Data Mining
資料探勘

Evaluation

Hung-Yu Kao, Fall 2019

RETRIEVAL/PREDICTION EVALUATION

Introduction

- Type of evaluation
  - Functional analysis phase, and Error analysis phase
  - Performance evaluation

- Performance evaluation
  - Response time/space required

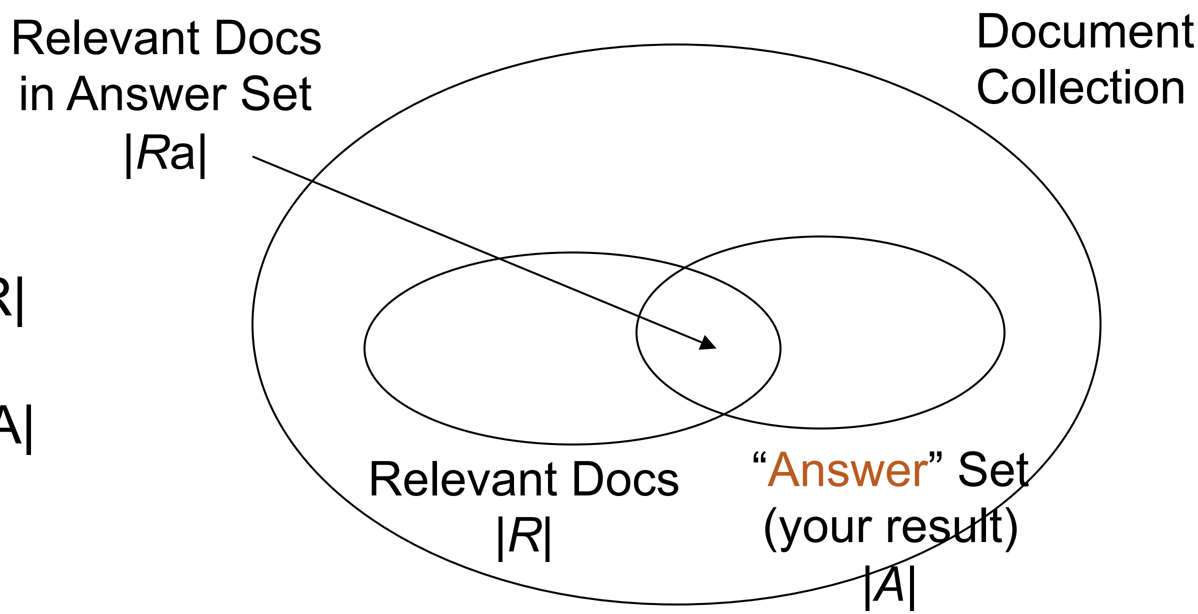
- Retrieval performance evaluation
  - The evaluation of how precise is the answer set
    - Or so-call Answer? Result? Ground truth?

Recall and Precision (for retrieval)

- Recall:
  - The fraction of the relevant documents (**R**) which has been retrieved
- Precision:
  - The fraction of the retrieved documents (**A**) which is relevant

Recall = |Ra| / |R|

Precision = |Ra| / |A|



Precision versus recall curve

- R_q={d3,d5,d9,d25,d39,d44,d56,d71,d89,d123}

<u>Ranking for query q:</u>

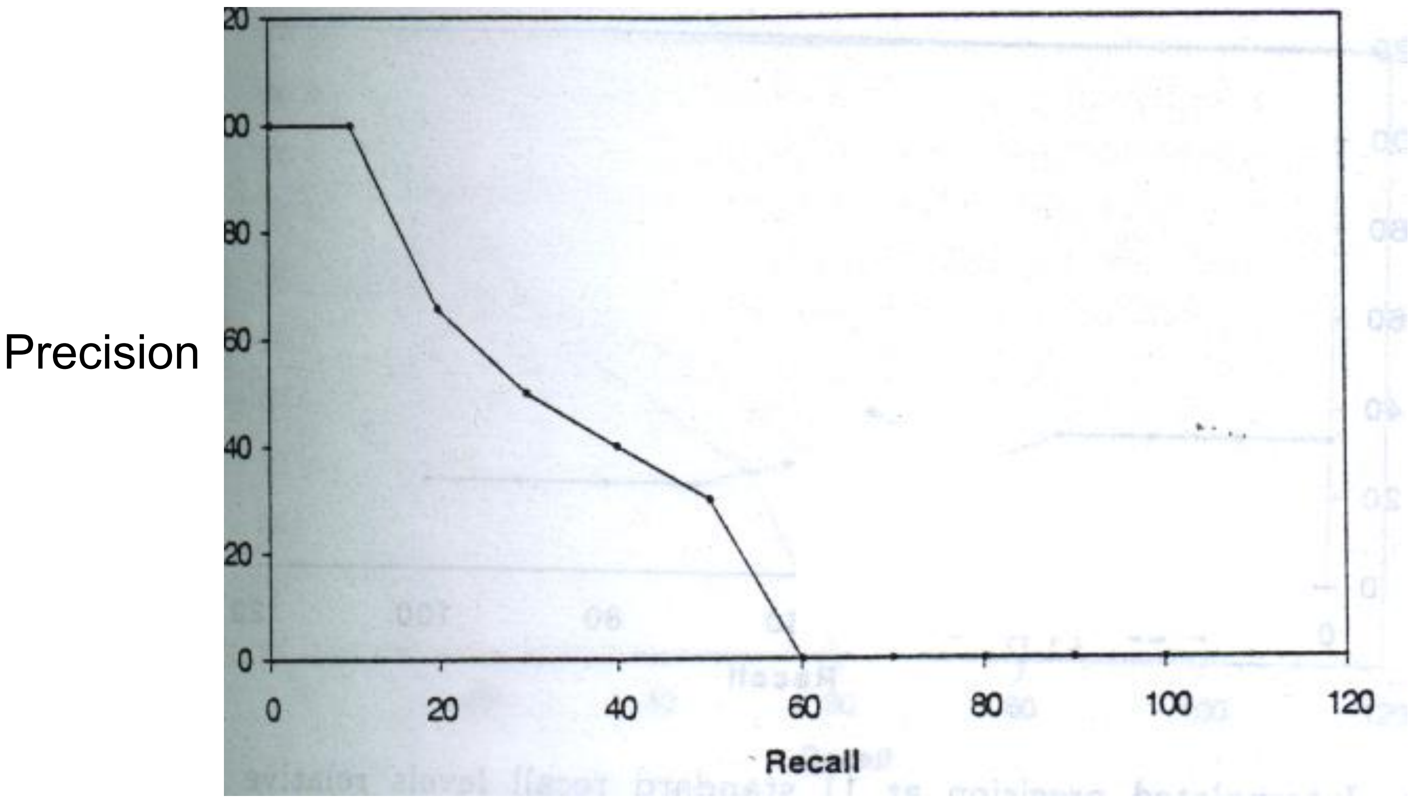
1.d123*	6.d9*	11.d38
2.d84	7.d511	12.d48
3.d56*	8.d129	13.d250
4.d6	9.d187	14.d11
5.d8	10.d25*	15.d3*

- P=100% at R=10%
- P= 66% at R=20%
- P= 50% at R=30%

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

Precision versus recall curve

- For a single query



Top-k precision / Precision at k (P@k)

- Precision evaluation in a ranking list
- The precision value of the top-k results
- Top-1, 2, 5, 10, … / P@1, P@2, P@5, P@10, …
- Frequently used in search engine evaluation

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

P@1= 100%
P@2= 50%
P@3= 66%
P@5= 40%
P@10=40%

Average Over Multiple Queries

$$\overline{P}(r) = \frac{1}{N_q} \sum_{i=1}^{N_q} P_i(r)$$

- $\overline{P}(r)$=average precision at the recall level *r*
- N_q= Number of queries used
- $P_i(r)$=The precision at recall level r for the i-th query

