
ការប្រើប្រាស់សម្រាប់បង្កើត / រចនា / ផ្តល់ព័ត៌មាន

Chapter 03

Evaluation of Information Retrieval Systems

Evaluation

- Why Evaluate?
- What to Evaluate?
- How to Evaluate?

Why System Evaluation?

-
- There are many retrieval models/
algorithms/ systems, which one is the best?
 - What is the best component for:
 - Ranking function (dot-product, cosine, ...)
 - Term selection (stopword removal, stemming...)
 - Term weighting (TF, TF-IDF,...)

What to Evaluate?

- How much of the information need is satisfied.
- How much was learned about a topic.
- Incidental learning:
 - How much was learned about the collection.
 - How much was learned about other topics.
- How inviting the system is.

Relevance

សំពីអ្នកដើរ និង ការស្វែងរកទូទៅ កំណត់ផែនដំ

- In what ways can a document be relevant to a query?
 - Answer precise question precisely. ចរចាប់បាន
 - Partially answer question.
 - Suggest a source for more information. ឈរអំពី
 - Give background information.
 - Remind the user of other knowledge.
 - Others ...

Which is the Best Rank Order?

តារាង ដែលបាន

8 algorithms a-h

a

R
R
R
R
R
R
R
R

b

R
R
R
R
R
R
R
R

c

R
R
R
R
R
R
R
R

d

R
R
R
R
R
R
R
R

e

R
R
R
R
R
R
R
R

f

R
R
R
R
R
R
R
R

g

R
R
R
R
R
R
R
R

h

R
R
R
R
R
R
R
R

Relevance

- How **relevant** is the document
 - for this user for this information need.
- Subjective, but
- Measurable to some extent
 - How often do people agree a document is relevant to a query
- How well does it answer the question?
 - Complete answer? Partial?
 - Background Information?
 - Hints for further exploration?

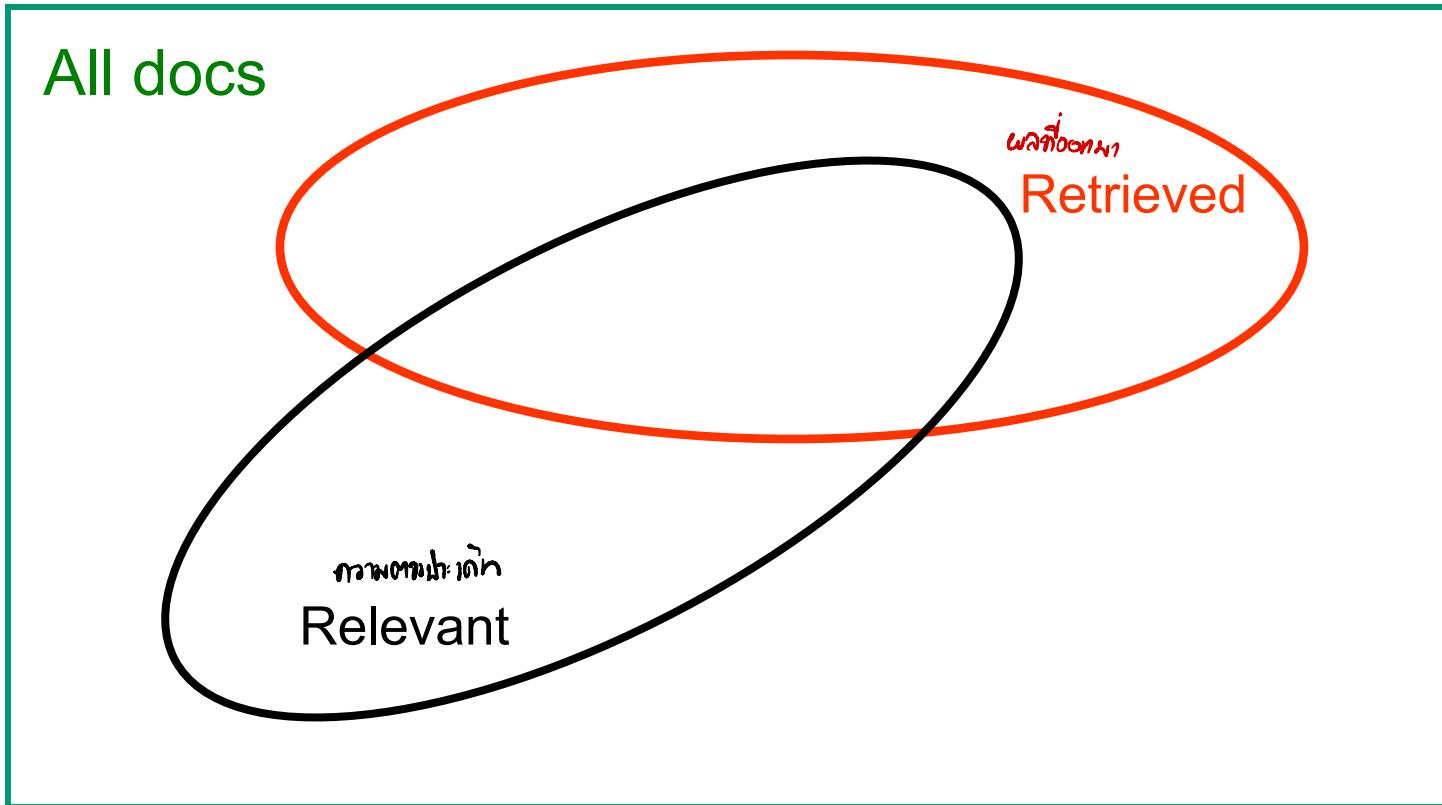
What to Evaluate?

