Metrics, Statistics, Tests

Tetsuya Sakai

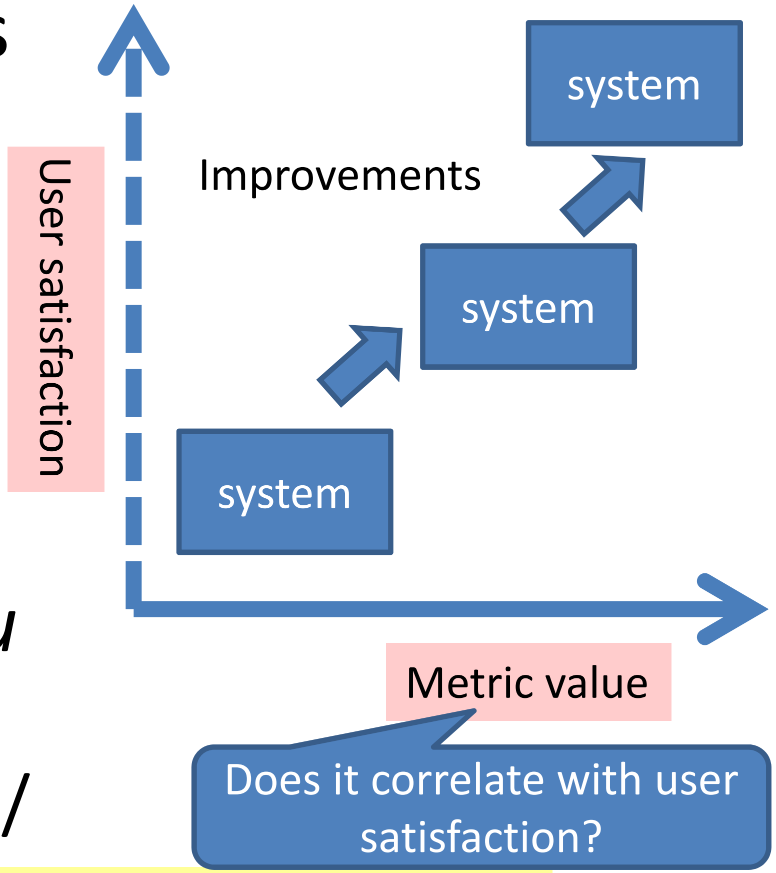
Microsoft Research Asia, P. R. China

@tetsuyasakai

February 6, 2013@PROMISE Winter School 2013 in Bressanone, Italy

Why measure?

- IR researchers’ goal: build systems that satisfy the user’s information needs.
- We cannot ask users all the time, so we need metrics as surrogates of user satisfaction/performance.
- *“If you cannot measure it, you cannot improve it.”* http://zapatopi.net/kelvin/quotes/



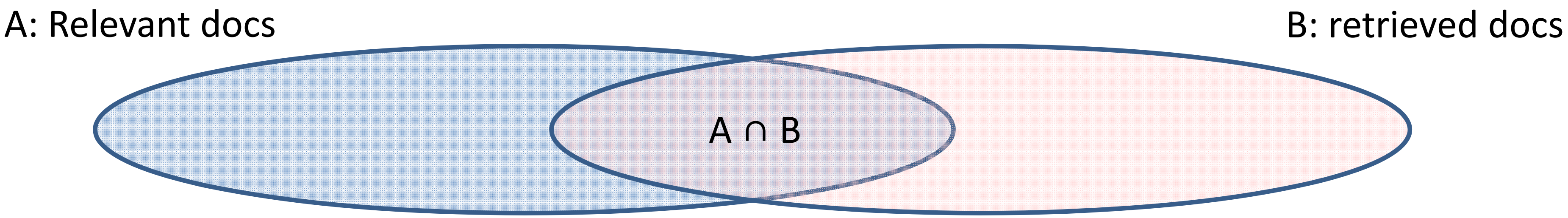
Does it correlate with user satisfaction?

An interesting read on IR evaluation: [Armstrong+CIKM09]
Improvements that don't add up: ad-hoc retrieval results since 1998

LECTURE OUTLINE

1. Traditional IR metrics
   - Set retrieval metrics
   - Ranked retrieval metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary

Do you recall recall and precision from Dr. Ian Soboroff’s lecture?



- E-measure $= (|A \cup B| - |A \cap B|)/(|A|+|B|)$

$= 1 - 1/(0.5*(1/Prec) + 0.5*(1/Rec))$

where $Prec=|A \cap B|/|B|$, $Rec=|A \cap B|/|A|$.

A generalised form

$= 1 - 1/(\alpha*(1/Prec) + (1-\alpha)*(1/Rec))$

$= 1 - (\beta^2 + 1)*Prec*Rec/(\beta^2 *Prec+Rec)$

where $\alpha = 1/(\beta^2 + 1)$. See [vanRijsbergen79].

F-measure [Chinchor MUC92]

- Used at the 4th Message Understanding Conference; much more widely used than E
- F-measure = 1 – E-measure

$= 1/(\alpha*(1/Prec) + (1-\alpha)*(1/Rec))$

$= (\beta^2 + 1)*Prec*Rec/(\beta^2 *Prec+Rec)$

where $\alpha = 1/(\beta^2 + 1)$.

- F with β=b is often expressed as F_b.
- $F_1 = 2*Prec*Rec/(Prec+Rec)$

i.e. harmonic mean of Prec and Rec

User attaches β times as much importance to Rec as Prec ($dE/dRec=dE/dPrec$ when $Prec/Rec=\beta$) [vanRijsbergen79]

LECTURE OUTLINE

1. Traditional IR metrics
   - Set retrieval metrics
   - Ranked retrieval metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary

Normalised Discounted Cumulative Gain [Jarvelin+TOIS02]

- Introduced at SIGIR2000, a variant of Pollack's sliding ratio [Pollack AD68; Korfhage97]
- Popular "Microsoft" version [Burges+ICML05]:

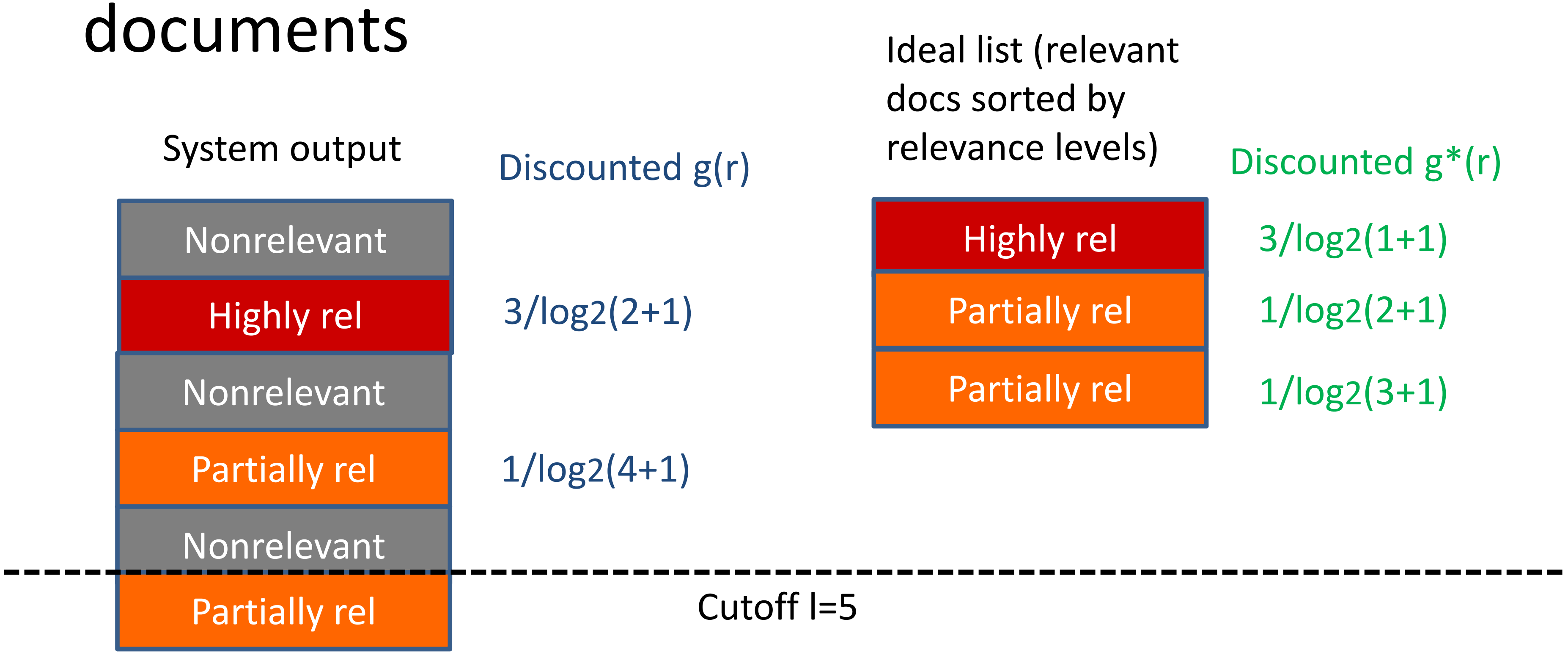
$$\mathrm{nDCG@}l = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)}$$