Interpolated precision

- $R_q=\{d_3, d_{56}, d_{129}\}$

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*

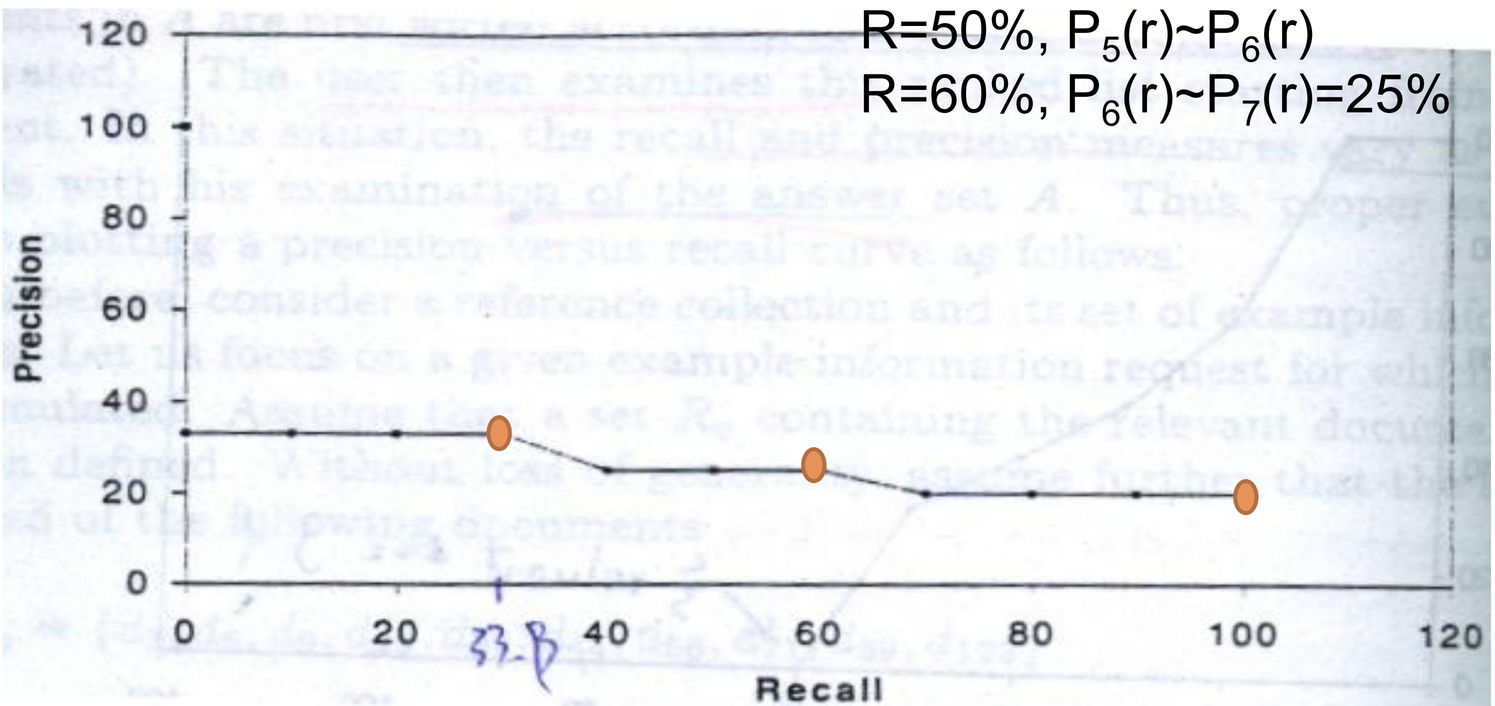
- P=33% at R=33%
- P= 25% at R=66%
- P= 20% at R=100%

- $P(r_j)=\max_{r_i \leqq r \leqq r_{j+1}} P(r)$

Interpolated precision

- Let r_j, $j \in \{0, 1, 2, \ldots, 10\}$, be a reference to the j-th standard recall level
- $P(r_j) = \max_{r_j \leqq r \leqq r_{j+1}} P(r)$

R=30%, $P_3(r)$~$P_4(r)$=33%
R=40%, $P_4(r)$~$P_5(r)$
R=50%, $P_5(r)$~$P_6(r)$
R=60%, $P_6(r)$~$P_7(r)$=25%



Single Value Summaries

- Average precision versus recall:
  - Compare retrieval algorithms over a *set of example queries*
- Sometimes we need to compare individual query's performance
  - Average precision可能會隱藏演算法中不正常的部分
  - 可能需要知道, 兩個演算法中，對某特定query的performance為何
- Need a single value summary
  - The single value should be interpreted as a summary of the corresponding precision versus recall curve

MAP: mean average precision

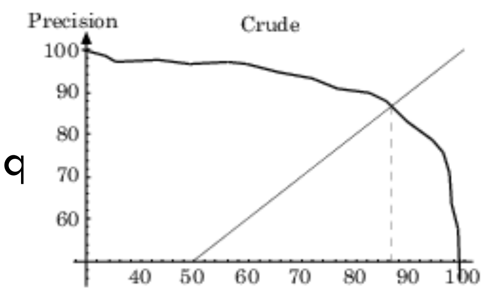
- Average of the precision value obtained for the top *k* documents, *each time a relevant doc is retrieved*
- Avoids interpolation, use of fixed recall levels
- MAP for query collection is arithmetic ave.
  - Macro-averaging: each query counts equally

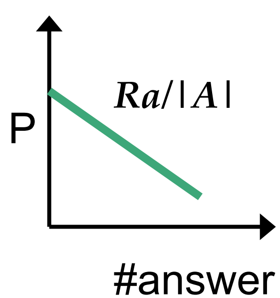
$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk})$$

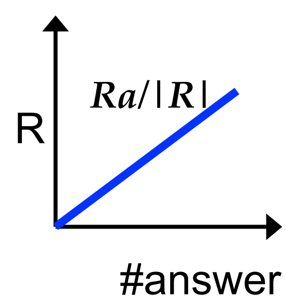
Single Value Summaries

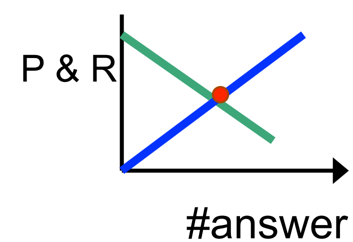
- Average Precision at Seen Relevant Documents
  - Averaging the precision figures obtained after each new relevant document is observed.
  - Example: (1+0.66+0.5+0.4+0.3)/5=0.57
  - 此方法對於很快找到相關文件的系統是相當有利的(相關文件被排在越前面, precision值越高)

- R-Precision (break-even point)
  - The precision at **the R-th position** in the ranking
  - R: the total number of relevant documents of the current q (total number in R_q)
  - E.g., RP=0.33 in the previous example









R-Precision vs. MAP

- MAP practice
  - System1 RNRNN NNNRR
  - System2 NRNNR RRNNN
  - What is the MAP of each system?
  - And their RP?