What can be measured that reflects users' ability to use system? (Cleverdon 66)

- Coverage of Information ครอบคลุมปัจจัย
- Form of Presentation รูปแบบแสดงผล
- Effort required/Ease of Use ต้องการใช้แรงงานน้อย
- Time and Space Efficiency ใช้เวลาและพื้นที่น้อย
- Recall $\frac{\text{จำนวนเอกสารที่ถูกเรียกคืน}}{\text{จำนวนเอกสารที่มี}} \times 100\%$ คือจำนวนเอกสารที่ถูกเรียกคืนในจำนวนเอกสารที่มีทั้งหมด
- proportion of relevant material actually retrieved จำนวนเอกสารที่ถูกเรียกคืนที่เป็นสาระจริงๆ
- Precision $\frac{\text{จำนวนเอกสารที่ถูกเรียกคืน}}{\text{จำนวนเอกสารที่ถูกเรียกคืนทั้งหมด}} \times 100\%$ (จำนวนเอกสารที่ถูกเรียกคืนที่เป็นสาระจริงๆ ในจำนวนเอกสารที่ถูกเรียกคืนทั้งหมด)
- proportion of retrieved material actually relevant จำนวนเอกสารที่ถูกเรียกคืนที่เป็นสาระจริงๆ ในจำนวนเอกสารที่ถูกเรียกคืนทั้งหมด

effectiveness

Relevant vs. Retrieved



Precision vs. Recall

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

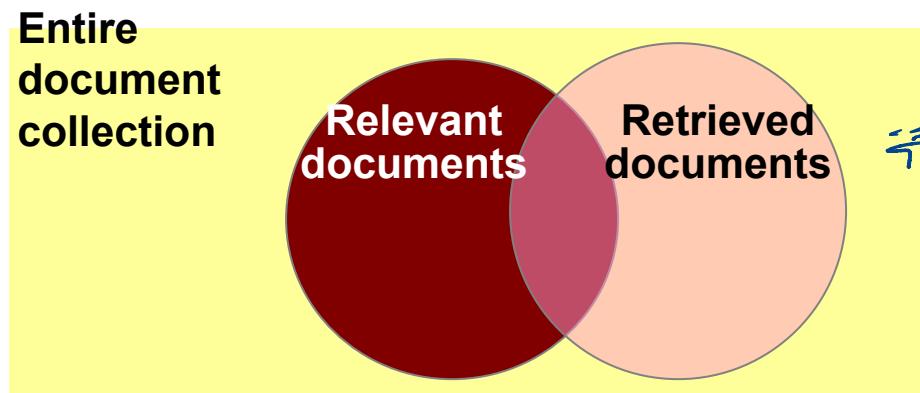
$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$

All docs

Relevant

Retrieved

Precision and Recall



	retrieved	not retrieved
relevant	<p>retrieved & relevant</p> <p>ପାଇଁ ପାଇଁ</p>	not retrieved but relevant
irrelevant	<p>retrieved & irrelevant</p> <p>ମା ହୋଇ ନାହିଁ</p>	Not retrieved & irrelevant
retrieved		not retrieved

$$recall = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

จำนวนหนังสือที่ถูกค้นหามา retrieved no
จำนวนหนังสือที่ถูกค้นหามา ค้นหามา

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

จำนวนหนังสือที่ถูกต้อง^{จำนวนหนังสือที่ถูกต้อง} จำนวนหนังสือที่ได้^{จำนวนหนังสือที่ได้}

Another common representation

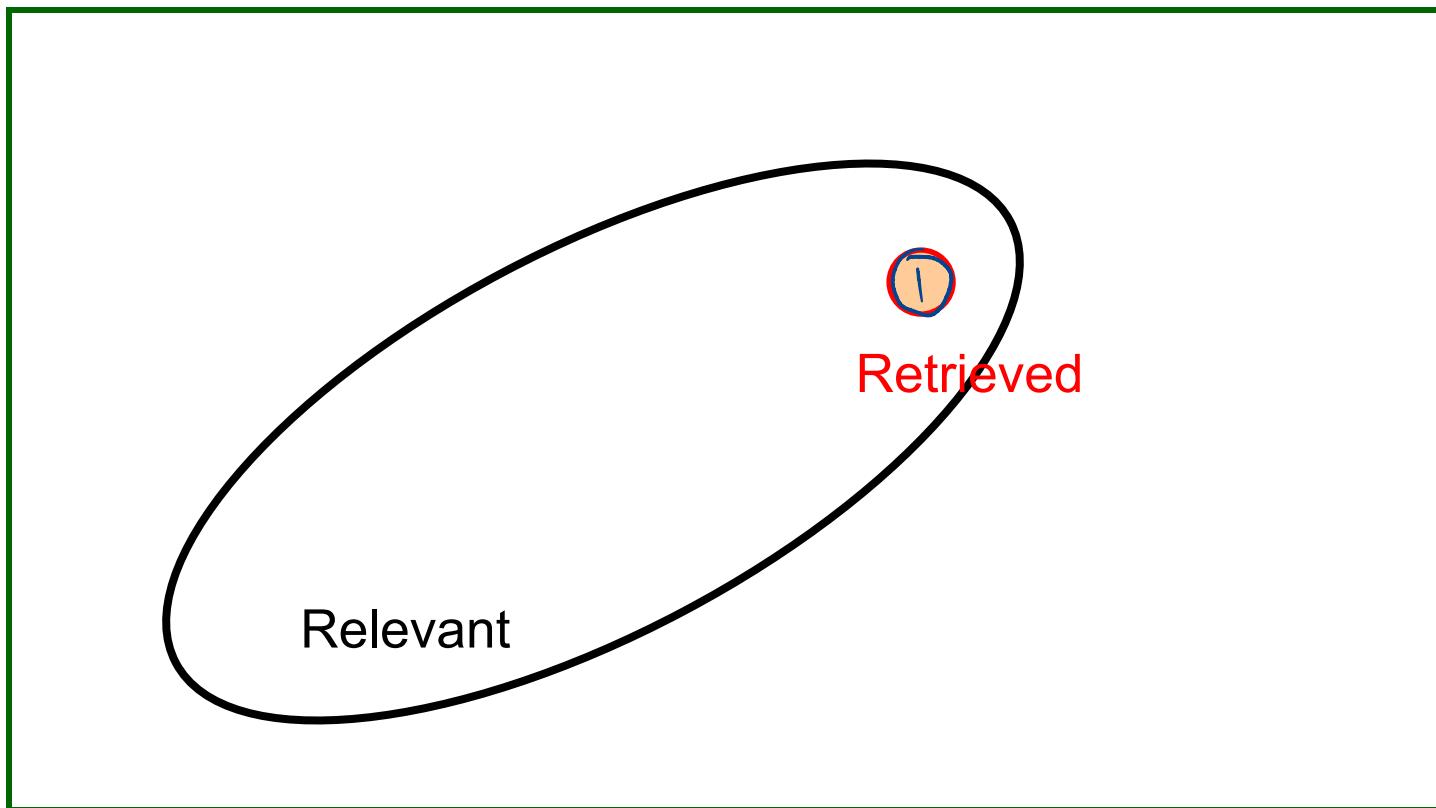
		relevant	irrelevant
Retrieved	relevant	A	B
	irrelevant	C	D

