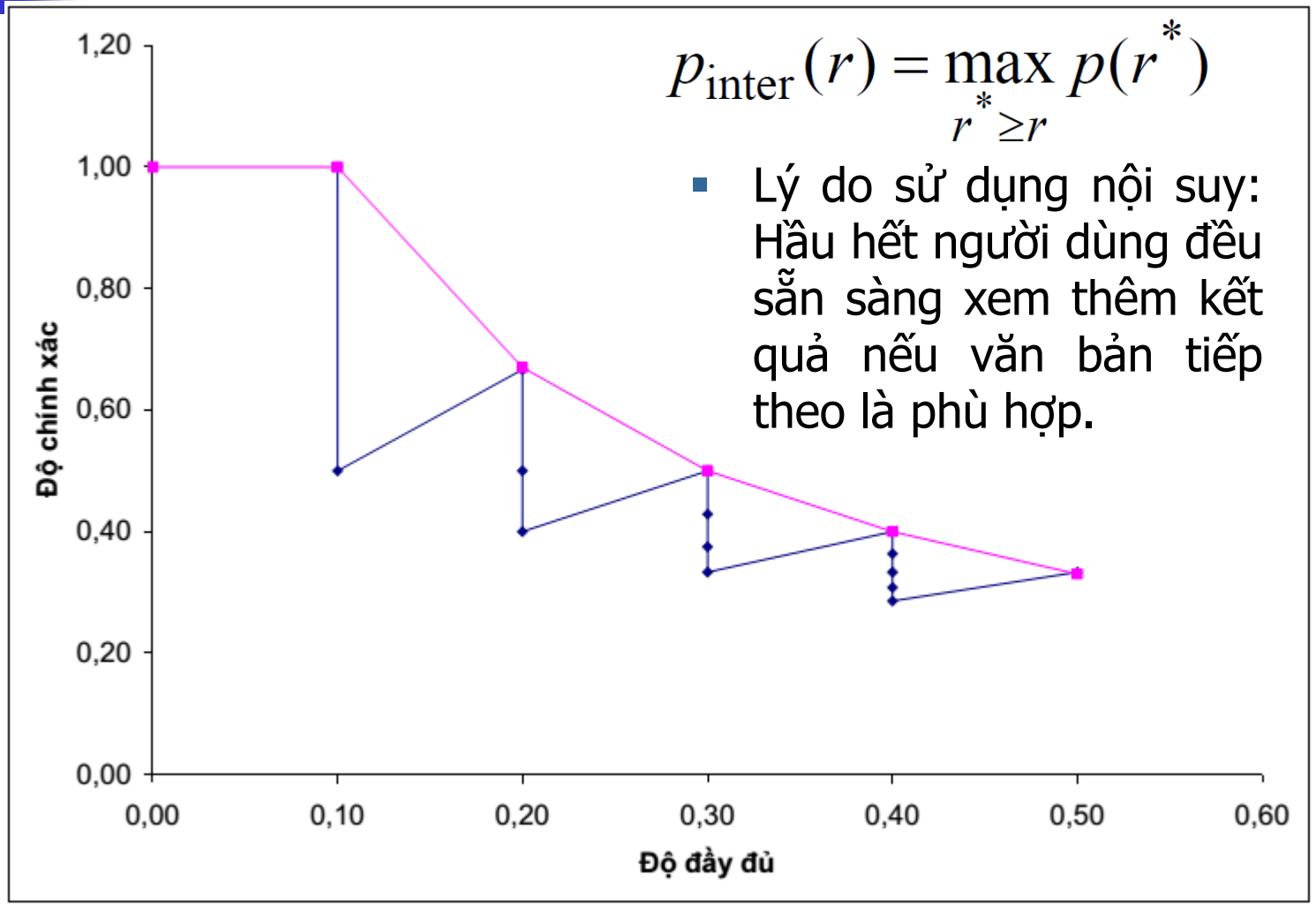


- Mỗi điểm trên đồ thị ứng với độ chính xác/đầy đủ cho k văn bản đầu tiên (k = 1, 2, 3, 4, . . .).