l: document cutoff (e.g. 10)
r: document rank
g(r): gain value at rank r
e.g. 1 if doc is partially relevant
3 if doc is highly relevant
g*(r) gain value at rank r of an ideal ranked list

Original Jarvelin/Kekalainen definition not recommended: a system that returns a relevant document at rank 1 and one that returns a relevant document at rank b are treated as equally effective, where b is the logarithm base (patience parameter). b's cancel out in the Burges definition.

nDCG: an example

Evaluating a ranked list at l=5 for a topic with 1 highly relevant and 2 partially relevant documents



System output

- Nonrelevant
- Highly rel — 3/log2(2+1)
- Nonrelevant
- Partially rel — 1/log2(4+1)
- Nonrelevant
- Partially rel

Discounted g(r)

Ideal list (relevant docs sorted by relevance levels)

- Highly rel — 3/log2(1+1)
- Partially rel — 1/log2(2+1)
- Partially rel — 1/log2(3+1)

Discounted g*(r)

Cutoff l=5

nDCG@5= 2.3235/4.1309 = 0.5625

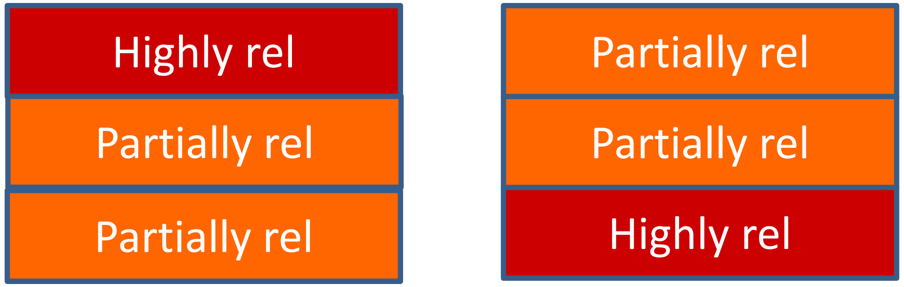
Average Precision

- Introduced at TREC (1992～), implemented in trec_eval by Buckley
- Like Prec and Rec,

cannot handle

graded relevance

Equally effective?



Highly rel
Partially rel
Partially rel

Partially rel
Partially rel
Highly rel

$AP=(1/R)\sum_r I(r)Prec(r)$

where $Prec(r)=rel(r)/r$

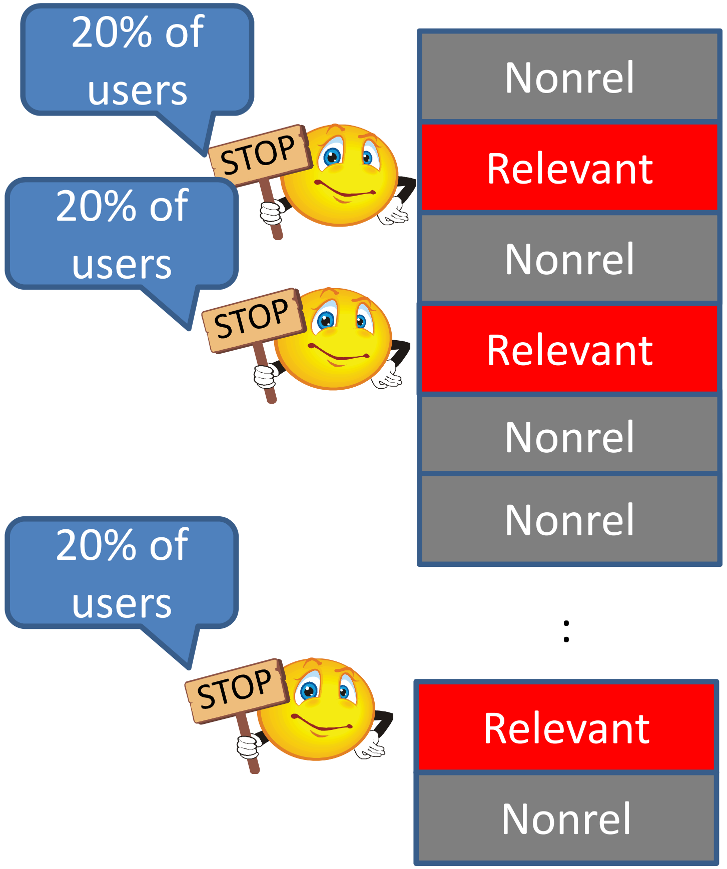
R: total number of relevant docs
I(r): flag indicating a relevant doc
rel(r): number of relevant docs within ranks [1,r]

11-point average precision (average over interpolated precision at recall=0, 0.1, ..,1) not recommended for precision oriented tasks, as it lacks the top heaviness of AP. A top heavy metric emphasises the top ranked documents.

User model for AP [Robertson SIGIR08]

- Different users stop scanning the ranked list at different ranks. They only stop at a relevant document.
- The user distribution is uniform across all (R) relevant documents.
- At each stopping point, compute utility (Prec).
- Hence AP is the expected utility for the user population.

Ranked list for a topic with R=5 relevant documents



Non-uniform stopping distributions have been investigated in [Sakai+EVIA08] .

Q-measure

[Sakai IPM07; Sakai+EVIA08]

- A graded relevance version of AP (see also Graded AP [Robertson+SIGIR10; Sakai+SIGIR11]).
- Same user model as AP, but the utility is computed using the blended ratio BR(r) instead of Prec(r).

