VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY
UNIVERSITY OF INFORMATION TECHNOLOGY

DS307

SOCIAL MEDIA ANALYSIS

Faculty of Information Science and Engineering
University of Information Technology, VNU-HCM

This course's contents

Evaluation of IR

This course's contents

How good are the returned documents?

Measures for a search engine

- How fast does it index?

Measures for a search engine

- How fast does it index?
  - e.g., number of bytes per hour

Measures for a search engine

- How fast does it index?
  - e.g., number of bytes per hour
- How fast does it search?

Measures for a search engine

- **How fast does it index?**
  - e.g., number of bytes per hour
- **How fast does it search?**
  - e.g., latency as a function of queries per second

Measures for a search engine

- **How fast does it index?**
  - e.g., number of bytes per hour
- **How fast does it search?**
  - e.g., latency as a function of queries per second
- **What is the cost per query?**

Measures for a search engine

- **How fast does it index?**
  - e.g., number of bytes per hour
- **How fast does it search?**
  - e.g., latency as a function of queries per second
- **What is the cost per query?**
  - in dollars

Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money

Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money
- However, the key measure for a search engine is **user happiness**.

Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money
- However, the key measure for a search engine is **user happiness**.
- What is user happiness?

Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money
- However, the key measure for a search engine is **user happiness**.
- What is user happiness?
- Factors include:

Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money
- However, the key measure for a search engine is **user happiness**.
- What is user happiness?
- Factors include:
  - Speed of response
  - Size of index
  - Uncluttered UI
  - We can measure
    - Rate of return to this search engine
    - Whether something was bought
    - Whether ads were clicked
  - Most important: relevance

Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money
- However, the key measure for a search engine is **user happiness**.
- What is user happiness?
- Factors include:
  - Speed of response
  - Size of index
  - Uncluttered UI
  - We can measure
    - Rate of return to this search engine
    - Whether something was bought
    - Whether ads were clicked
  - **Most important: relevance**

Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money
- However, the key measure for a search engine is **user happiness**.
- What is user happiness?
- Factors include:
  - Speed of response
  - Size of index
  - Uncluttered UI
  - We can measure
    - Rate of return to this search engine
    - Whether something was bought
    - Whether ads were clicked
  - Most important: relevance
  - (actually, maybe even more important: it's free)

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements
  - Benchmark document collection
  - A benchmark suite of queries
  - An assessment of the relevance of each query-document pair

Relevance: query vs. information need

- Relevance to what? The query?
  - **Information need i**: "I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine."
- Translated into:
  - **Query q**: [red wine white wine heart attack]
- So what about the following document:
  - **Document d′**: At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.

Relevance: query vs. information need

- User happiness can only be measured by relevance to an information need, not by relevance to queries.

- Sloppy terminology here and elsewhere in the literature: we talk about query–document relevance judgments even though we mean information-need–document relevance judgments.

Relevance: query vs. information need

- Precision (P) is the fraction of retrieved documents that are relevant:

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- Recall (R) is the fraction of relevant documents that are retrieved:

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

Relevance: query vs. information need

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

Relevance: query vs. information need

	THE TRUTH	
	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

Relevance: query vs. information need

	THE TRUTH: Relevant	THE TRUTH: Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

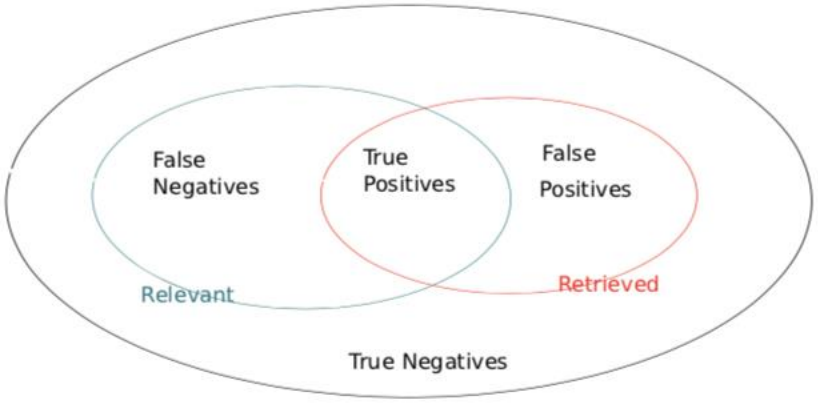
Relevance: query vs. information need

		THE TRUTH	
WHAT THE SYSTEM THINKS		Relevant	Nonrelevant
	Retrieved	true positives (TP)	false positives (FP)
	Not retrieved	false negatives (FN)	true negatives (TN)