- Relevant = A+C
- Retrieved = A+B
- Collection size = A+B+C+D
- Precision = $A \div (A+B)$
- Recall = $A \div (A+C)$
- Miss = $C \div (A+C)$ ความผิดพลาดที่ถูกต้อง / ความผิดพลาดที่ไม่ถูกต้อง
- False alarm (fallout) = $B \div (B+D)$ ความผิดพลาดที่ถูกต้อง / ความผิดพลาดที่ไม่ถูกต้อง

Retrieved vs. Relevant Documents

Very high precision, very low recall

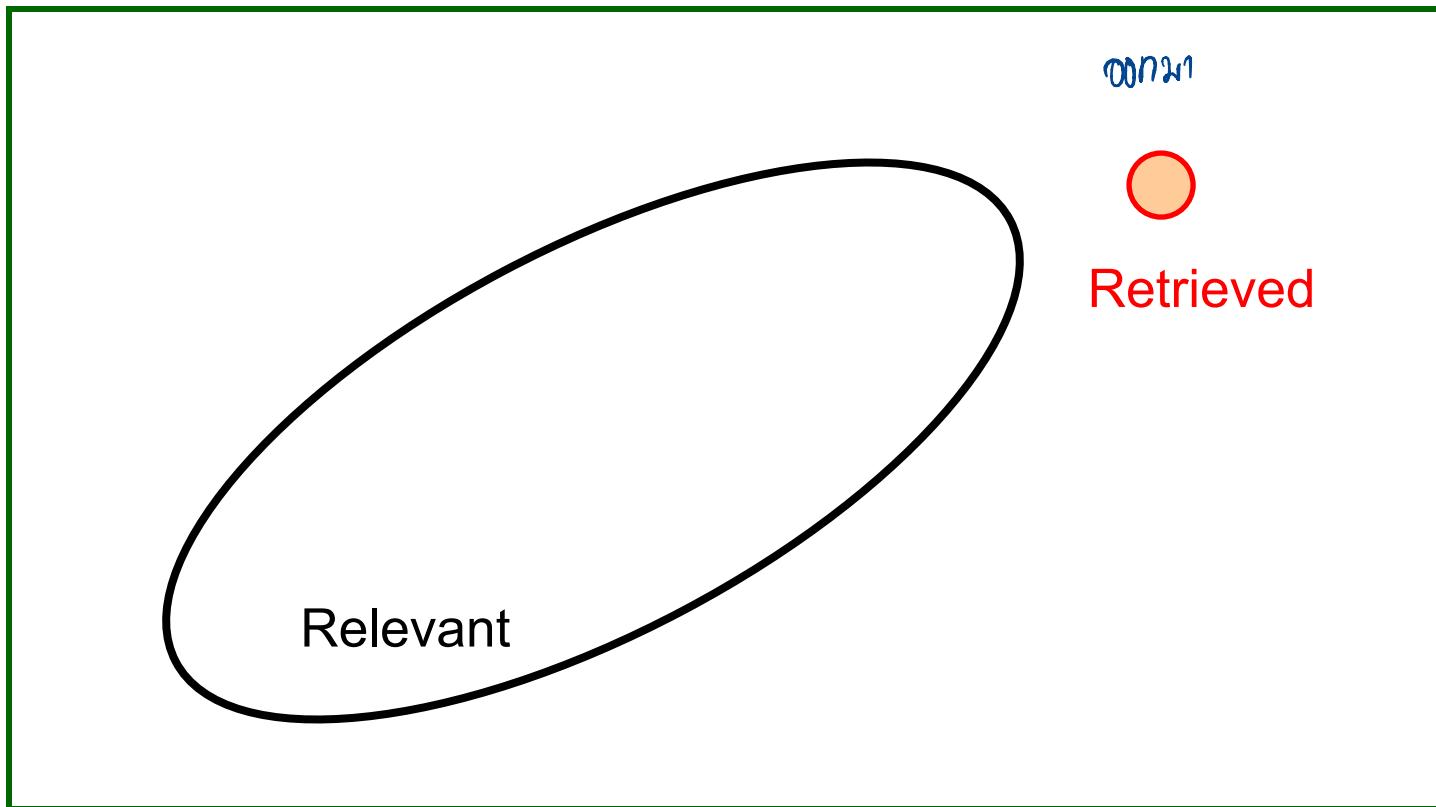
نامناسب



Retrieved vs. Relevant Documents

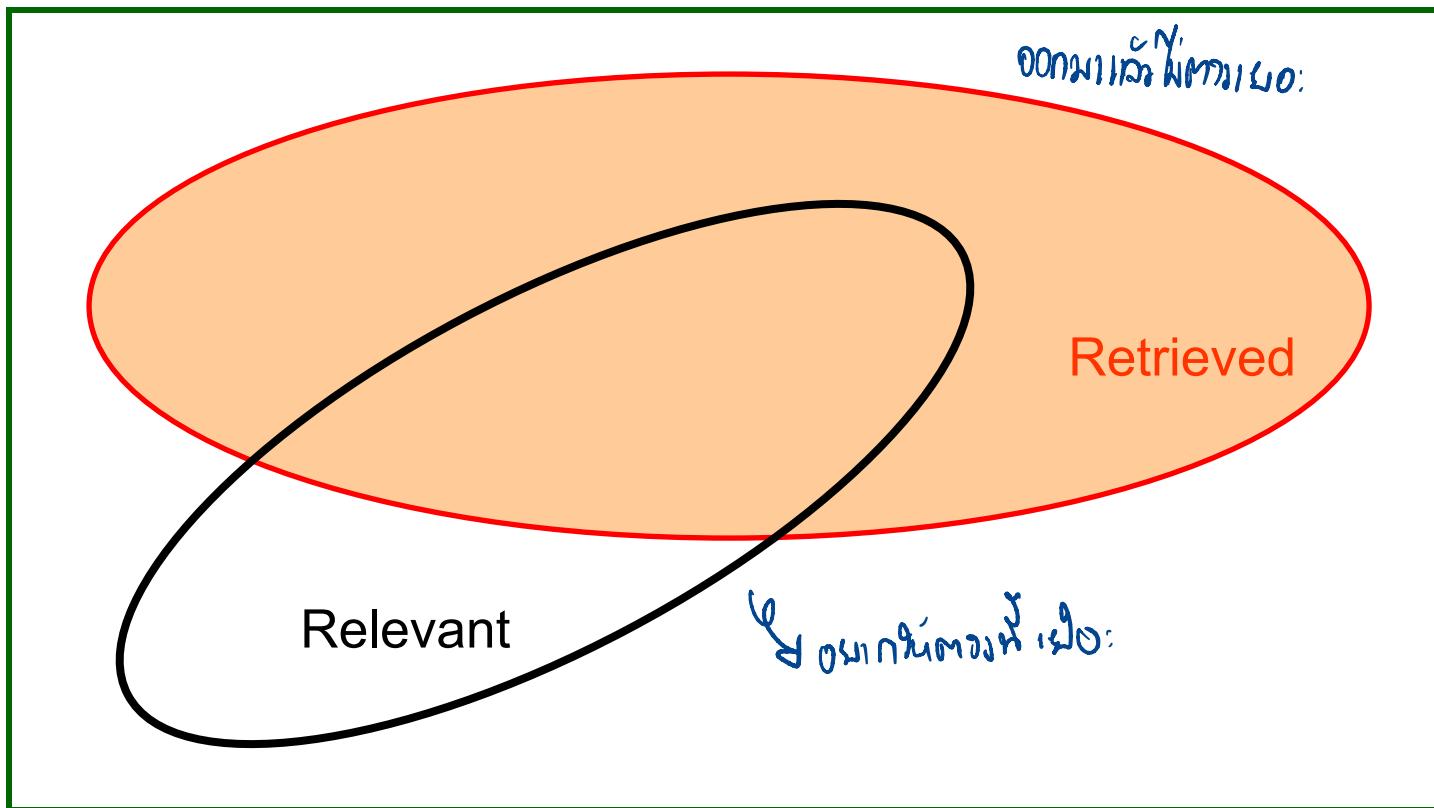
Very low precision, very low recall (0 in fact)

missed document



Retrieved vs. Relevant Documents

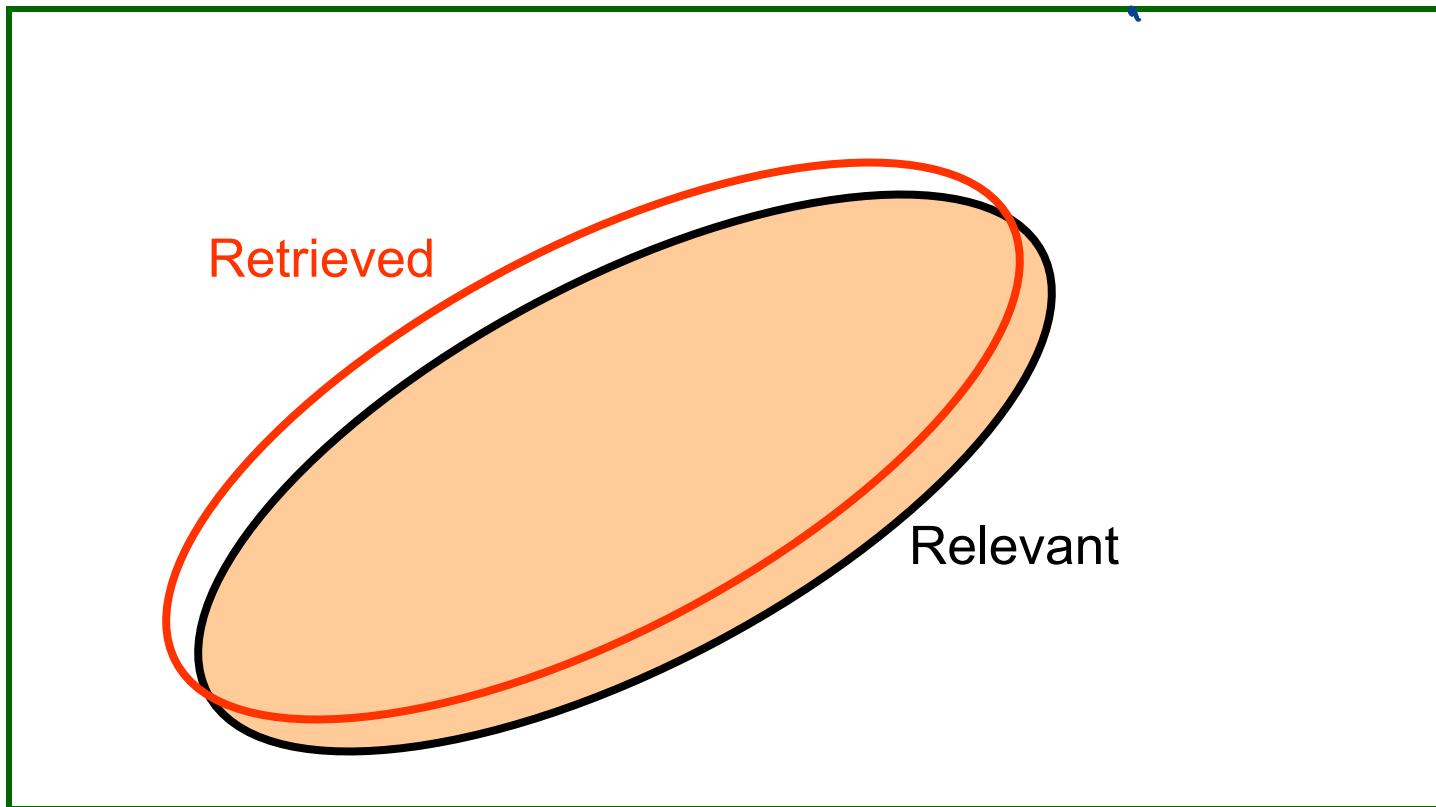
High recall, but low precision



Retrieved vs. Relevant Documents

High precision, high recall (at last!)