$Q=(1/R)\ \Sigma_r\ I(r)BR(r)$

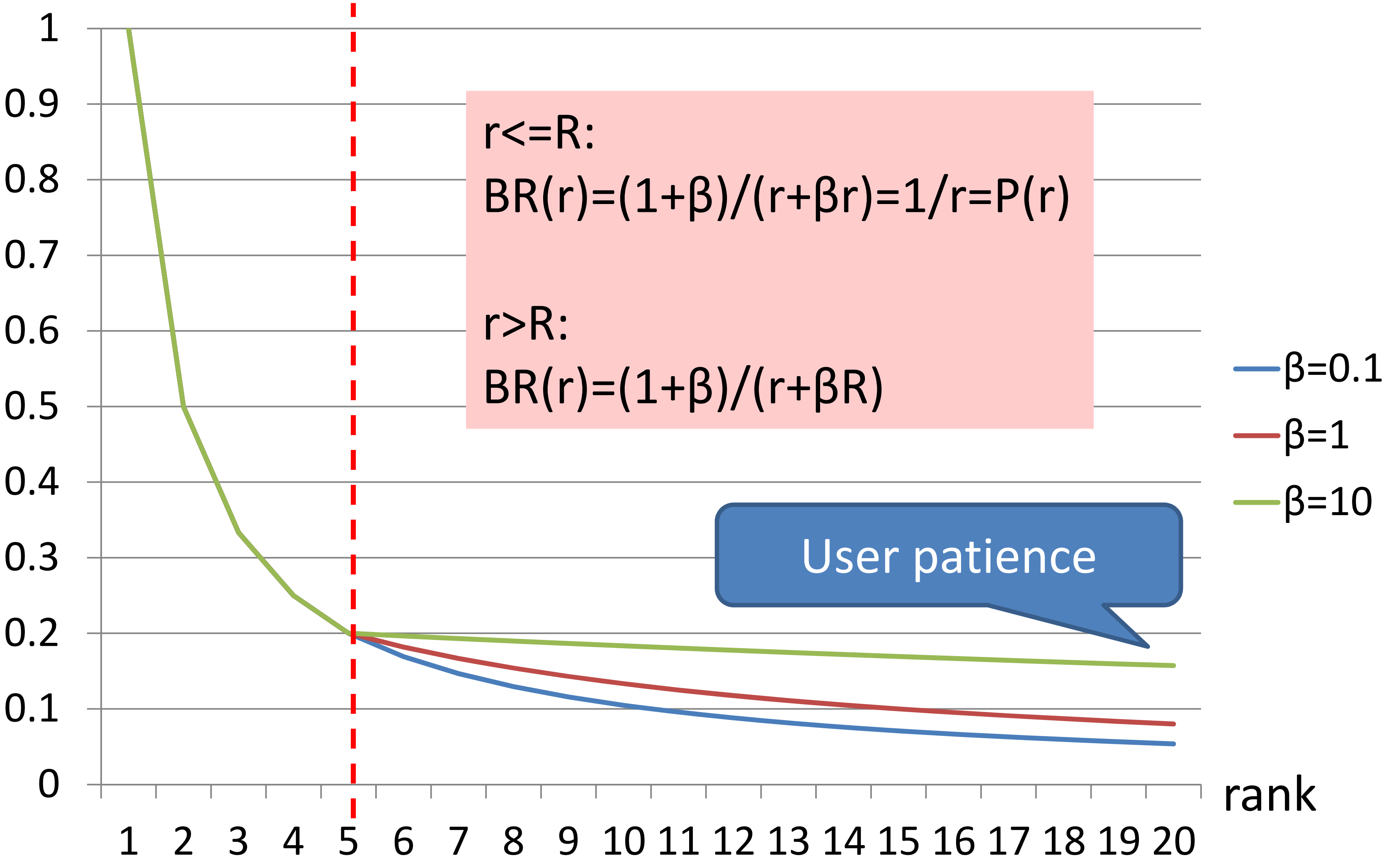
where BR(r)

$=(\ rel(r) + \beta\ \Sigma^r_{k=1} g(k)\ )/(\ r + \beta\ \Sigma^r_{k=1} g^*(k)\ )$

β: patience parameter
(when β=0, BR=Prec, hence Q=AP;
when β is large, Q is tolerant to rel docs retrieved at low ranks)

Combines Precision and normalised cumulative gain (nCG) [Jarvelin+TOIS02]

Value of the first relevant document at rank r according to BR(r) (binary relevance, R=5)



r<=R:
BR(r)=(1+β)/(r+βr)=1/r=P(r)

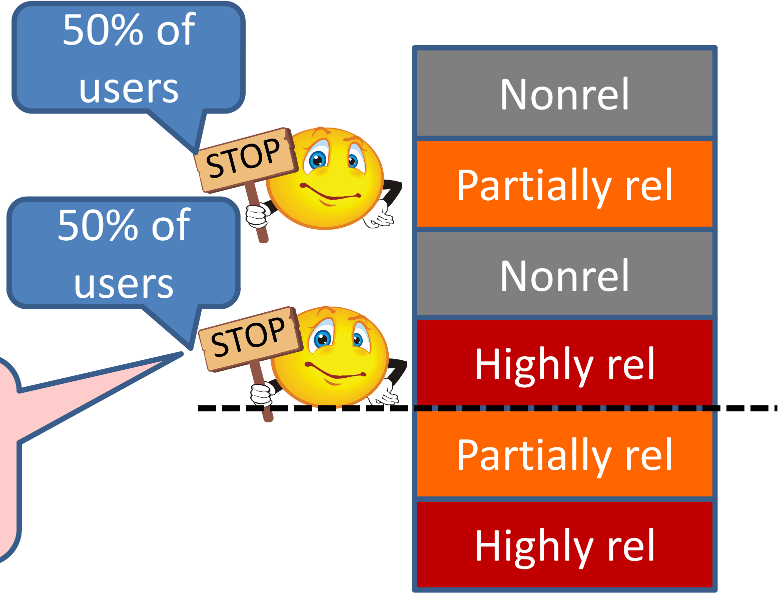
r>R:
BR(r)=(1+β)/(r+βR)

User patience

P+ [Sakai AIRS06; Sakai WWW12]

- Most IR metrics are for informational search intents (user wants as may relevant docs as possible), but P+ is suitable for navigational intents (user wants just one very good doc).
- Same as Q, except that the user distribution is uniform across rel docs above the preferred rank rp, not all rel docs.

$$P+ = (1/\mathrm{rel}(rp)) \sum_{r=1}^{rp} I(r)BR(r)$$



Preferred rank: rank of the most relevant doc in the list that is closest to the top.
In this example, rp=4.

Expected Reciprocal Rank

[Chapelle+CIKM09; Chapelle+IRJ11]

Also quite suitable for navigational intents, as it has the diminishing return property, i.e. whenever a relevant doc is found, the value of a new relevant doc is discounted.

$$\mathrm{ERR} = \Sigma_r \underbrace{\mathrm{dsat}(r-1)\ \mathrm{Pr}(r)}_{\text{Probability that the user is finally satisfied at r}}\ \underbrace{(1/r)}_{\text{Utility at r}}$$

where

$$\mathrm{dsat}(r) = \Pi^{r}_{k=1}(1-\mathrm{Pr}(k))$$

Pr(r): probability that doc at rank r is relevant
≒ prob that the user is satisfied with doc at r
dsat(r): prob that the user is dissatisfied with docs [1,r]

Pr(r) could be set based on gain values
e.g. 1/4 for partially relevant; 3/4 for highly relevant

Rank-Biased Precision [Moffat+TOIS08]

- Moffat and Zobel argue that recall shouldn't be used: RBP is precision that considers ranks
- RBP does not range fully between [0,1]

e.g. When R=10 and p=.95, the RBP for a best possible ranked list is only .4013 [Sakai+IRJ08].

- User model: after examining doc at rank r, will examine next doc with probability p or stop with probability 1-p. Unlike ERR, disregards doc relevance.

$$RBP = (1-p)\ \Sigma_r\ p^{r-1}\ g(r)/gain(H)$$

gain(H): gain for the highest relevance level H (e.g. 3 for highly relevant)

Time-Biased Gain [Smucker SIGIR12]

- Instead of document ranks, TBG uses time to reach rank r for discounting the information value.
- TBG has the diminishing return property.