MRR: Mean Reciprocal Rank

- the multiplicative inverse of the rank of the first correct answer

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}.$$

Query	Results	Rank	MRR
1	X X O	3	1/3
2	X O X	2	1/2
3	O X X	1	1

Precision-Recall Averages
-- for multiple categories

- Microaveraging

$$P^{\mu} = \frac{\sum_{c=1}^{k} TP_c}{\sum_{c=1}^{k} (TP_c + FP_c)}$$

$$R^{\mu} = \frac{\sum_{c=1}^{k} TP_c}{\sum_{c=1}^{k} (TP_c + FN_c)}$$

重視量

- Macroaveraging

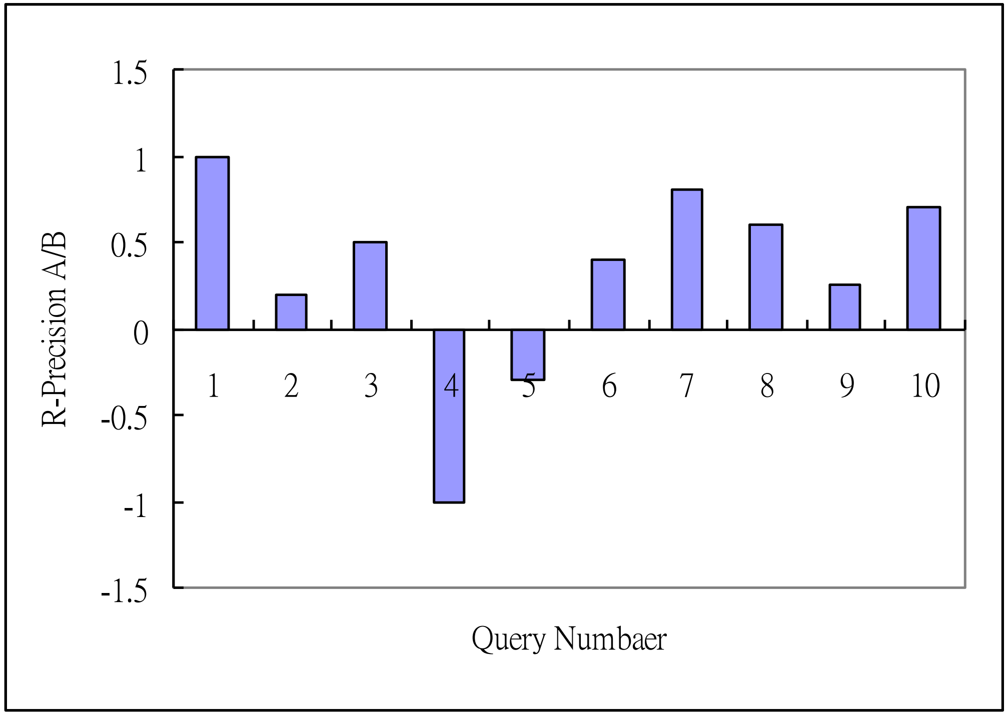
$$P^{M} = \frac{1}{K} \sum_{c=1}^{K} P_c$$

$$R^{M} = \frac{1}{K} \sum_{c=1}^{K} R_c$$

重視種類

Precision Histograms

- Use R-precision measures to compare the retrieval history of two algorithms through visual inspection
- $RP_{A/B}(i) = RP_A(i) - RP_B(i)$



Summary Table Statistics

- 將所有query相關的single value summary 放在table中
  - the number of queries ,
  - total number of documents retrieved by all queries,
  - total number of relevant documents were effectively retrieved when all queries are considered
  - total number of relevant documents retrieved by all queries…

Search Result Comparison (polling)

Rank	LAMIS-LN-CB-HR-TW			HITS		
	ID	Ans	URL	ID	Ans	URL
1	554	●	/wp-dyn/articles/A4931-2002Apr17.html	65	○	/wp-dyn/sports/leaguesandsports/nhl/
2	131	○	/wp-srv/front.htm	66	●	/wp-dyn/articles/A5164-2002Apr17.html
3	1	○	/	397	●	/wp-dyn/articles/A5101-2002Apr17.html
4	484	○	/wp-dyn/print/sports/inside/	398	●	/wp-dyn/articles/A5731-2002Apr18.html
5	9	○	/wp-dyn/sports/	399	●	/wp-dyn/articles/A4954-2002Apr17.html
6	420	○	/wp-dyn/sports/leaguesandsports/nba/	405	○	/wp-dyn/sports/leaguesandsports/mlb/
7	405	○	/wp-dyn/sports/leaguesandsports/mlb/	420	○	/wp-dyn/sports/leaguesandsports/nba/
8	319	○	/wp-dyn/print/metro/	67	●	/wp-dyn/articles/A4919-2002Apr17.html
9	286	○	/wp-dyn/world/latestap/	396	●	/wp-dyn/articles/A4942-2002Apr17.html
10	7	○	/wp-dyn/world/	394	●	/wp-dyn/articles/A4713-2002Apr17.html
11	160	●	/wp-dyn/metro/traffic/	467	●	/wp-dyn/articles/A4887-2002Apr17.html
12	314	●	/traffic	478	●	/wp-dyn/articles/A4712-2002Apr17.html
13	4	●	/wp-dyn/metro/traffic/index.html	480	●	/wp-dyn/articles/A4823-2002Apr17.html
14	184	●	/ac2/wp-dyn/metro/traffic	481	●	/wp-dyn/articles/A5475-2002Apr17.html
15	23	○	/wp-dyn/digest/	390	○	/wp-dyn/sports/leaguesandsports/nba/19992000/
16	8	○	/wp-dyn/metro/	400	●	/wp-dyn/articles/A4955-2002Apr17.html
17	10	○	/wp-dyn/business/	391	○	/wp-dyn/sports/leaguesandsports/nfl/20002001/
18	543	○	/wp-dyn/business/latestap/	388	○	/wp-dyn/sports/leaguesandsports/mlb/2000/
19	6	○	/wp-dyn/nation/	389	○	/wp-dyn/sports/leaguesandsports/mls/2000/
20	229	○	/wp-dyn/nation/specials/attacked/	393	○	/wp-dyn/sports/leaguesandsports/wnba/2000/

○: a TOC page ●: a not-TOC page

right answer

wrong answer

Precision and Recall 的適用性

- Maximum recall值的產生，需要知道所有文件相關的背景知識
- Recall and precision是相對的測量方式，兩者要合併使用比較適合
  - Application dependent
- Recall + Precision = Constant ?
  - **Average of Recall and Precision**

Alternative Measures

- The Harmonic Mean, F-measure (Rijsbergen, 1979)
  - $$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$,介於0,1

$$F_\beta = \frac{1+\beta^2}{\frac{\beta^2}{\text{Recall}} + \frac{1}{\text{Precision}}} = (1+\beta^2)(\text{Precision} * \text{Recall}) / (\beta^2 * \text{Precision} + \text{Recall})$$

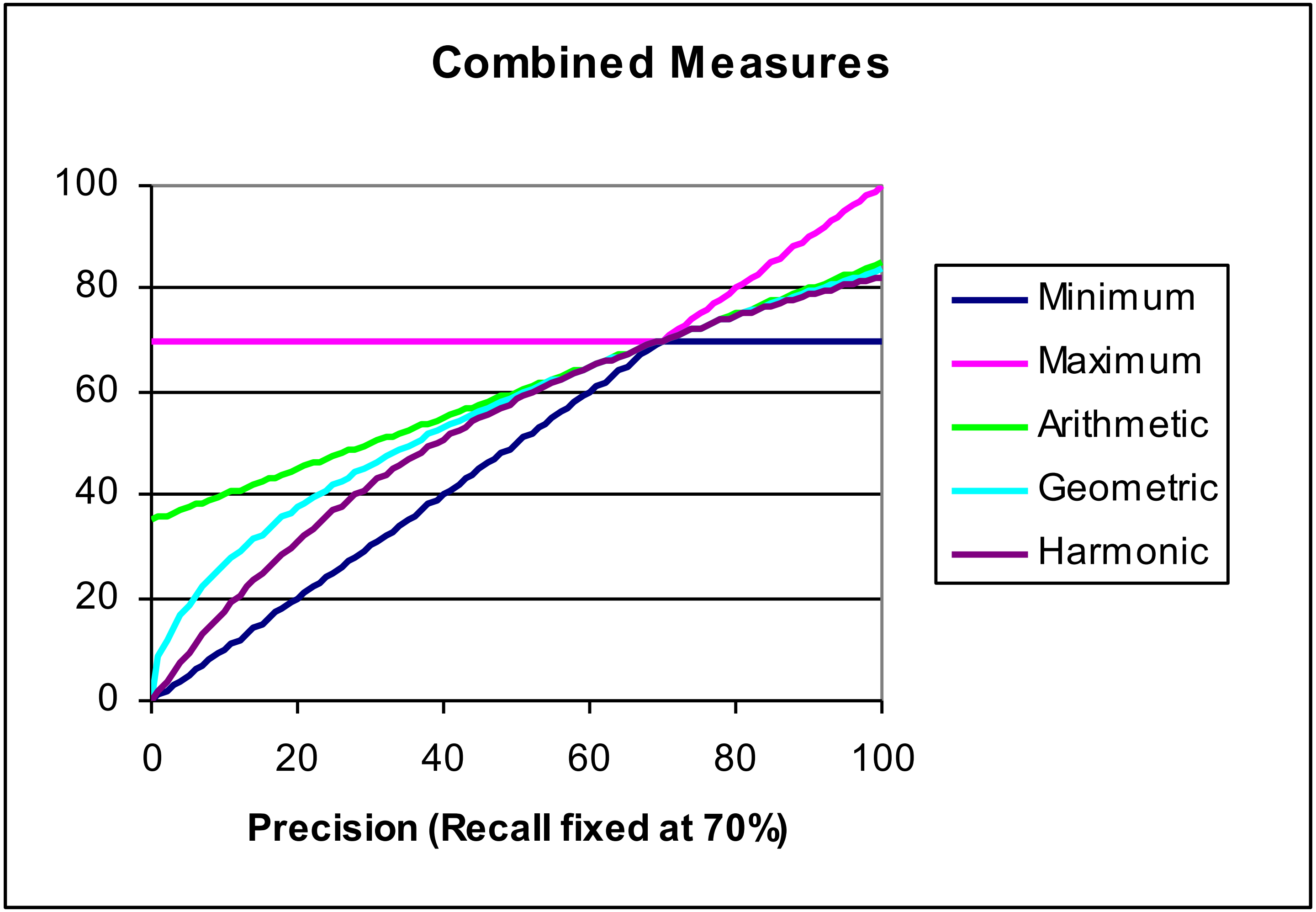
- 加入喜好比重 (effectiveness measure)
- The E Measure-
  $$E(j) = 1 - \frac{1+b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$
  - b=1, $E(j)=F(j)$
  - b>1, more interested in precision
  - b<1, more interested in recall