Nội dung chính

- Vấn đề đánh giá kết quả tìm kiếm
- Độ chính xác, độ đầy đủ
- Độ đo F
- Đường cong P/R
  - Vẽ đường cong P/R
  - Chuẩn hóa đường cong P/R
  - So sánh đường cong P/R và ROC
- Các giá trị trung bình

Độ chính xác nội suy



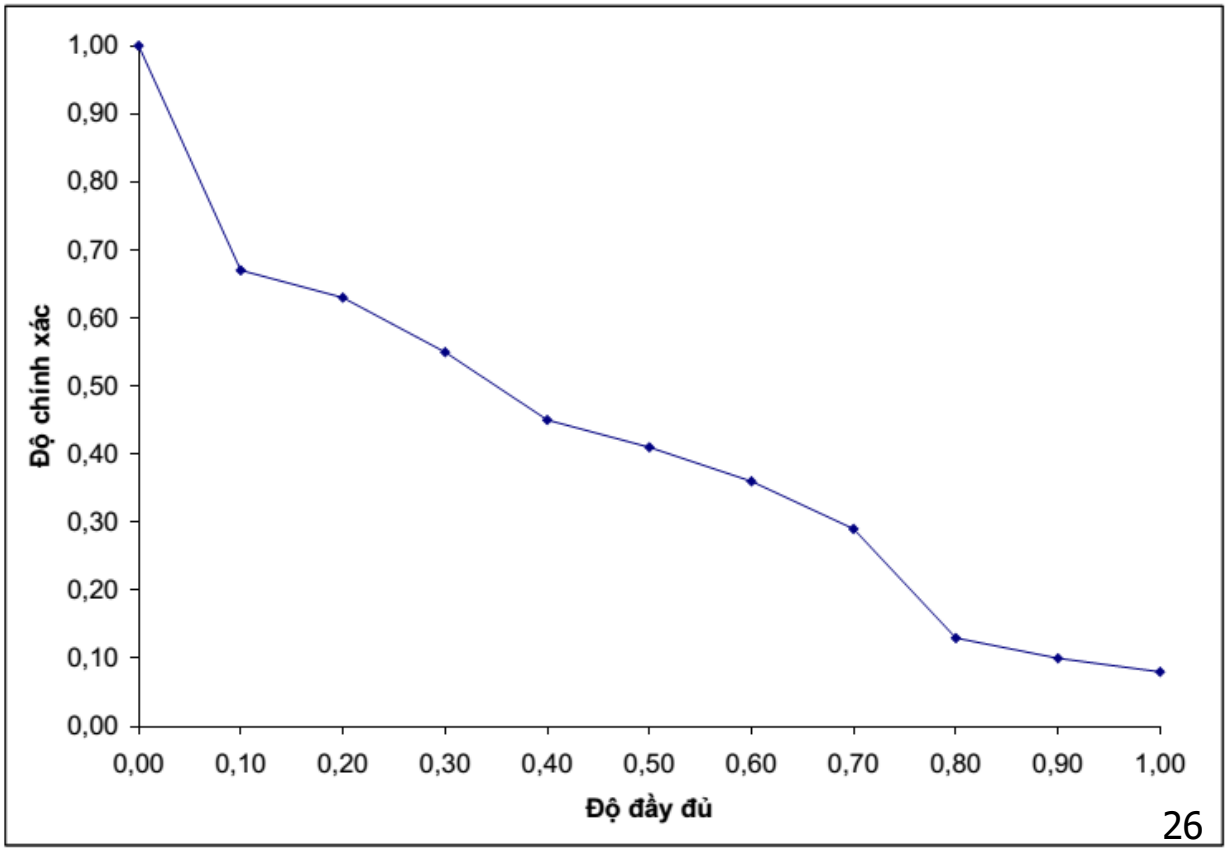
$$p_{\text{inter}}(r) = \max_{r^* \geq r} p(r^*)$$

- Lý do sử dụng nội suy: Hầu hết người dùng đều sẵn sàng xem thêm kết quả nếu văn bản tiếp theo là phù hợp.