TBG in [Smucker SIGIR12] is binary-relevance-based, with parameters estimated from a user study and a query log:

$$\text{TBG} = \Sigma_r\ I(r) * \underbrace{.4928}_{\text{Gain of a relevant doc}} * \underbrace{\exp(-T(r)\ \ln 2/224\ )}_{\text{Decay function where h=224 is its half life}}$$

where T(r) is the estimated time to reach r

$$= \Sigma_{m=1}^{r-1}\ \underbrace{4.4}_{\text{Time to read a snippet}} + \underbrace{(0.018\ lm + 7.8) * Pclick(m)}_{\text{Time to read a document of length lm}}$$

(Pclick=.64 if relevant, .39 otherwise)

Traditional ranked retrieval metrics summary

	AP	nDCG	Q	P+	ERR	RBP	TBG
Graded relevance	☹	😃	😃	😃	😃	😃	☹
Intent type	Inf	Inf	Inf	Nav	Nav	Inf	Inf
Normalised	YES	YES (nDCG) NO (DCG)	YES	YES	NO (ERR) YES (nERR)	NO	NO
User model	😃	☹	😃	😃	😃	😃	😃
Diminishing return	☹	☹	☹	☹	😃	☹	😃
Document length	☹	☹	☹	☹	☹	☹	😃
Discriminative power	😃	😃	😃	☹	☹	☹	☹

Discriminative power will be explained later

Normalisation and averaging

- Usually an arithmetic mean over a topic set is used to compare systems e.g. AP->Mean AP (MAP)
- Normalising a metric before averaging implies that every topic is of equal importance, no matter how R varies
- Not normalising implies that every user effort (e.g. finding one relevant document) is of equal importance – but topics with large R will dominate the mean, and different topics will have different upperbounds
- Alternatives: median, geometric mean (equivalent to taking the log of the metric and then averaging) to emphasise the lower end of the metric scale
e.g. GMAP [Robertson CIKM06]

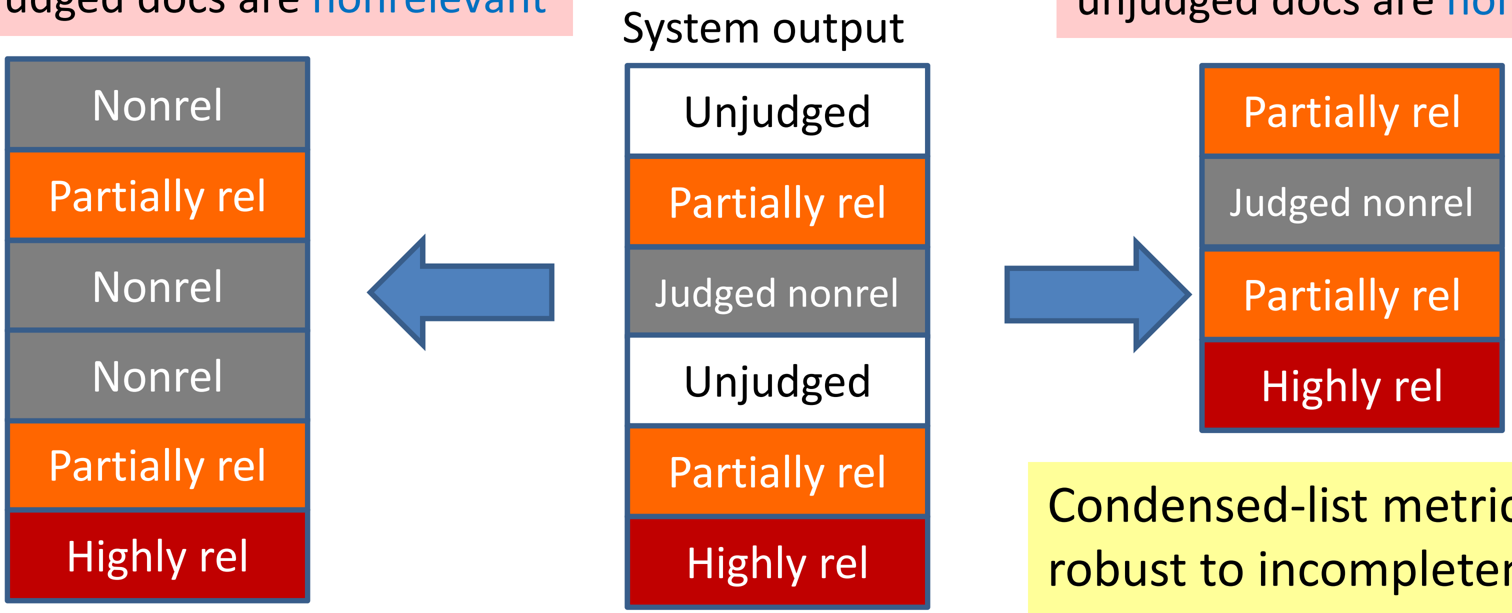
Condensed-list metrics

[Sakai SIGIR07; Sakai CIKM08; Sakai+IRJ08]

Modern test collections rely on pooling: we have many unjudged docs, not just judged nonrelevant docs
i.e. relevance assessments are incomplete

Standard evaluation: assume unjudged docs are nonrelevant

Condensed-list evaluation: assume unjudged docs are nonexistent



System output

Unjudged
Partially rel
Judged nonrel
Unjudged
Partially rel
Highly rel

Nonrel
Partially rel
Nonrel
Nonrel
Partially rel
Highly rel

Partially rel
Judged nonrel
Partially rel
Highly rel

Condensed-list metrics are more robust to incompleteness than standard metrics.

But condensed-list metrics overestimate systems that did not contribute to the pool, while standard metrics underestimate them [Sakai CIKM08; Sakai+AIRS12a]

"Binary Preference" was probably the first condensed-list metric in the literature but…

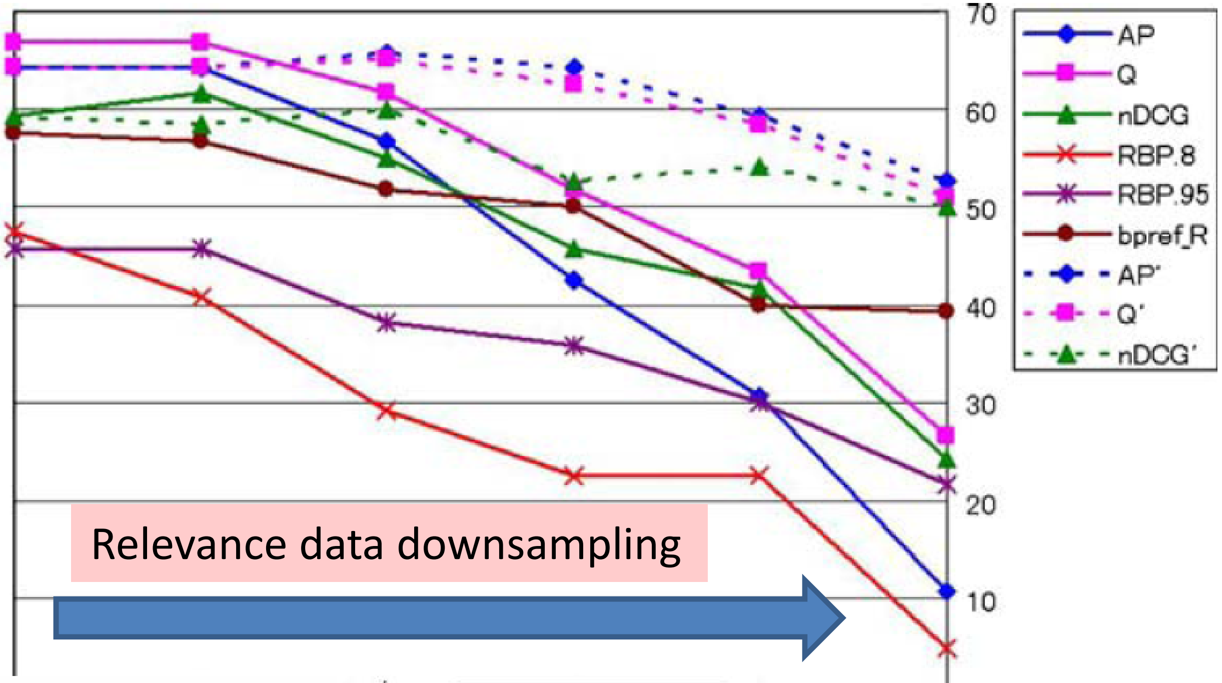
- [Buckley+SIGIR04] proposed bpref, which is in fact a variant of condensed-list Average Precision. It lacks the top heaviness of AP and is less robust to incompleteness.
  See [Sakai SIGIR07; Sakai +IRJ08].
- [Buttcher+SIGIR07] used Ahlgren/Gronqvist RankEff but this metric is in fact a known variant of bpref called bpref_N (bpref_allnonrel in trec_eval). See [Sakai CIKM08].
- Hence bpref and bpref_N are not recommended.