F-measure examples

Method	Precision	Recall	average	F-1
1	0.5	0.6	0.55	0.545
2	0.4	0.7	0.55	0.509

Method	Precision	Recall	average	F-1
1	0.4	0.7	0.55	0.509
2	0.5	0.7	0.60	0.583
3	0.4	0.8	0.60	0.533
4	0.45	0.7	0.575	0.547

F_1 and other averages

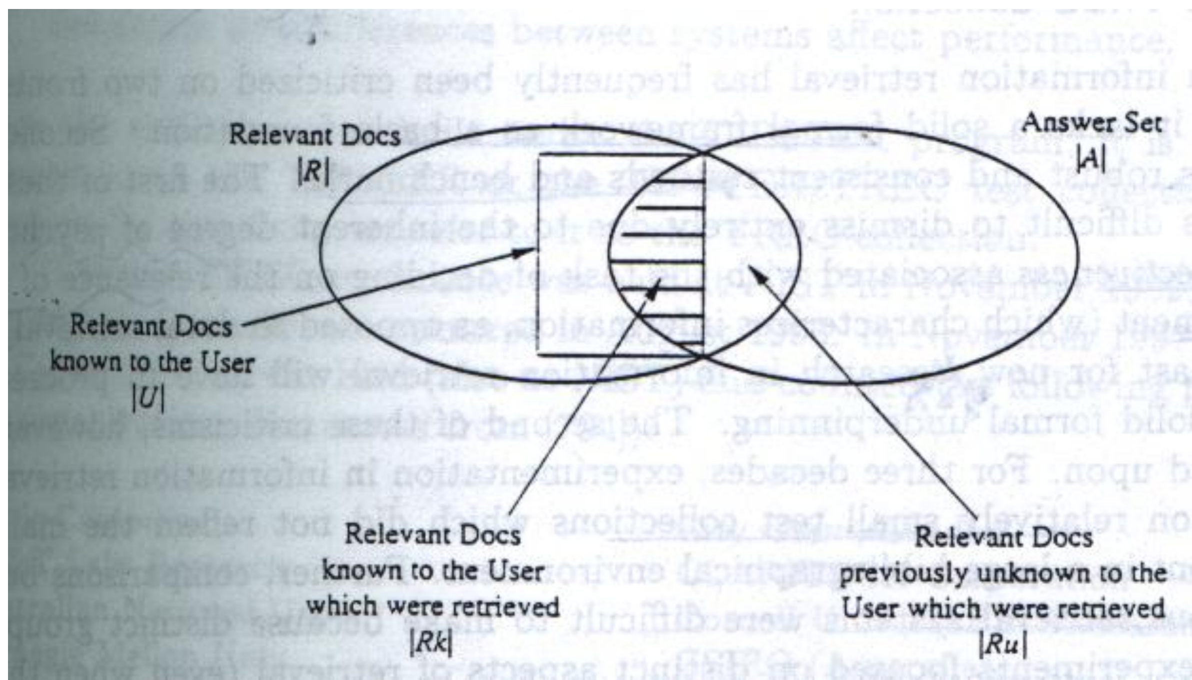


User-Oriented Measure

- 假設：Query與使用者有相關,不同使用者有不同的 relevant docs
  - Coverage=$|R_k|/|U|$
  - Novelty=$|R_u|/(|R_u|+|R_k|)$

◆Coverage越高,系統找到使用者期望的文件越多

◆Noverlty越高,系統找到許多使用者之前不知道相關的文件越多

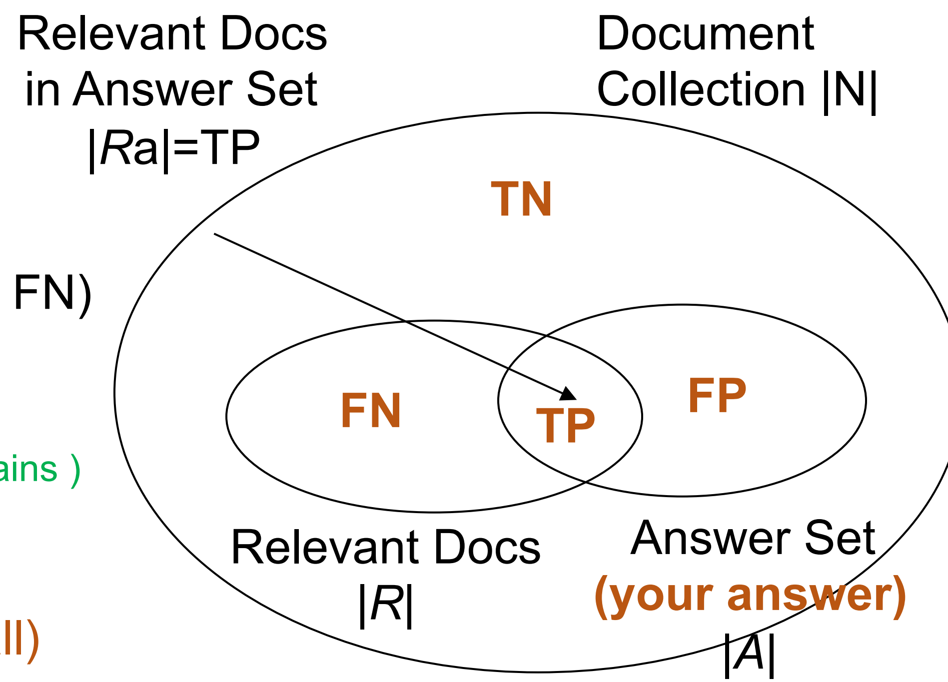


Alternative Measures / confusion matrix (contingency matrix?)

(predict) Answer	Relevant -	Relevant +
-	TN	FN
+	FP	TP

FP: type I error, alpha error

FN: type II error, beta error



Recall (sensitivity) = |Ra| / |R| = TP / (TP + FN)

Precision = |Ra| / |A| = TP / (TP + FP)

Accuracy = (**TN** + **TP**) / |N| (For balanced domains)

classification error, *E = 1 – A*

Specificity = TN / (TN + FP) (negative recall)

(not useful for Web search, TN is always so large)

An example (10000 sick + 10000 healthy)

		HIV Infected	
		+	-
ELISA	+	9990 (TP)	10 (FP)
	-	10 (FN)	9990 (TN)
		10,000 TP+FN	10,000 FP+TN
		Sensitivity = TP/(TP+FN) 9990/(9990+10) =.999 or 99.9%	Specificity= TN/(FP+TN) 9990/(9990+10) =.999 or 99.9%

2% :Sick 20/ Healthy 980

	+	-
+	8	10
-	12	970

Sensitivity: 8 / (8 + 12)
Specificity: 970 / (970+10)

A sensitivity of 100% means that the test recognizes all sick people as such

*A specificity of **100%** means that the test recognizes all healthy people as healthy*

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If a model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost Matrix