Những giá trị chuẩn của độ đầy đủ

Độ đầy đủ	Độ chính xác nội suy
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

trung bình ≈ 0.425



Nội dung chính

- Vấn đề đánh giá kết quả tìm kiếm
- Độ chính xác, độ đầy đủ
- Độ đo F
- Đường cong P/R
  - Vẽ đường cong P/R
  - Chuẩn hóa đường cong P/R
  - So sánh đường cong P/R và ROC
- Các giá trị trung bình

Bảng nhầm lẫn cho tập kết quả

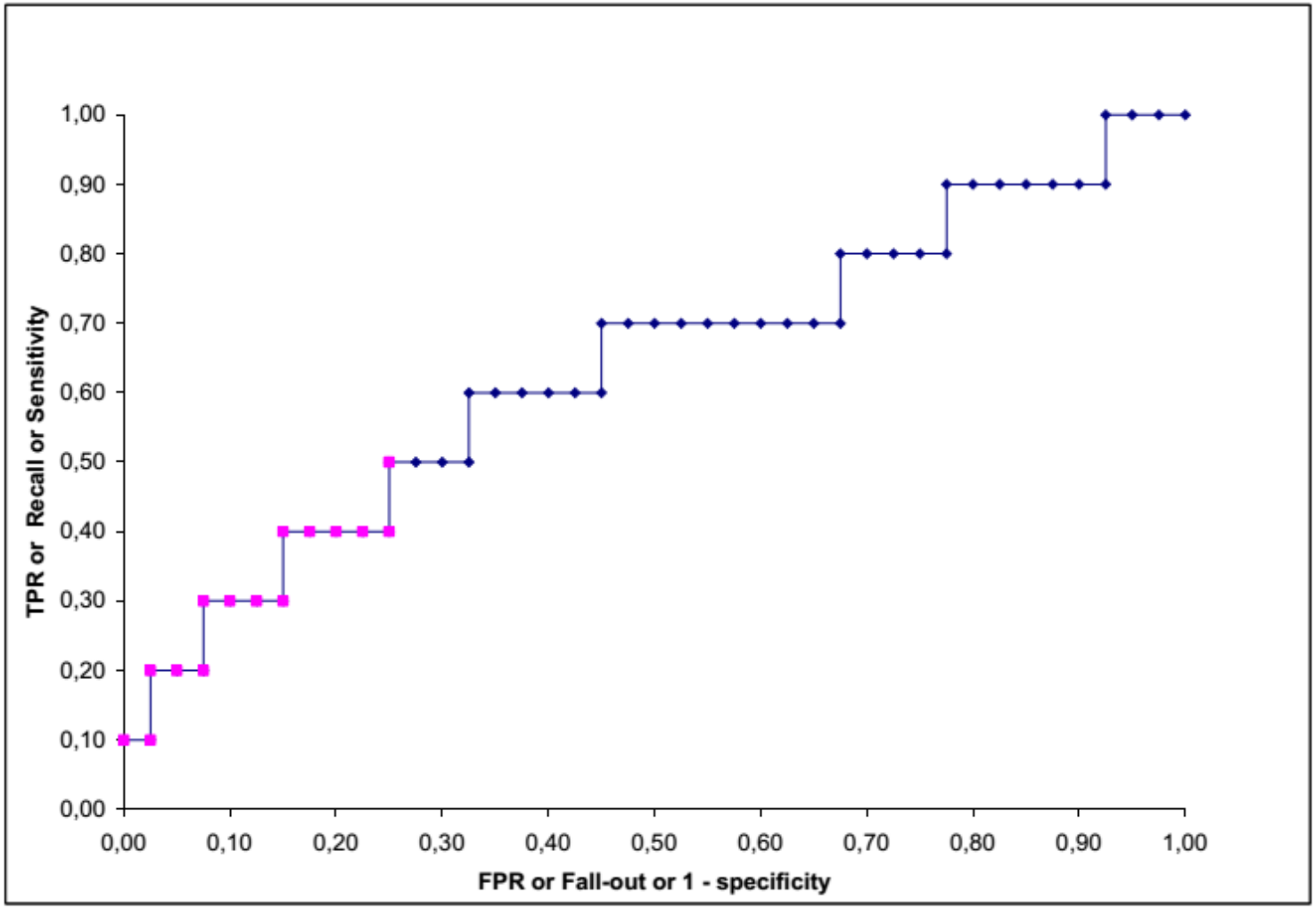
	Phù hợp	Không phù hợp
Trả về	TP	FP
Không trả về	FN	TN