Relevance: query vs. information need

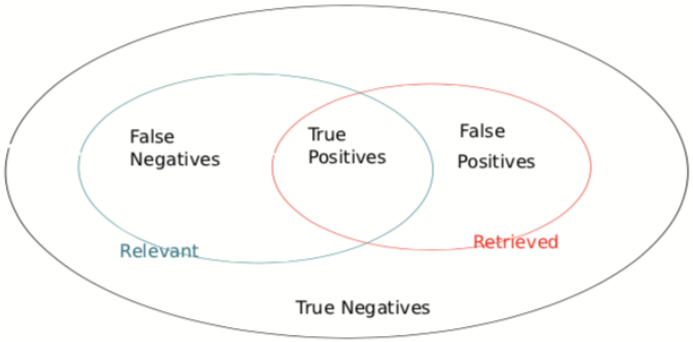
THE TRUTH

WHAT THE SYSTEM THINKS		Relevant	Nonrelevant
	Retrieved	true positives (TP)	false positives (FP)
	Not retrieved	false negatives (FN)	true negatives (TN)



Relevance: query vs. information need

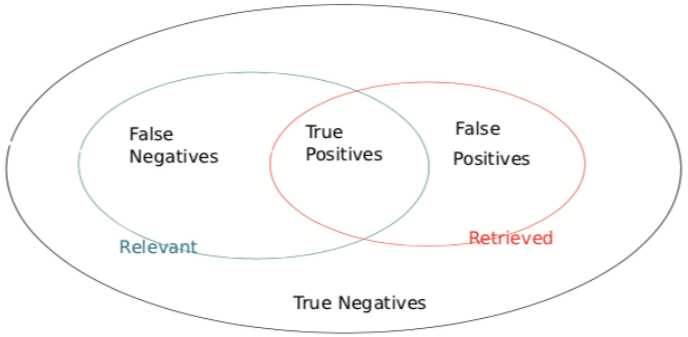
		THE TRUTH: Relevant	THE TRUTH: Nonrelevant
WHAT THE SYSTEM THINKS	Retrieved	true positives (TP)	false positives (FP)
	Not retrieved	false negatives (FN)	true negatives (TN)



$$P = TP/(TP + FP)$$

Relevance: query vs. information need

		THE TRUTH: Relevant	THE TRUTH: Nonrelevant
WHAT THE SYSTEM THINKS	Retrieved	true positives (TP)	false positives (FP)
	Not retrieved	false negatives (FN)	true negatives (TN)



$$P = TP/(TP + FP)$$
$$R = TP/(TP + FN)$$

Precision/Recall Trade-off

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs has 100% recall!
- The converse is also true (usually): It's easy to get high precision for very low recall.

A combined measure: F

- F allows us to trade off precision against recall

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

α ∈ [0,1] and thus β2 ∈ [0,∞]

Most frequently used: balanced F with β = 1 or α = 0.5

This is the harmonic mean of P and R: 1/F = ½.(1/P + 1/R)

Example for precision, recall and F1

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

P =20/(20+40)=1/3

R =20/(20+60)=1/4

$$F1 = 2.\frac{1}{\frac{1}{\frac{1}{3}}+\frac{1}{\frac{1}{4}}} = 2/7$$

Accuracy

Why do we use complex measures like precision, recall, and F ?

Accuracy

- ✓ Why do we use complex measures like precision, recall, and F ?
- ✓ Why not something simple like accuracy?
- ✓ Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- ✓ In terms of the contingency table above, accuracy = (TP + TN)/(TP + FP + FN + TN).