More on handling incomplete and biased relevance assessments:
[Yilmaz+CIKM06] [Aslam+CIKM07] [Carterette SIGIR07] [Webber+SIGIR09]..

[Sakai+IRJ08]

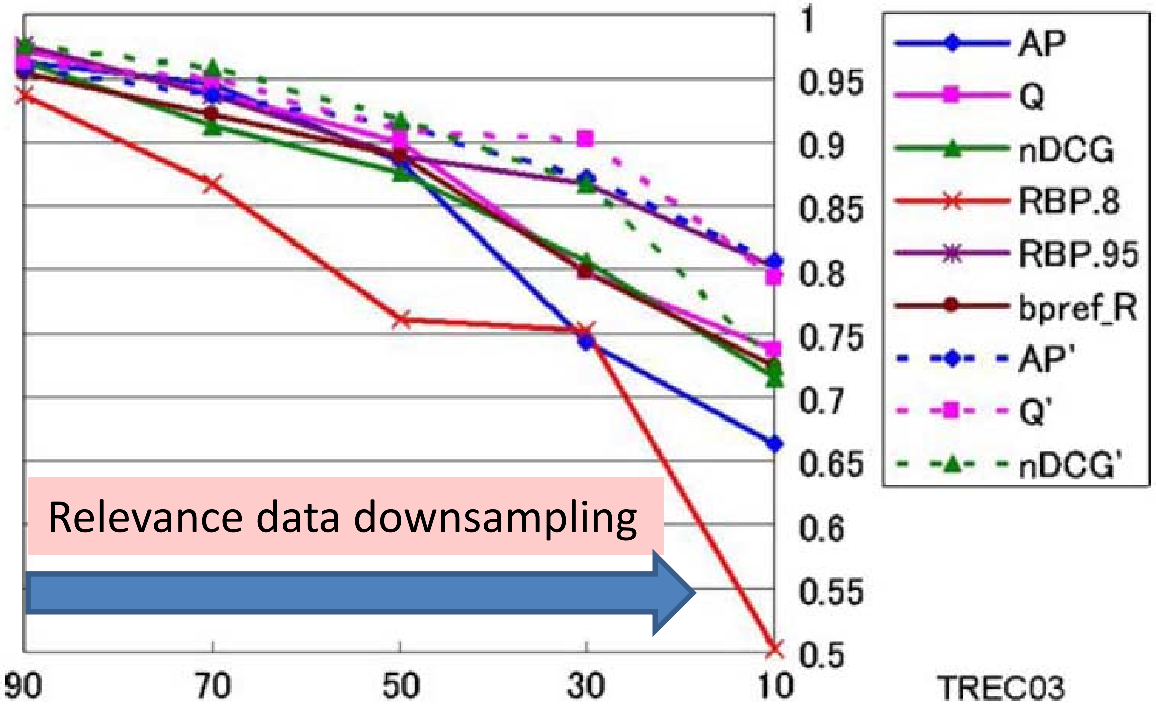
Condensed-list versions of AP, Q, nDCG (AP’, Q’, nDCG’) are relatively robust to incompleteness

Discriminative power (number of significant differences obtained)



Relevance data downsampling

Rank correlation with system ranking based on full relevance data



Relevance data downsampling

Condensed-list AP (AP’) is also known as Induced AP [Yilmaz+CIKM06]

LECTURE OUTLINE

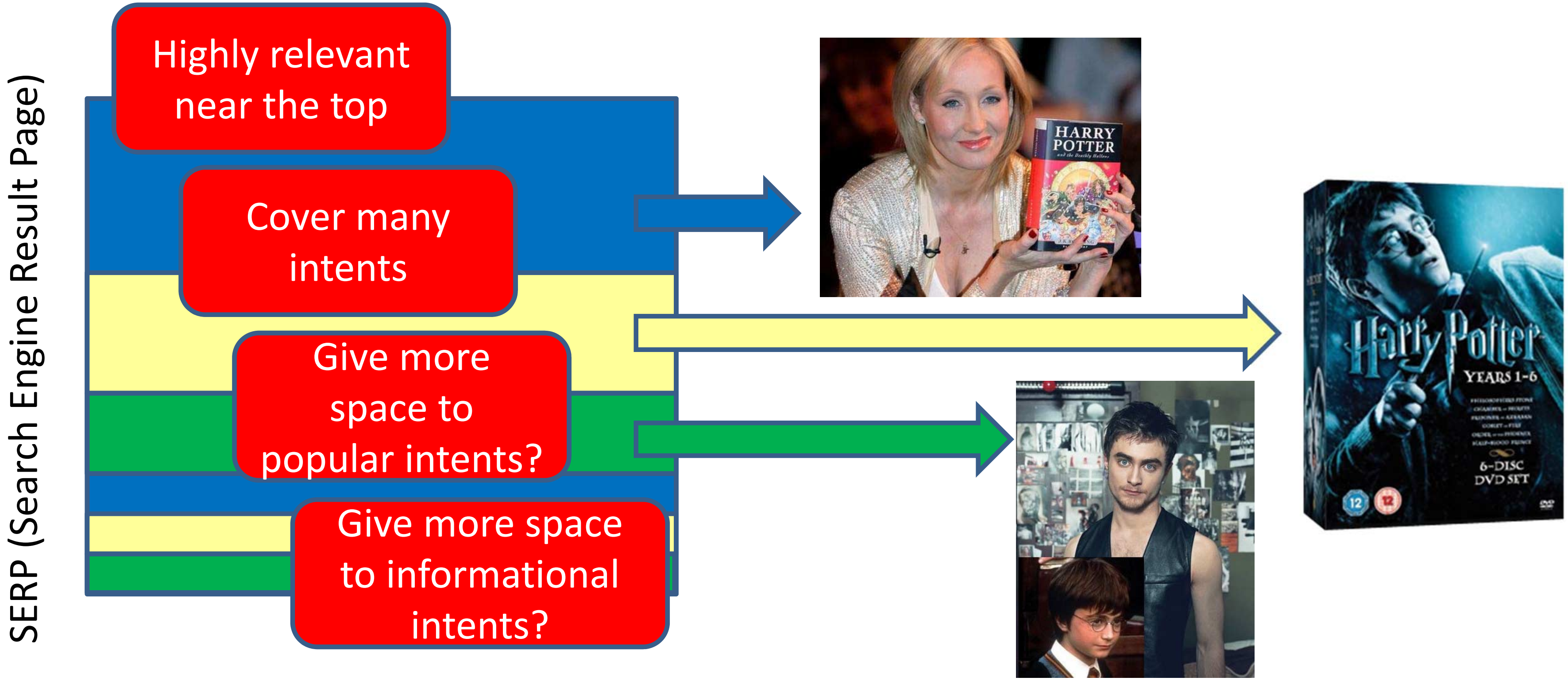
1. Traditional IR metrics
2. Advanced IR metrics