- Văn bản được xác định phù hợp/không phù hợp dựa trên bộ dữ liệu kiểm thử;
- Hệ thống quyết định trả về/không trả về;
- Các giá trị trong bảng nhầm lẫn lần lượt là # văn bản:
  - TP: được trả về và phù hợp;
  - FP: được trả về và không phù hợp;
  - FN: không được trả về và phù hợp;
  - TN: không trả về và không phù hợp;

Đường cong ROC và P/R

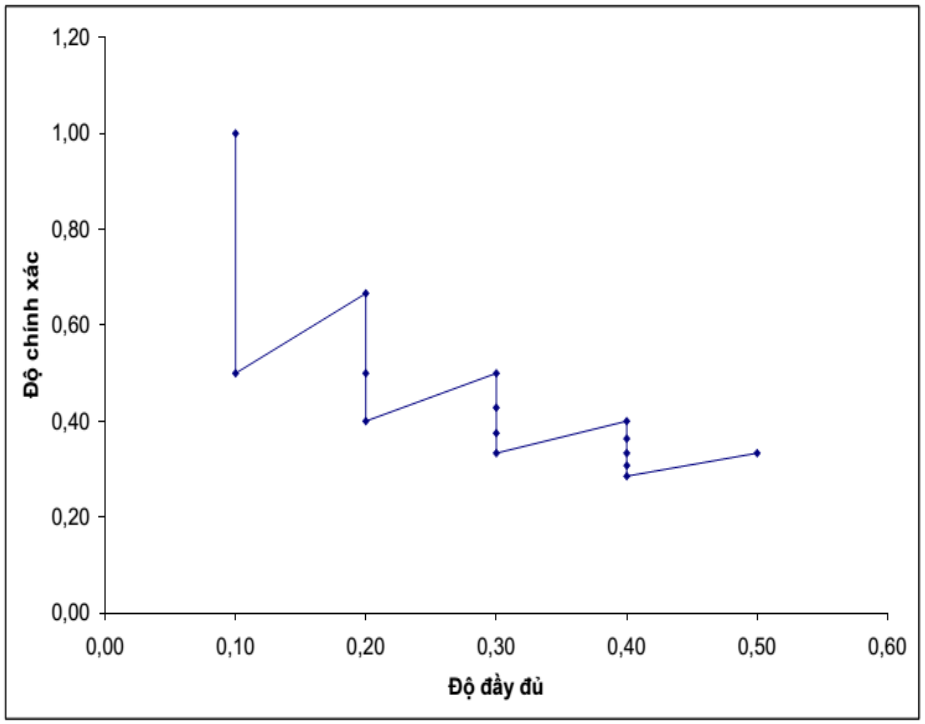
- ROC: TPR/FPR
- PR: Precision/Recall
- TPR = Recall = TP/(TP+FN) = p(trả về|phù hợp)
- FPR = Fall-out = FP/(FP+TN) = p(trả về|không phù hợp)
- Precision = TP/(TP+FP) = p(phù hợp|trả về)
- Fall-out = FP/(FP+TN) = p(trả về|không phù hợp)
- Các đại lượng khác:
  - Specificity = TN/(FP+TN) = p(không trả về|không phù hợp)
  - Sensitivity = TP/(TP+FN) = p(trả về|phù hợp)
- FPR = Fall-out = 1 – specificity
- TPR = Recall = Sensitivity