ສະບັບ
໨

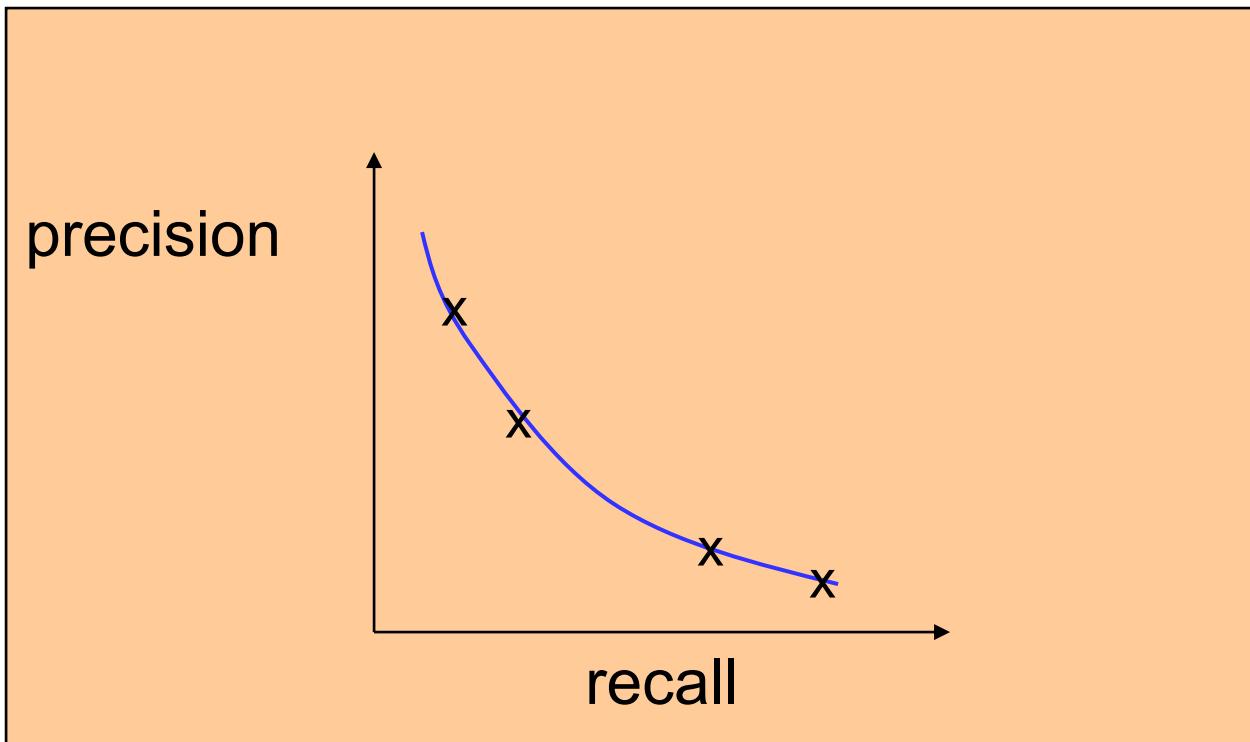


Average Recall/Precision Curve

- Typically average performance over a large **set** of queries.
- Compute average precision at each standard recall level across all queries.
- Plot average precision/recall curves to evaluate overall system performance on a document/query corpus.

Precision/Recall Curves

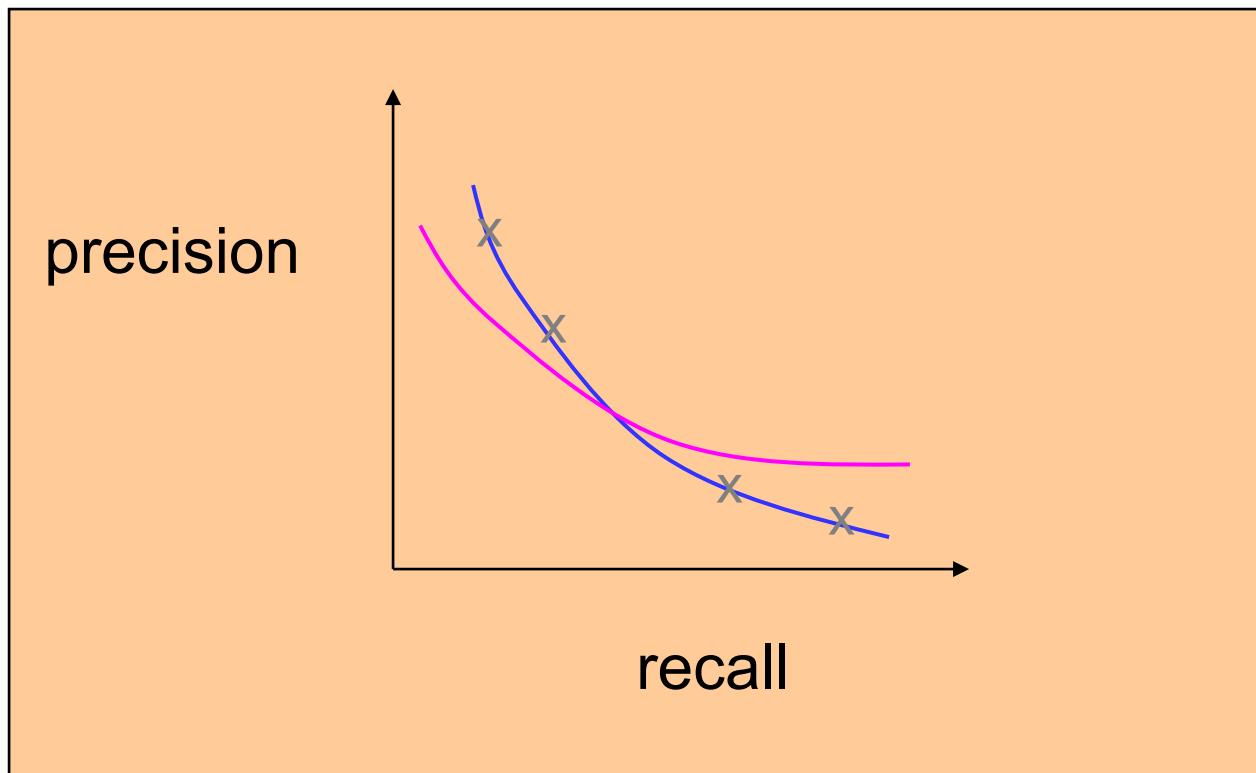
- There is a tradeoff between Precision and Recall 
- So measure Precision at different levels of Recall
- Note: this is an AVERAGE over MANY queries