- Diversified search metrics
- Session, summarisation and QA metrics

3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary

Diversified search

- Given an ambiguous/underspecified query, produce a single Search Engine Result Page that satisfies different user intents!
- Challenge: balancing relevance and diversity



SERP (Search Engine Result Page)

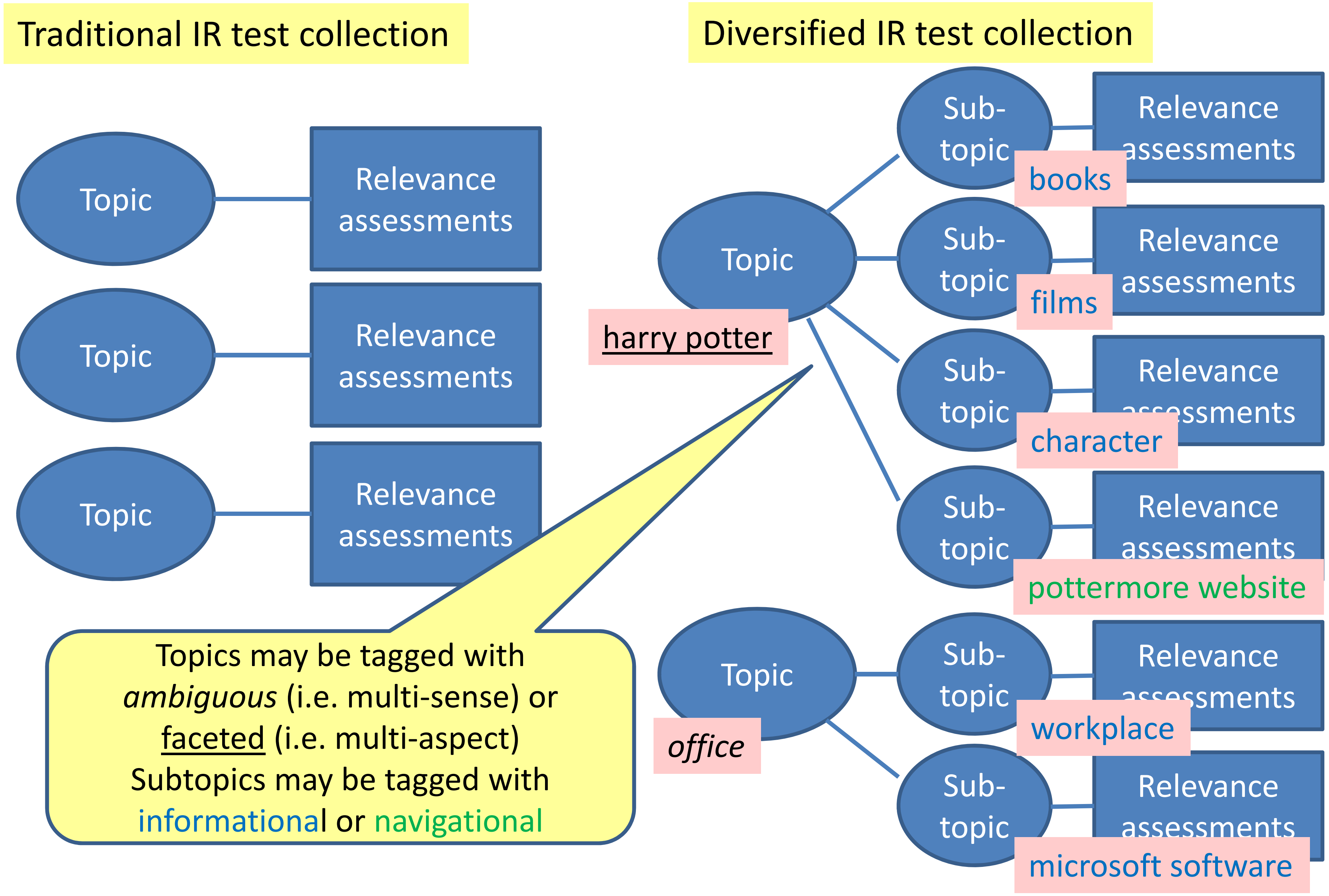
Highly relevant near the top

Cover many intents

Give more space to popular intents?

Give more space to informational intents?

Diversified search test collections



Traditional IR test collection

Topic — Relevance assessments

Topic — Relevance assessments

Topic — Relevance assessments

Diversified IR test collection

Topic (harry potter)

Sub-topic — Relevance assessments (books)

Sub-topic — Relevance assessments (films)

Sub-topic — Relevance assessments (character)

Sub-topic — Relevance assessments (pottermore website)

Topic (*office*)

Sub-topic — Relevance assessments (workplace)

Sub-topic — Relevance assessments (microsoft software)

Topics may be tagged with *ambiguous* (i.e. multi-sense) or faceted (i.e. multi-aspect)
Subtopics may be tagged with informational or navigational

α-nDCG

[Clarke+SIGIR08; Clarke+WSDM11]

- Replaces the gain of nDCG by novelty-biased gain

$$ng(r) = \Sigma_{i=1}^{m} I_i(r)\ (1-\alpha)^{rel_i(r-1)}$$

m: number of “nuggets” (intents)
$I_i(r)$: relevance flag for i-th nugget
α: probability that user “finds” a nonexistent nugget in doc
$rel_i(r)$: number of docs relevant to i-th nugget in [1,r]

Graded relevance of a doc = number of nuggets covered by doc (Cannot handle graded relevance assessments)

Discounts gain based on relevant information already seen (diminishing return) e.g. α=.5
If doc at r=1 is nonrelevant to i, discount factor for r=2 is (1-0.5)^0=1 .
If doc at r=1 is relevant to i, it’s (1-0.5)^1=0.5.

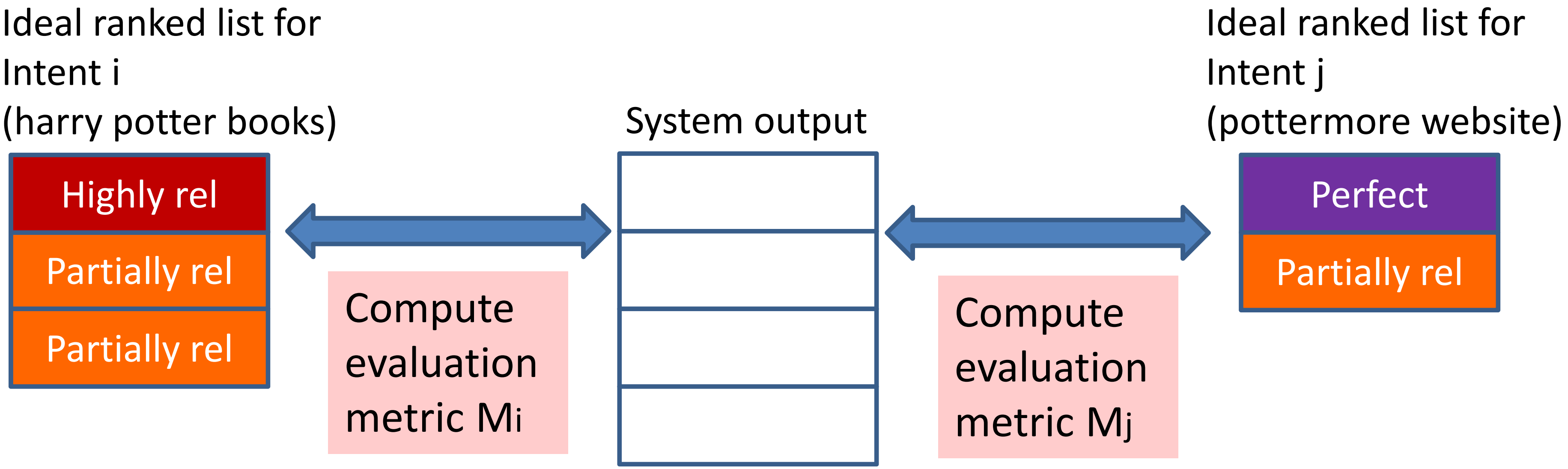
But probability that user misses an existing nugget in doc is 0…