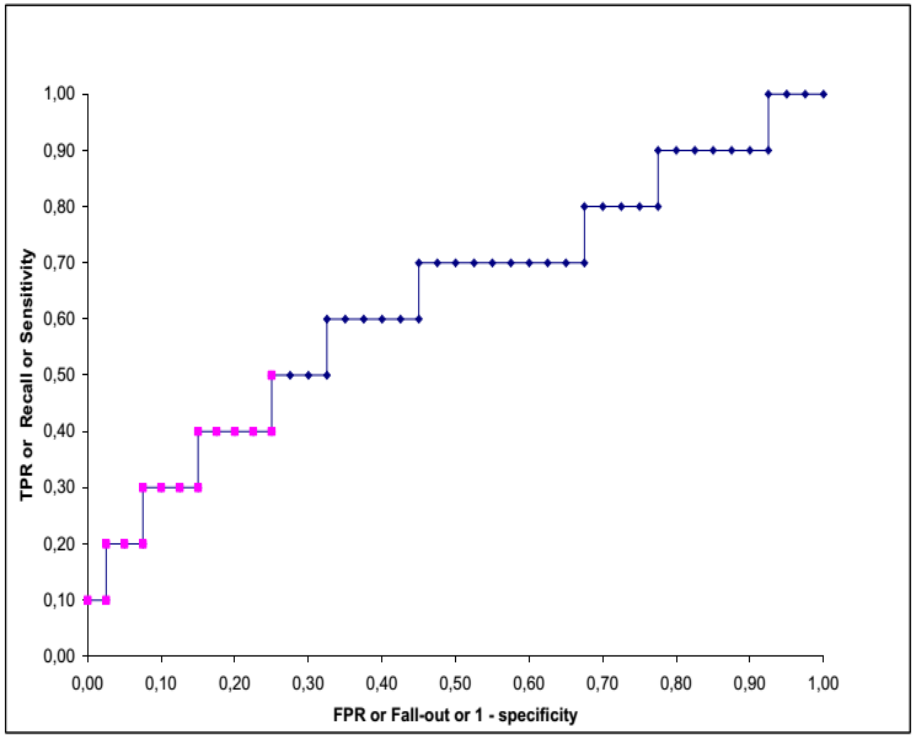
Đường cong ROC

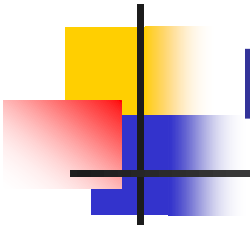


- Tương đương đồ thị độ chính xác/độ đầy đủ trong so sánh các thuật toán
- Thường chỉ quan tâm tới một khoảng nhỏ ở góc thấp bên trái ứng với độ nhạy (sensitivity) < 0.4.

Đường cong P/R và ROC







Nội dung chính

- Vấn đề đánh giá kết quả tìm kiếm
- Độ chính xác, độ đầy đủ
- Độ đo F
- Đường cong P/R
  - Vẽ đường cong P/R
  - Chuẩn hóa đường cong P/R
  - So sánh đường cong P/R và ROC
- Các giá trị trung bình

Độ chính xác trung bình

- Ký hiệu vị trí của những văn bản phù hợp trong danh sách kết quả là:
  - K_1, K_2, … K_R
- Độ chính xác trung bình:

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

- Ví dụ: d1*, d2, d3*, d4, d5* và R= 3

$$AP = \frac{1}{3}\cdot\left(\frac{1}{1}+\frac{2}{3}+\frac{3}{5}\right)\approx 0.76$$

Độ chính xác trung bình: AP: Average Precision

Bình quân độ chính xác trung bình

- Mean Average Precision (MAP)

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

Trong đó R_i là số văn bản trong bộ dữ liệu phù hợp với truy vấn q_i.