	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

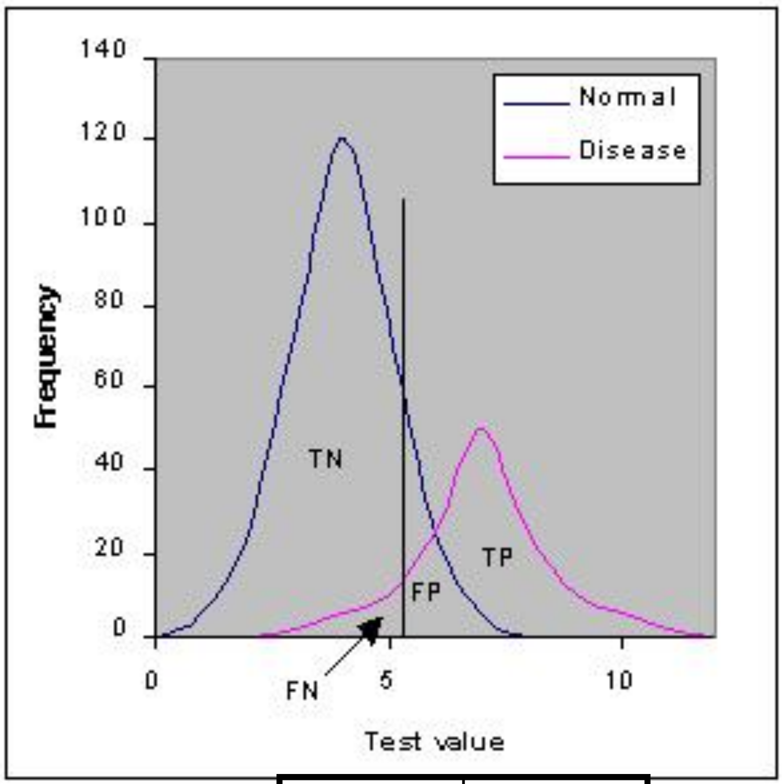
Count	PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

ROC curve

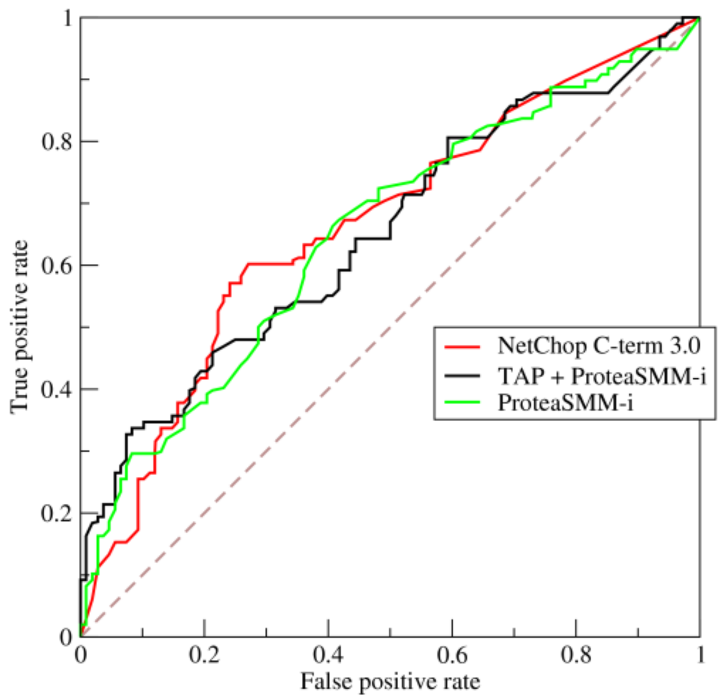
- **receiver operating characteristic (ROC,**接收器運作指標曲線)
- 起源研究軍事雷達的敵我偵測能力, 1954年情報理論研討會
- is a graphical plot of the ***sensitivity*** vs. ***(1 - specificity)*** for a binary classifier system as its discrimination threshold is varied
  - TPR (TP/(TP + FN)) vs. FPR (FP/(FP + TN))



(predict) Answer	Relevant	
	-	+
-	TN	FN
+	FP	TP

ROC curve

- equivalently by plotting the fraction of true positives vs. the fraction of false positives.
- the area under the ROC curve, or "AUC".
- What's the meaning of the dotted line?
- If we don't know all negative data / positive data?

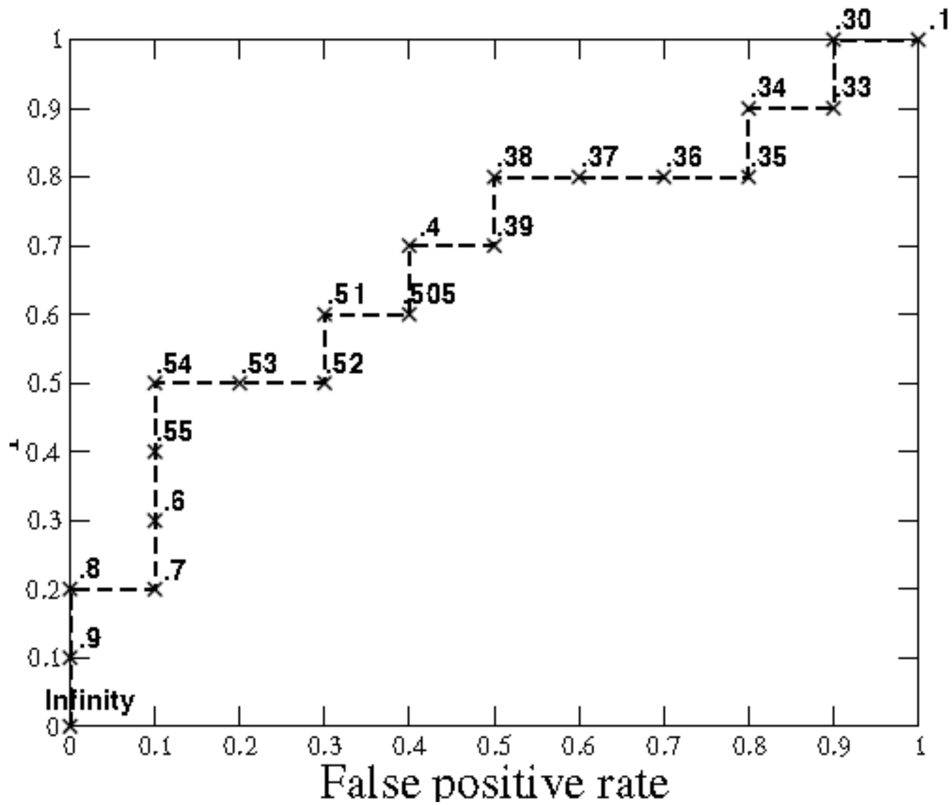


Wikipedia: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC curve: example

(ROC Graphs: Notes and Practical Considerations for Researchers,Tom Fawcett 2004)