Used at the TREC web track diversity task

Intent-Aware metrics

[Agrawal+WSDM09; Chapelle+IRJ11]

harry potter



Ideal ranked list for Intent i (harry potter books)

System output

Ideal ranked list for Intent j (pottermore website)

Compute evaluation metric M_i

Compute evaluation metric M_j

$$\text{M-IA} = P(i|q)M_i + P(j|q)M_j$$

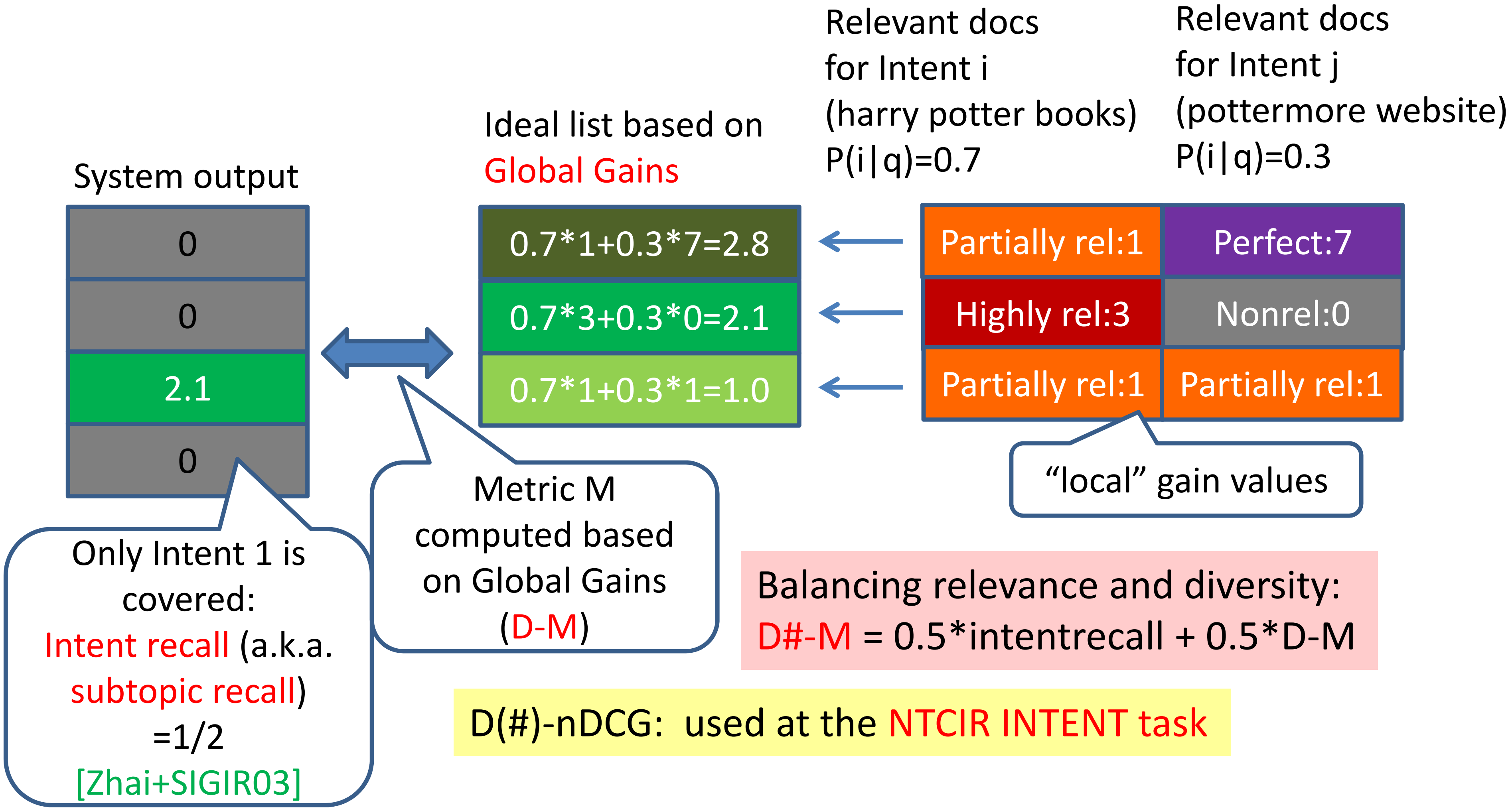
where $P(\cdot|q)$ is the intent probability (popularity)

ERR-IA: used at the TREC web track diversity task

D-measures

[Sakai+SIGIR11; Sakai+IRJ13]

harry potter



System output

0
0
2.1
0

Ideal list based on Global Gains

0.7*1+0.3*7=2.8
0.7*3+0.3*0=2.1
0.7*1+0.3*1=1.0

Relevant docs for Intent i (harry potter books) P(i|q)=0.7	Relevant docs for Intent j (pottermore website) P(i|q)=0.3
Partially rel:1	Perfect:7
Highly rel:3	Nonrel:0
Partially rel:1	Partially rel:1

“local” gain values

Metric M computed based on Global Gains (D-M)

Only Intent 1 is covered: Intent recall (a.k.a. subtopic recall) =1/2 [Zhai+SIGIR03]