Accuracy

- ✓ Why do we use complex measures like precision, recall, and F ?
- ✓ Why not something simple like accuracy?
- ✓ Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- ✓ In terms of the contingency table above, accuracy = (TP + TN)/(TP + FP + FN + TN).

Though experiment

- Computing precision, recall and F1-score for this result set:

	relevant	not relevant
retrieved	**18**	2
not retrieved	82	**1,000,000**

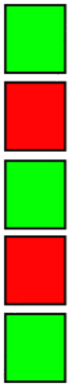
- You then get 99.99% accuracy on most queries.

Rank-Based Measures

- Binary relevance
  - **Precision@K (P@K)**
  - **Mean Average Precision (MAP)**
  - **Mean Reciprocal Rank (MRR)**
- Multiple levels of relevance
  - Normalized Discounted Cumulative Gain (NDCG)

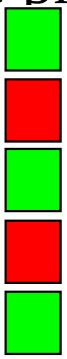
Precision@K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K
- Ex:
  - Prec@3 of 2/3
  - Prec@4 of 2/4
  - Prec@5 of 3/5

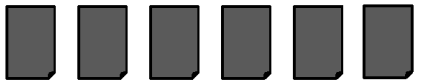


- In similar fashion we have Recall@K

Mean Average Precision

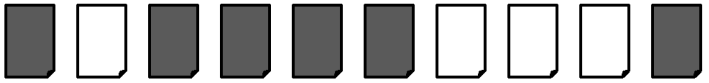
- Consider rank position of each ***relevant*** doc
  - $K_1, K_2, \ldots K_R$
- Compute Precision@K for each $K_1, K_2, \ldots K_R$
- Average precision = average of P@K
- Ex:  has AvgPrec of $\frac{1}{3} \cdot \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$
- MAP is Average Precision across multiple queries/rankings

Average Precision



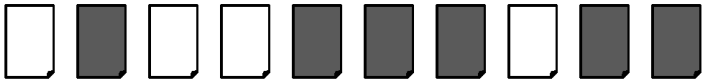
= the relevant documents

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

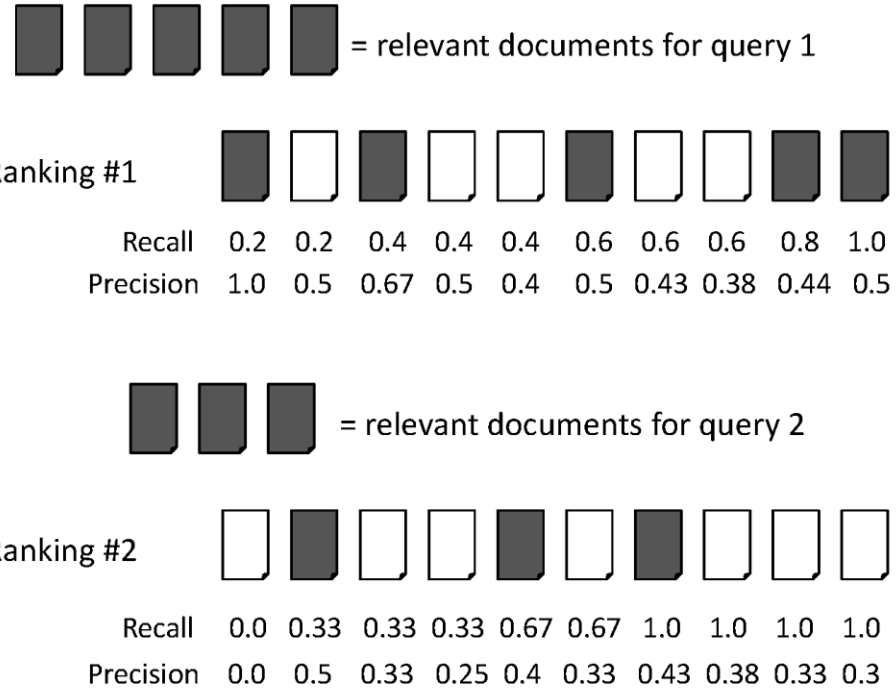


Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

$$\text{Ranking } \#1: \ (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking } \#2: \ (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

MAP



$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

Q&A

Thank you!