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



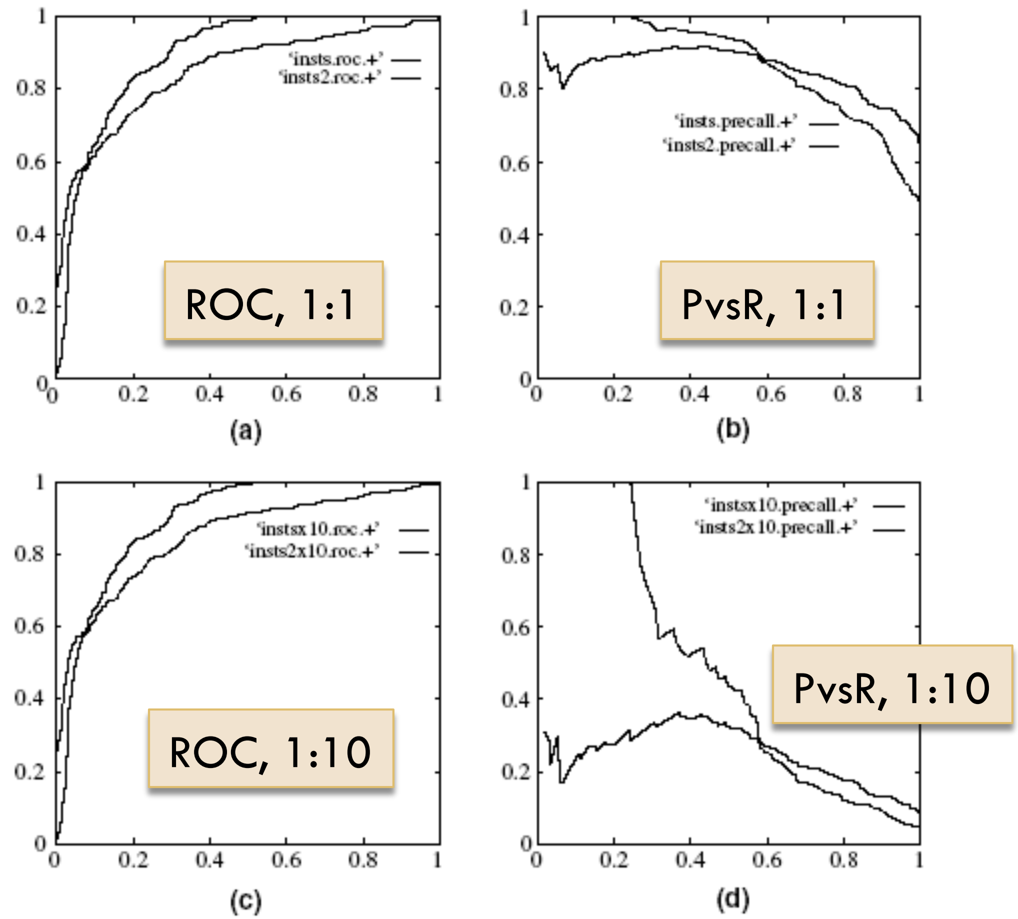
pppppppppppnnnnnnnnnn

nnnnnnnnnnpppppppppp

pnpnpnpnpnpnpnpnpnpn

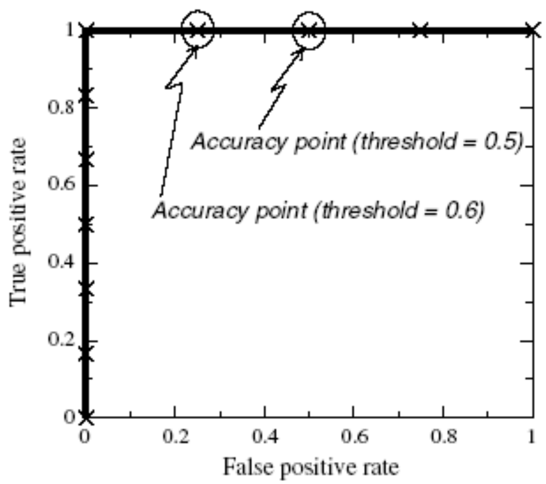
ROC curve -- issue 1

- An attractive property: ROC curves are insensitive to changes in class distribution (*Pattern Recognition letters 2006*)
- TPR and FPR are all strict columnar ratio



ROC curve -- issue 2

- ROC measures the ability of a classifier to produce good relative scores.
  - A good classifier need only produce relative accurate scores that serve to discriminate positive and negative instances



Inst no.	Class True	Class Hyp	Score
1	p	Y	0.99999
2	p	Y	0.99999
3	p	Y	0.99993
4	p	Y	0.99986
5	p	Y	0.99964
6	p	Y	0.99955
7	n	Y	0.68139
8	n	Y	0.50961
9	n	N	0.48880
10	n	N	0.44951

Questions

- Q:What is the relationship between the value of F1 and the break-even point?

- Q: Prove that the F1 is equal to the Dice coefficient of the retrieved and relevant document sets.
  - Dice(X, Y)=2|X∩Y|/|X|+|Y|

Questions

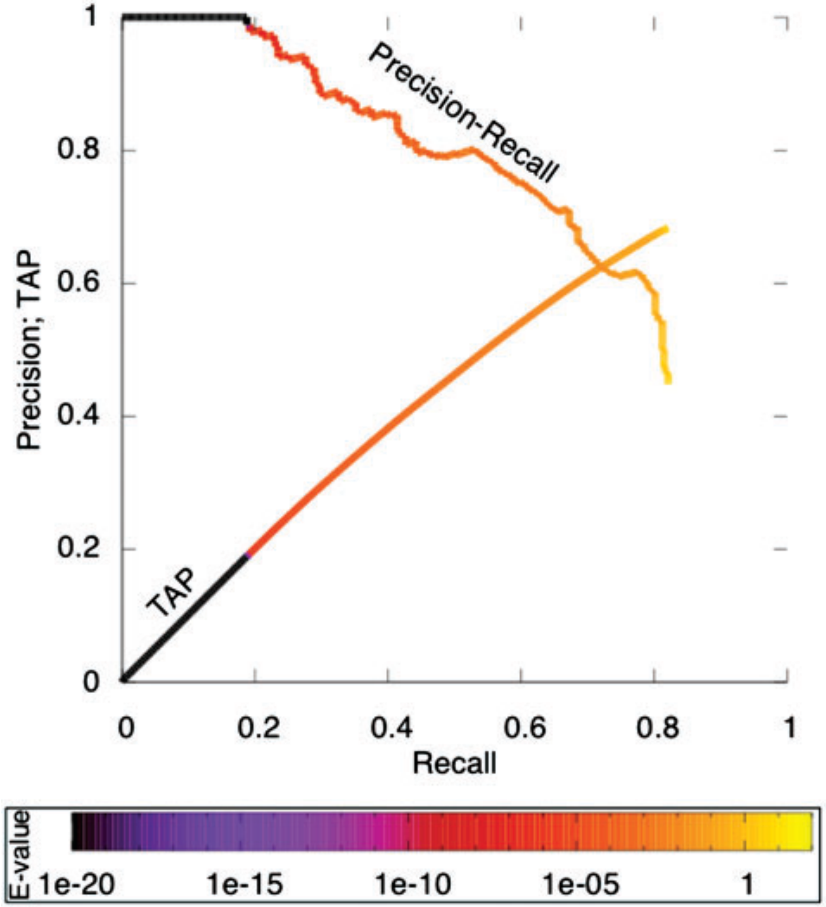
- Q:What is the relationship between the value of F1 and the break-even point?
- A: at break-even point F1=P=R.

- Q: Prove that the F1 is equal to the Dice coefficient of the retrieved and relevant document sets.
  - $Dice(X, Y)=2|X \cap Y|/|X|+|Y|$
- A:
  - $F1=2PR/(P+R)$, $P=tp/(tp+fp)$, $R=tp/(tp+fn) \rightarrow F1=2tp/(2tp+fp+fn)$
  - $|x|=\mathbf{tp}+fp$, $|y|=\mathbf{tp}+fn \rightarrow Dice(x, y)=tp/(2tp+fp+fn)$

TAP-K: Threshold Average Precision

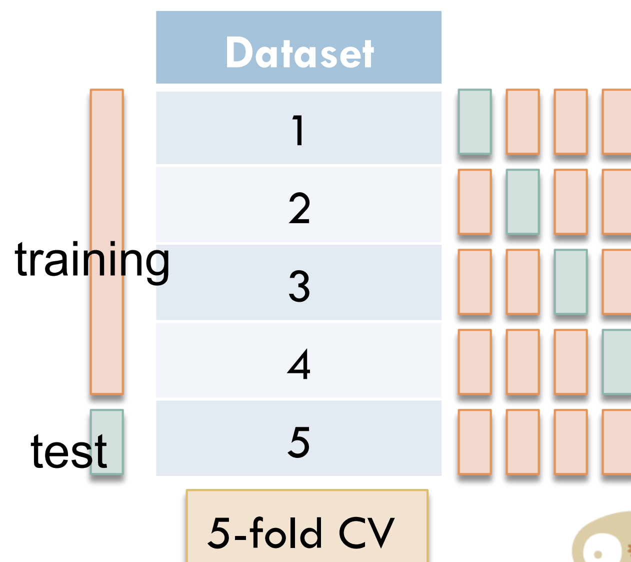
(bioinformatics, 2010 May)

- a measure of retrieval designed for bioinformatics
- ROC_n curve
  - Pooled negative data
- E-value added



Methods of Estimation

- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Random subsampling
  - Repeated holdout
- Cross validation
  - Partition data into k disjoint subsets
  - k-fold: train on k-1 partitions, test on the remaining one
  - Leave-one-out (LOOCV): k=n
- Stratified sampling
  - oversampling vs undersampling
- Bootstrap
  - Sampling with replacement



Test of Significance

- Given two models:
  - Model M1: accuracy = 85%, tested on 30 instances
  - Model M2: accuracy = 75%, tested on 5000 instances

- Can we say M1 is better than M2?
  - How much confidence can we place on accuracy of M1 and M2?
  - Can the difference in performance measure be explained as a result of random fluctuations in the test set?