Precision/Recall Curves

- Difficult to determine which of these two hypothetical results is better:

2 algorithm នាយករដ្ឋបាល ក្នុងវិជ្ជាគារណ៍អាជីវកម្ម



5 เอกสาร ที่ตรงประเด็น



การจัดลำดับ #1



กระบวนการหา מהเอกสารที่ตรงเป้าเก็ทไปต่อ
หากโจทย์ถาม Inn เอกสารที่ตรงเป้าเดิมๆ

#เอกสารที่ 1 2 3 4 5 6 7 8 9 10

ออกมาก



P

$1/1$

$1/2$

$2/3$

$2/4$

$2/5$

$3/6$

$3/7$

$3/8$

$4/9$

$5/10$

R

$1/5$

$1/5$

$2/5$

$2/5$

$2/5$

$3/5$

$3/5$

$3/5$

$4/5$

$5/5$

ลงมาเก็บที่บน

វិធាន់បង្កើត algorithms ពួរដោរ នៃកំណត់លទ្ធផល

ប្រើប្រាស់ពាណិជ្ជកម្មពីការស្វែងរកព័ត៌មានពីការស្វែងរកពីកម្ពស់ A និង B

A



P	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5
---	-----	-----	------	-----	-----	-----	------	------	------	-----

R	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
---	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

mean Prec ឡាតាំង

វិធាន់បង្កើត គឺ

$$AVGPrec_A = 62.2\%$$

B



P	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5
---	-----	-----	------	------	-----	-----	------	------	------	-----

R	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
---	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

$$AVGPrec_B = 52\%$$

$AVGPrec_A > AVGPrec_B \therefore A \text{ គឺចុះថ្មីជាបន្ទាន់ពីការស្វែងរកពីកម្ពស់} B$

Precision and Recall example



Ranking #1



Ranking #2



Computing Recall/Precision Points: An Example

n	doc #	relevant
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	

Let total # of relevant docs = 5
Check each new recall point:

$$R=1/5=0.2; \quad P=1/1=1$$

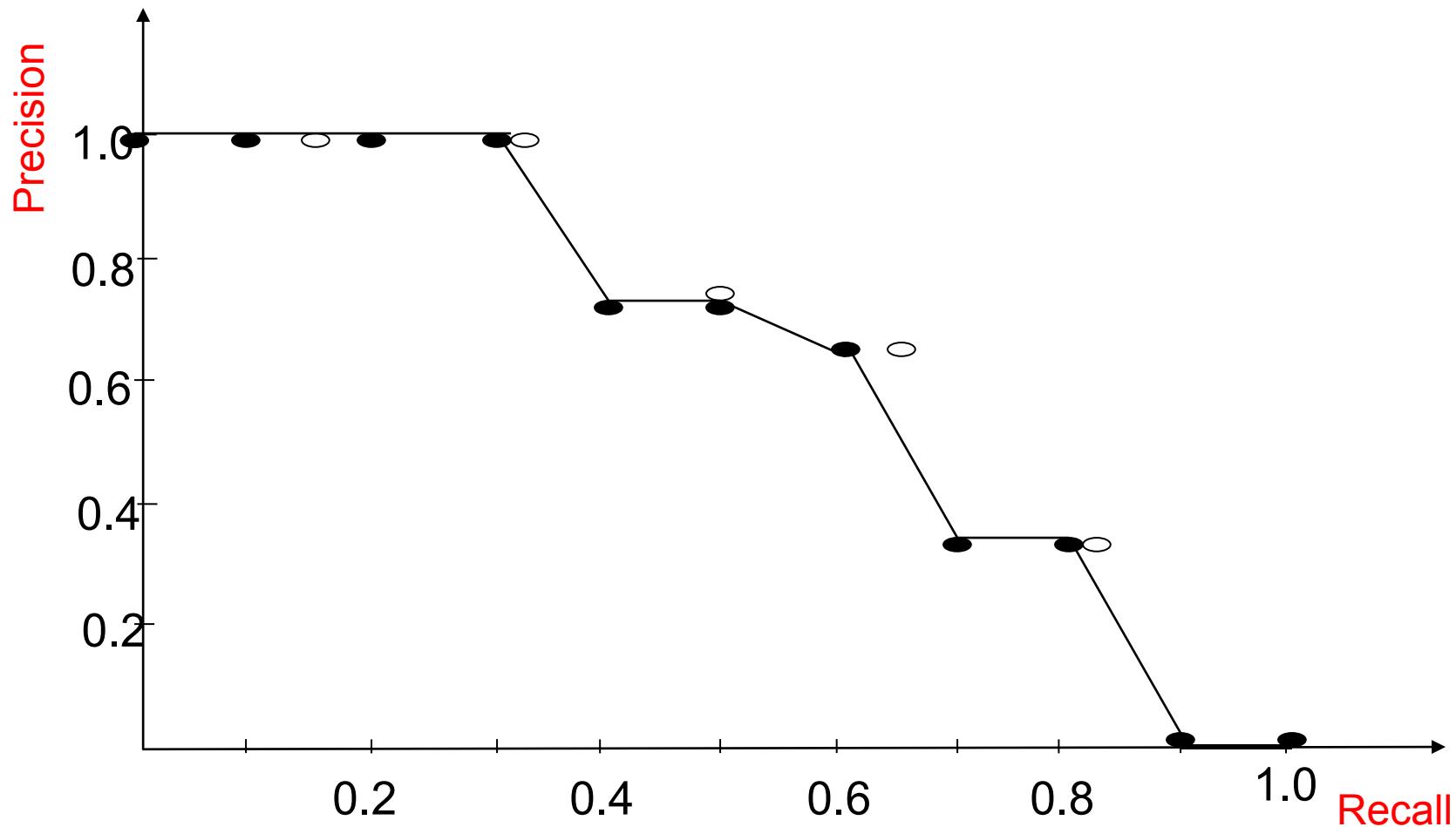
$$R=2/5=0.4; \quad P=2/2=1$$

$$R=3/5=0.6; \quad P=3/4=0.75$$

$$R=4/5=0.8; \quad P=4/6=0.67$$

$$R=5/5=1.0; \quad p=5/13=0.38$$

Interpolating a Recall/Precision Curve: An Example



Precision versus recall curve

ਪ੍ਰਚੱਲਿਤ ਕਰਨਾ

- $R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

Ranking for query q:

1.d ₁₂₃ *	6.d ₉ *	11.d ₃₈
2.d ₈₄	7.d ₅₁₁	12.d ₄₈
3.d ₅₆ *	8.d ₁₂₉	13.d ₂₅₀
4.d ₆	9.d ₁₈₇	14.d ₁₁
5.d ₈	10.d ₂₅ *	15.d ₃ *

- **P = 1 at R = 0.1**
- **P = 0.66 at R = 0.2**
- **P = 0.5 at R = 0.3**
- **P = 0.4 at R = 0.4**
- **P = 0.33 at R = 0.5**

**Usually based on 11 standard
recall levels: 0%, 10%, ..., 100%**

Precision versus recall curve

- For a single query



Precision and Recall

- Precision

ความสามารถในการค้นหานักเขียน

อันดับสูงสุด

ไม่ทราบเบอร์ = pre ตื้น
ทราบเบอร์ = pre ลึก

- The ability to retrieve top-ranked documents that are mostly relevant.

ที่มีความเกี่ยวข้องมากที่สุดในบัญชี

- Recall

ความสามารถในการค้นหานักเขียนที่เกี่ยวข้องที่มีผลลัพธ์มากที่สุด

ตรวจสอบที่ถูกจับมา = เก็บทุกเบอร์

- The ability of the search to find all of the relevant items in the corpus.

Determining Recall is Difficult

ນັ້ນມາວິທີ ດົບອານຸ

- Total number of relevant items is sometimes not available:
 - Sample across the database and perform relevance judgment on these items.
 - Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total relevant set.

Average Precision

- Often want a single-number effectiveness measure
 - E.g., for a machine-learning algorithm to detect improvement
- Average precision is widely used in IR
- Calculate by averaging precision when recall increases

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

AvgPrec= 62.2%

Recall	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
Precis.	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5

AvgPrec= 52.0%

Precision and Recall second example



= the relevant documents (as before)

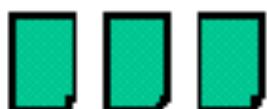
Ranking #1



Recall 0.2 0.2 0.4 0.4 0.4 0.6 0.6 0.6 0.8 1.0

Precis. 1.0 0.5 0.67 0.5 0.4 0.5 0.43 0.38 0.44 0.5

AvgPrec= 62.2%



= different query's relevant documents

Ranking #3



Recall 0.0 0.33 0.33 0.33 0.67 0.67 1.0 1.0 1.0 1.0

Precis. 0.0 0.5 0.33 0.25 0.4 0.33 0.43 0.38 0.33 0.3

AvgPrec= 44.3%

R- Precision

- Precision at the R-th position in the ranking of results for a query that has R relevant documents.

n	doc #	relevant
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	

ເພື່ອ R = # of relevant docs = 5

ມີອຳນວຍກຳດົກ ແກ້ໄຂ (n)
ຕັດກໍ n ມີການ
ການຝາກດົກ 5
ເລື່ອມວວກຕາມກຳນົດ
n otn

$$R\text{-Precision} = 3/5 = 0.60$$

ຕ້າຕວັນກຳກົດອຸ່ນໃຫຍງ 0.60% = 60% ມີຄວາມ

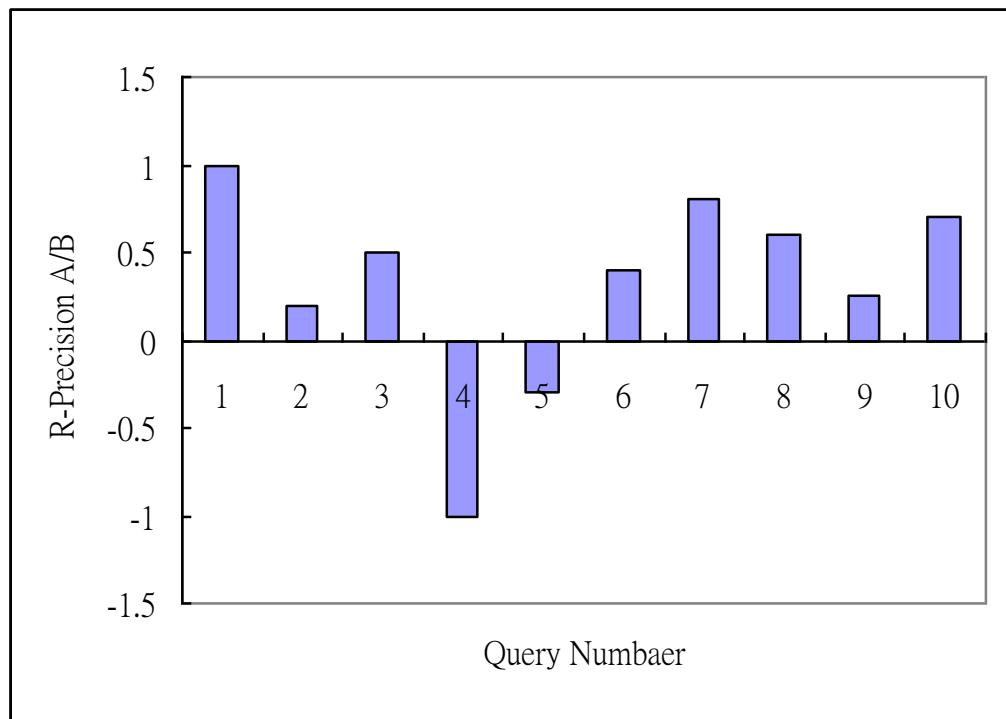
Precision Histograms

ກົດ algo : 10 ດີກາມ ສູນໄຟ ຕັ້ງໃຫຍ່ຈຳ R-pre ເພື່ອກຳນົດວ່າທຸກໆ
algo

- Use R-precision measures to compare the retrieval history of two algorithms through visual inspection