Ví dụ MAP

***** Văn bản phù hợp với truy vấn 1

Xếp hạng # 1	*	–	*	–	–	*	–	–	*	*
Độ đầy đủ	0,2	0,2	0,4	0,4	0,4	0,6	0,6	0,6	0,8	1,0
Độ chính xác	1,0	0,5	0,67	0,5	0,4	0,5	0,43	0,38	0,44	0,5

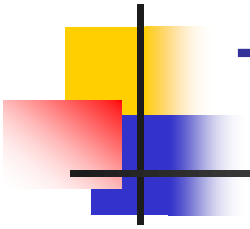
*** Văn bản phù hợp với truy vấn 2

Xếp hạng # 2	–	*	–	–	*	–	*	–	–	–
Độ đầy đủ	0,0	0,33	0,33	0,33	0,67	0,67	1,0	1,0	1,0	1,0
Độ chính xác	0,0	0,5	0,33	0,25	0,4	0,33	0,43	0,38	0,33	0,3

$AP_1 = (1,0 + 0,67 + 0,5 + 0,44 + 0,5)/5 = 0,62$

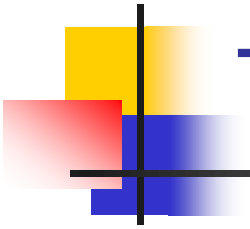
$AP_2 = (0,5 + 0,4 + 0,43)/3 = 0,44$

$MAP = (0,62 + 0,44)/2 = 0,53$



Tính ổn định của độ đo

- Trên một bộ dữ liệu kiểm thử hệ thống có thể trả về kết quả kém chất lượng với một số truy vấn nhưng lại trả về kết quả rất tốt với những truy vấn khác.
- Biên độ giao động của độ đo đối với một hệ thống trên những truy vấn khác nhau có thể lớn hơn nhiều so với những hệ thống khác nhau trên cùng truy vấn.
  - Truy vấn có độ khó khác nhau.



Tính ổn định của độ đo

- Lấy trung bình trên tất cả truy vấn làm tăng tính ổn định của độ đo
- Cần nhiều truy vấn hơn cho các độ đo kém ổn định
  - AP: 25 (đủ), 50 (tốt)
  - P@10: 150 – 200 (tốt)

[Modern Information Retrieval]

Bài tập 7.1

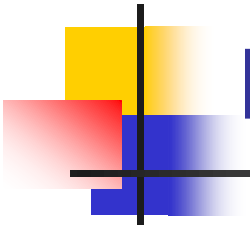
Tính độ chính xác, độ đầy đủ và F_1 cho tập kết quả sau:

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

Bài tập 7.2

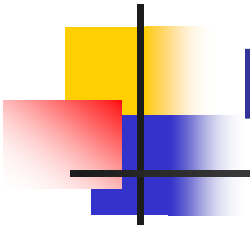
Công cụ tìm kiếm Snoogle luôn trả lời "tìm thấy 0 kết quả thỏa mãn", cho truy vấn bất kỳ. Vì sao Snoogle thể hiện rằng độ chính xác khái quát không hữu ích trong đánh giá kết quả tìm kiếm?

$$AC = (TP+TN)/(TP+TN+FP+FN)$$



Bài tập 7.3

Miền giá trị của độ chính xác nội suy ở mức độ đầy đủ = 0 là gì?



Bài tập 7.4

Có luôn tồn tại điểm cân bằng (break-even point) giữa P và R? Nếu luôn tồn tại: hãy chứng minh, hoặc lấy một ví dụ phủ định.

Điểm cân bằng là điểm i sao cho:

$$P@i = R@i$$