Need statistically evaluation to compare different models under different tests

How to evaluate a ranked list?

- The ground truth is ranked / partially preferred
- DCG: Discounted cumulative gain
  - Kalervo Jarvelin, ACM TOIS 2002
  - measures the usefulness, or *gain*, of a document based on its position in the result list
- Correlation coefficient measurement
  - Person's Correlation coefficient
  - Kendall-tau correlation coefficient (1938)
  - Cohen's Karpa correlation coefficient (1960)

DCG: Discounted cumulative gain

- measures the usefulness, or *gain*, of a document based on its position in the result list.
- The gain is accumulated cumulatively
  - from the top of the result list to the bottom
  - discounted at lower ranks
- **CG** (cumulative gain) at a particular rank position *p* is defined as $CG_p = \sum_{i=1}^{p} rel_i$
  - rel_i is the graded relevance of the result at position *i*
  - **Independent** with the result order

DCG

- ***Discounted CG*** *at a position p is defined as*

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

  - ***Dependent*** *with the result order*

- DCG without score
  - Use ranks as default scores
  - For example
    - Ground truth ranking: abcde
    - Result ranking: adecb → 52134

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2 (i+1)}$$

Another discounted, score functions

DCG example

- D1, D2, D3, D4, D5 with relevance score **2, 1, 0, 2, 0** (2: highly relevance, 1: relevance, 0: non-relevance)

- DCG_5 of this list = $2 + (1/1 + 0/\log_2 3 + 2/\log_2 4 + 0/\log_2 5)$ $= 2 + 1 + 1 = 4$
- **Ideal order** (2,2,1,0,0 perfect) $IDCG_5 = 2 + 2 + 1/\log_2 3 = 4.63$
- **NDCG=Normalized** $DCG_5 = DCG_5 / IDCG_5 = 4/4.63 = 0.86$

- What are NDCGs of lists (1, 2, 2, 0, 0) and (2, 1, 0, 2, 0) ?

Kendall-tau

- measure the association between two measured quantities
- (#concordant - #discordant) / (n(n+1)/2)
- E.g.,
  - Ground truth :12345, Result list: 21534
    - #concordant = 7, #discordant = 3, Kendall-tau = (7-3)/10 = 0.4
  - Try another list 21345

Discordant pairs:{1,2}, {3,5}, {4,5}

- Sensitive to few bad ranked results
- Compare: Rand Index

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Cohen's Kappa correlation coefficient

- measures the agreement between **two raters** who each classify N items into C mutually exclusive categories

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

  - Pr(a): relative observed agreement among raters
  - Pr(e): the hypothetical probability of chance agreement (random agreement)
  - k=1 complete agreement
  - Pr(e) up, then *k* down
- Change C to fit your application

Cohen's Kappa correlation coefficient

- Agreement Pr(a) = (10+15)/30=0.83
- Pr(e)
  - P(A=Y)=10/30=0.33
  - P(B=Y)=15/30=0.5
  - P(A=Y, B=Y) = 0.33*0.5 = 0.17
  - P(A=N,B=N) = 0.66*0.5 = 0.33
  - → Pr(e) = 0.17 + 0.33 = 0.5
- K = (0.83–0.5) / (1-0.5) = 0.66

		B	
		Y	N
A	Y	10	0
	N	5	15

Poor agreement = Less than 0.20
Fair agreement = 0.20 to 0.40
Moderate agreement = 0.40 to 0.60
Good agreement = 0.60 to 0.80
Very good agreement = 0.80 to 1.00

Cohen's Kappa correlation coefficient

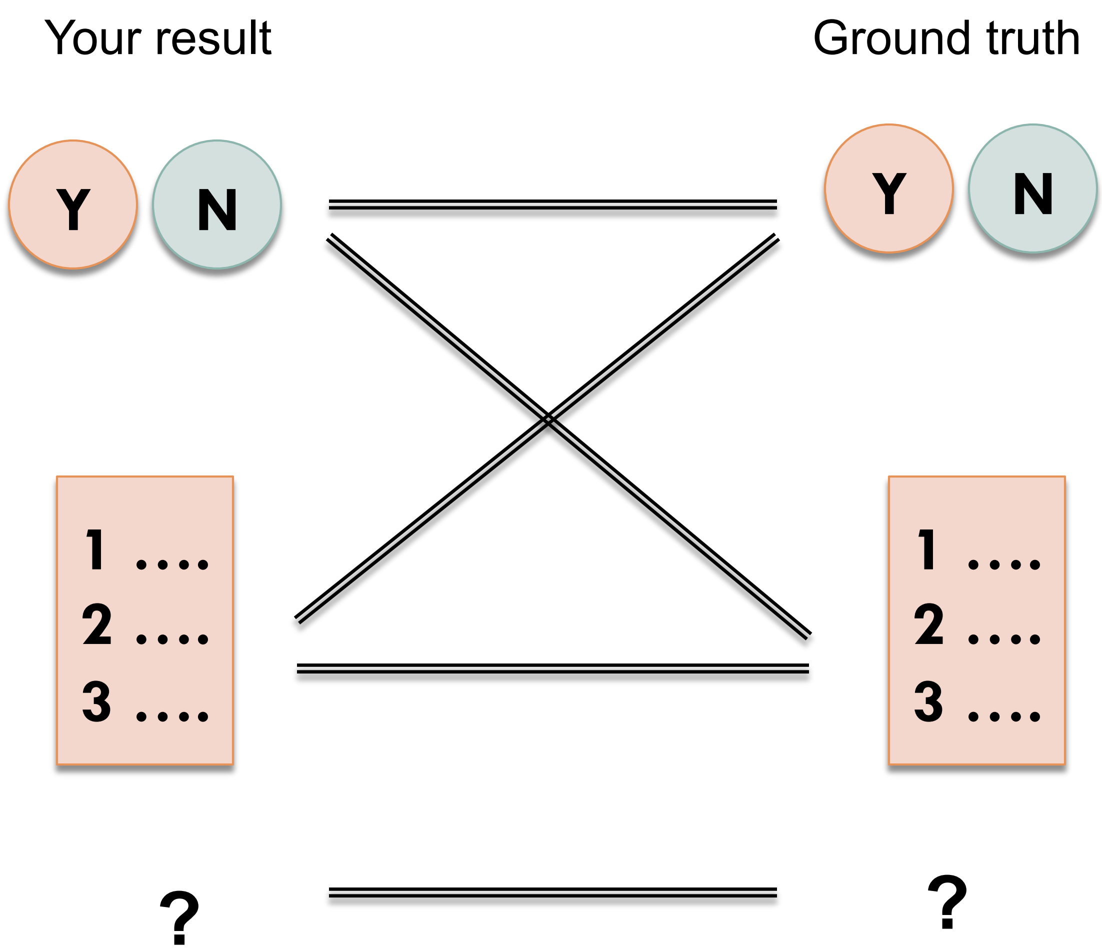
- Inconsistent example

1	Y	N
Y	45	15
N	25	15

2	Y	N
Y	25	35
N	5	35

- $Pr(a) = 0.6$ in two cases
- $Pr_1(e) = 0.54$, $Pr_2(e) = 0.46$
- $k_1 = 0.13$, $k_2 = 0.26$

Applicability



Applicability

- For **each** following evaluation criteria, please **briefly describe ONE** prediction system in which the criterion is important.
- NDCG
- Recall
- Top-1 precision
- F1
- Novelty

REFERENCE COLLECTION

KDD CUP (http://www.kdd.org/kddcup/)

- KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by ACM SIGKDD from 1997.
- Topics: data mining, machine learning, information retrieval / extraction
  - 2019:Transportation recommendation, temporal relational prediction, RL for Malaria
  - 2018: Fresh Air: forecast air quality indices (AQIs) of the future 48 hours
  - 2017: Highway tollgate traffic flow prediction
  - 2016: Given a research field, predict the most influential institutes
  - 2015: Predicting dropouts in MOOCs (1st place $10,000)
  - 2014: Predicting Excitement at DonorsChoose.org (*NLP data inside*)
  - 2013: author classification / prediction from citation (*NLP data inside*)
  - 2012: following prediction / CTR prediction for Ads (largest data)
  - 2011: Music rating prediction
  - 2010: Student performance evaluation
  - 2009: Customer relationship prediction
  - 2008: Breast cancer
  - ….
  - 2002: BioMed document; plus gene role classification

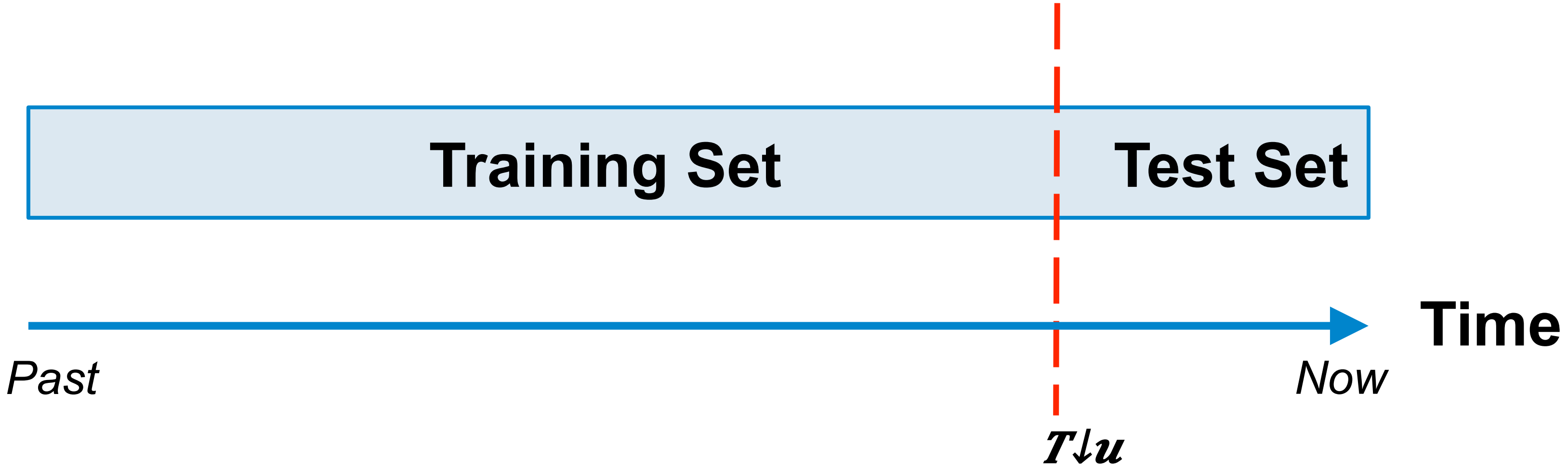
KDDCUP 2011Music Rating Prediction Dataset

- Contains large number of users/items/time data
  - 260 million ratings
  - 1 million users
  - 0.5 million items
  - 8 years

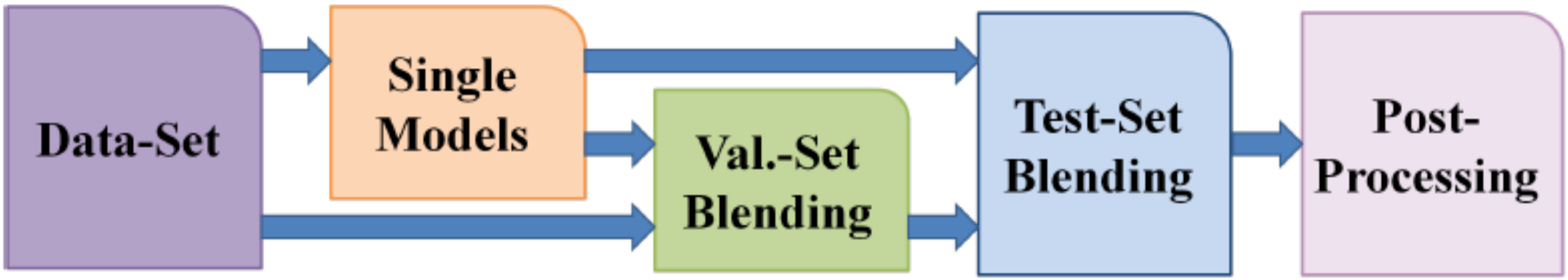
- 4 types of item
  - Genres, Artist, Album, Track

Dataset

- Goal
  - Predict user's ratings in the last four times
  - Predict a item will be rated or not



NTU



model	# used	best	average	worst	contribution
MF	81	22.90	23.92	26.94	**0.3645**
pPCA	2	24.46	24.61	24.75	0.0014
pLSA	7	24.83	25.53	26.09	0.0042
R-Boltz. machine	8	22.80	24.75	26.08	0.0314
k-NN	18	22.79	25.06	42.94	0.0298
regression	10	24.13	28.01	35.14	0.0261

Val.-Set Blending 95

KDDCup 2012

- 50 days data of 2 M active users from 4.25 億微博用戶
- 6 千被推荐用户、3 億條推薦紀錄及其 3 M follow actions
- 70 M training records, 30 M testing records

Kaggle (www.kaggle.com)

- The Home of Data Science
- Prediction problem / competition platform

Active Competitions

Tradeshift Text Classification Classify text blocks in documents	27 days 135 teams $5,000
Africa Soil Property Prediction Challenge Predict physical and chemical properties of soil using spectral measurements	8.0 days 1204 teams $8,000
American Epilepsy Society Seizure Prediction ... Predict seizures in intracranial EEG recordings	34 days 271 teams $25,000
CIFAR-10 - Object Recognition in Images Identify the subject of 60,000 labeled images	5.0 days 224 teams Knowledge
Learning Social Circles in Networks Model friend memberships to multiple circles	14 days 165 teams Knowledge
Sentiment Analysis on Movie Reviews Classify the sentiment of sentences from the Rotten Tomatoes dataset	4 months 480 teams Knowledge

Lessons Learned from the Hunt...
3rd Place interview from the ...
Learning from the best
11th hour win of Greek Media ...
First Place in Allstate Purch...
It's ML Conference Season

215355 players
649190 entries

2014 KDD Cup

Completed • $2,000 • 472 teams

KDD Cup 2014 - Predicting Excitement at DonorsChoose.org

Thu 15 May 2014 – Tue 15 Jul 2014 (3 months ago)

Dashboard

Private Leaderboard - KDD Cup 2014 - Predicting Excitement at DonorsChoose.org

This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts? Let us know.

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	'STRAYA *	0.67814	213	Tue, 15 Jul 2014 00:21:34 (-0.2h)
2	↓1	DataRobot *	0.67320	220	Tue, 15 Jul 2014 23:32:50 (-2d)
3	↑30	ChaoticExperiments (KIRAN R) *	0.67297	69	Tue, 15 Jul 2014 19:35:05 (-2d)
4	↓1	dkay & bmax & James King	0.66473	239	Tue, 15 Jul 2014 23:26:11 (-2.1d)
5	↓1	Triskelion,Yan, KazAnova & Shize	0.65949	225	Tue, 15 Jul 2014 23:29:42 (-0.4h)
6	↑35	Giulio, orchid, Luca & Ben	0.65919	264	Tue, 15 Jul 2014 18:51:21 (-0.4h)
7	↑2	:-)	0.65372	123	Tue, 15 Jul 2014 22:41:25 (-4d)

UCI Data Repository

- UC Irvine Machine Learning Repository
  - http://archive.ics.uci.edu/ml/
  - https://www.kaggle.com/uciml
- 351 datasets
- Famous datasets
  - Iris: 1105860 hits
  - Adult: 766735 hits
  - Wine: 584298 hits