$$RP_{A/B}(i) = RP_A(i) - RP_B(i)$$

ເປົ້າ + ໄສະກັບ A
ເປົ້າ - ໄສະກັບ B



F-Measure

- One measure of performance that takes into account both recall and precision.
 - Harmonic mean of recall and precision:

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

↑ บันทวนสำหรับ
↑ ต่อต้าน

↑ ต่อต้าน

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.

E Measure (parameterized F Measure)

- A variant of F measure that allows weighting emphasis on precision over recall:

E តម្លៃគោលការណ៍

$$E = \frac{(1 + \beta^2)PR}{\beta^2 P + R} = \frac{(1 + \beta^2)}{\frac{\beta^2}{R} + \frac{1}{P}}$$

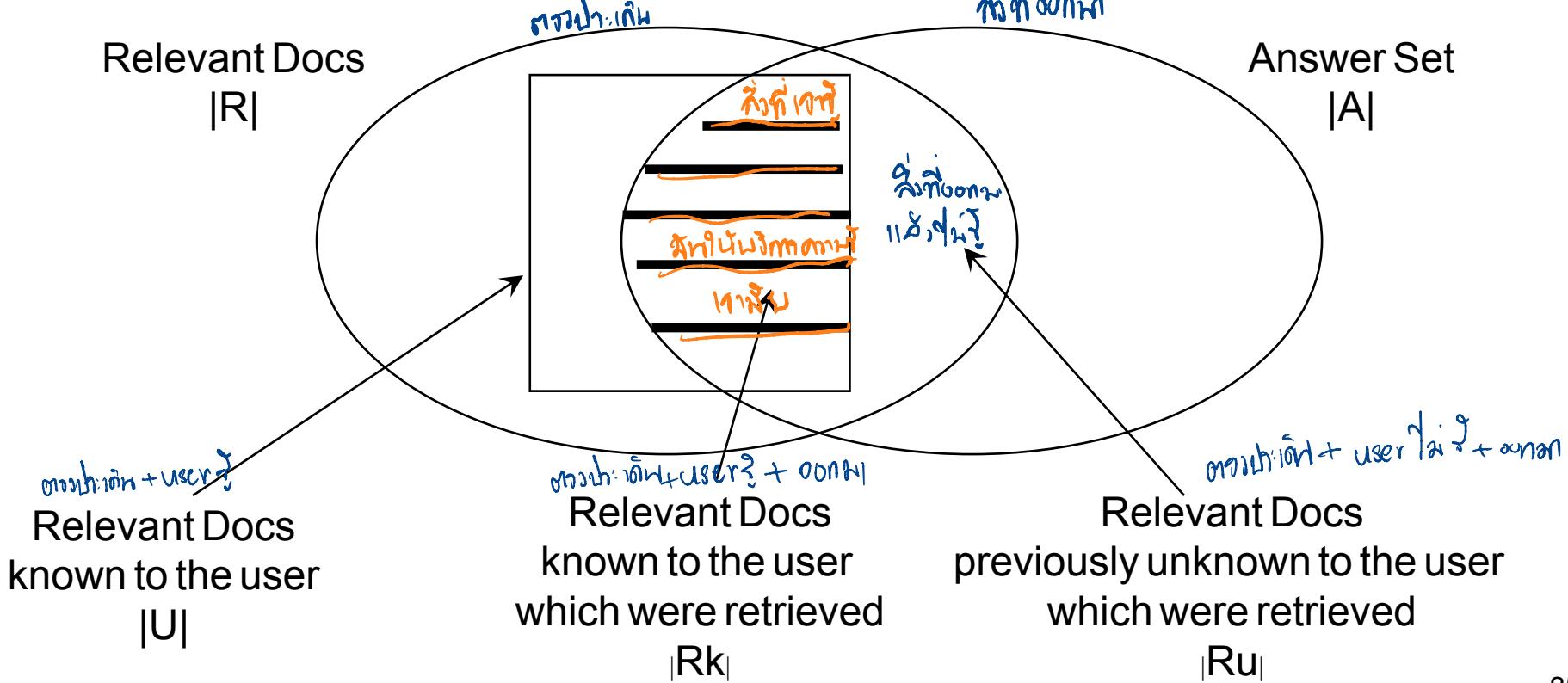
↑ ព័ត៌មានលម្អិត
↓ សំខាន់ៗ សម្រាប់
 អនុវត្តការណ៍

- Value of β controls trade-off:
 - $\beta = 1$: Equally weight precision and recall ($E=F$).
 - $\beta > 1$: Weight precision more.
 - $\beta < 1$: Weight recall more.

User-Oriented Measure

กิจกรรม การเขียน ฝึก

- Coverage=|R_k|/|U| ក្រោះសេចក្តី / ក្រោះអំពី
 - Novelty=|R_u|/(|R_u|+|R_k|) ក្នុងសេចក្តី / ការដាក់រាយក្នុងសេចក្តី



Fallout Rate

မာန်
ခို့ကြေး
Pre Avg
F
E

- Problems with both precision and recall:
 - Number of irrelevant documents in the collection is not taken into account.
 - Recall is undefined when there is no relevant document in the collection.
 - Precision is undefined when no document is retrieved.

$$Fallout = \frac{\text{no. of nonrelevant items retrieved}}{\text{total no. of nonrelevant items in the collection}}$$