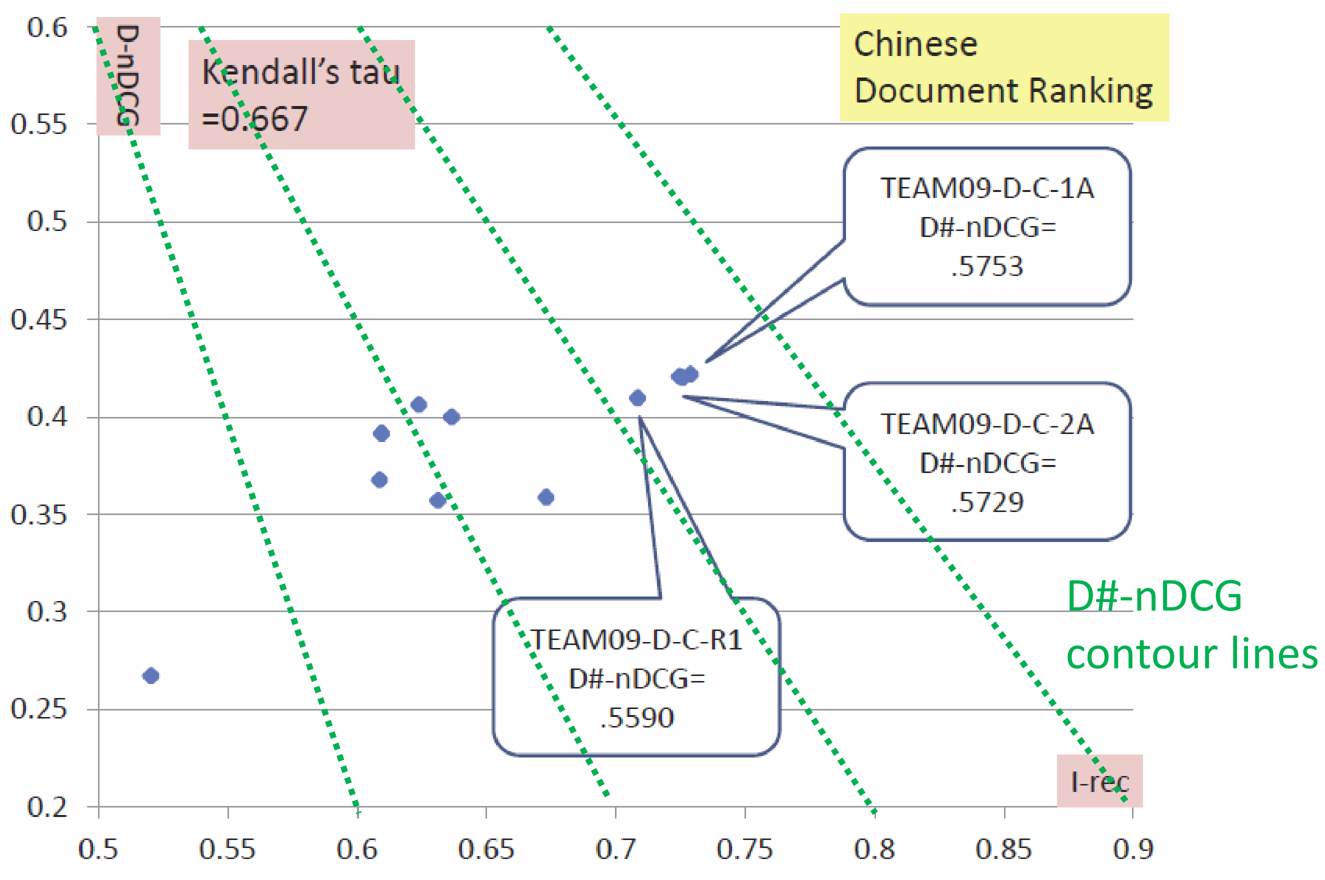
Balancing relevance and diversity:
D#-M = 0.5*intentrecall + 0.5*D-M

D(#)-nDCG: used at the NTCIR INTENT task

D#-nDCG at work

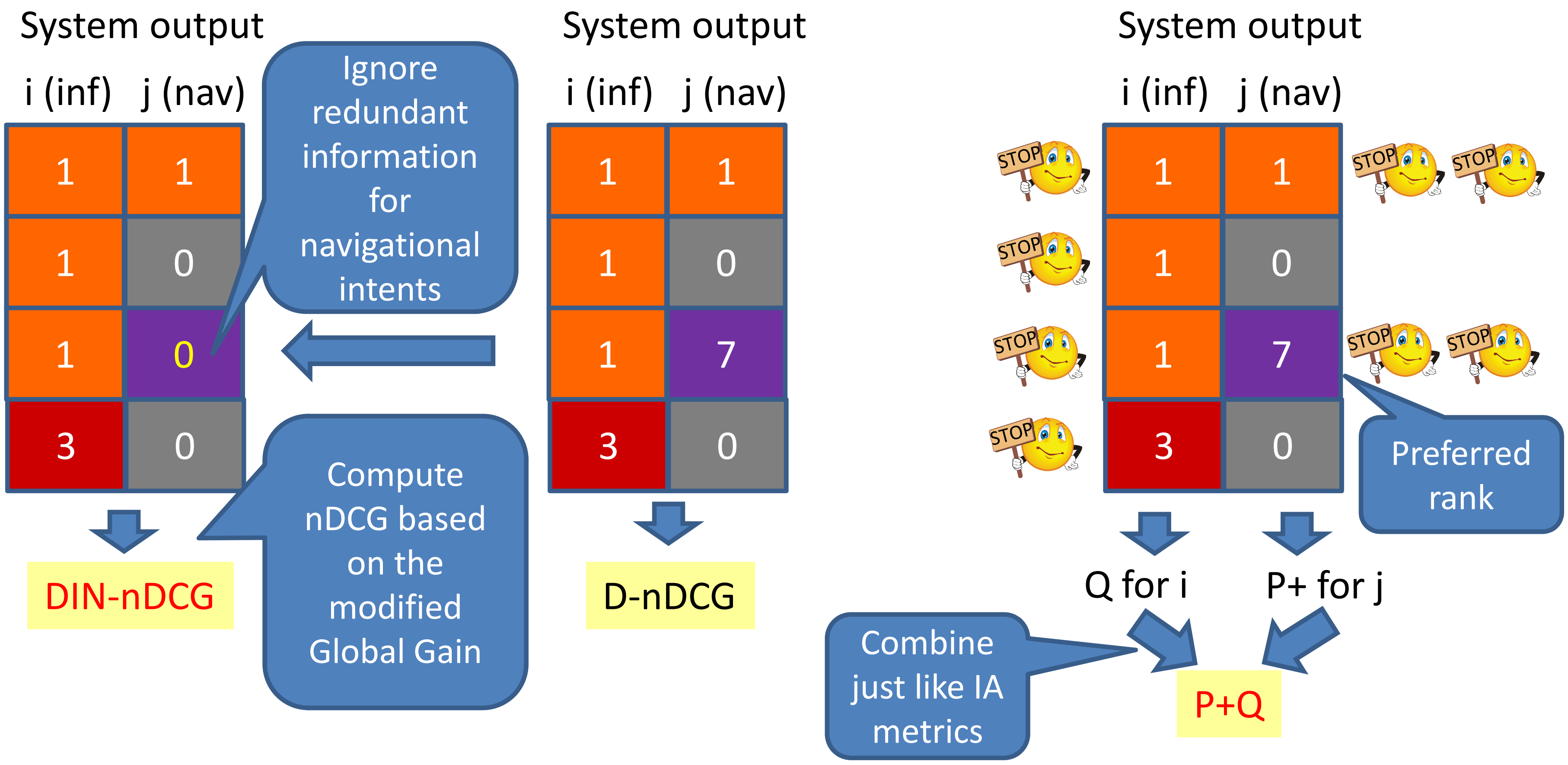
Example from the NTCIR-10 INTENT-2 task

(to be concluded at the NTCIR-10 conference in June 2013)



DIN-nDCG and P+Q [Sakai WWW12]

Unlike α-nDCG, IA metrics and D-measures, considers whether each intent is informational or navigational (do not reward redundant information for nav intents).



System output

i (inf)	j (nav)
1	1
1	0
1	0
3	0

DIN-nDCG

Ignore redundant information for navigational intents

Compute nDCG based on the modified Global Gain

System output

i (inf)	j (nav)
1	1
1	0
1	7
3	0

D-nDCG

System output

i (inf)	j (nav)
1	1
1	0
1	7
3	0

Preferred rank

Q for i

P+ for j

Combine just like IA metrics

P+Q

Diversity metrics summary

[Sakai+SIGIR11; Sakai WWW12; Sakai+IRJ13]

	α-nDCG	IA metrics	D#	DIN#	P+Q#
Graded relevance	😟	😄	😄	😄	😄
Computational complexity	😟	😄	😄	😄	😄
Maximum value is 1	😟	😟	😄	😟	😟
Intent popularity	😮 [Clarke+ WSDM11]	😄	😄	😄	😄
Informational/ navigational	😟	😟	😟	😄	😄
Discriminative power	😄	😟	😄	😄	😄
Concordance test	😟	😟	😄	😄	😄

Discriminative power and concordance test will be explained later

LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics

- Diversified search metrics
- Session, summarisation and QA metrics

3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary

Session DCG

[Jarvelin+ECIR08; Kanoulas+ SIGIR11]

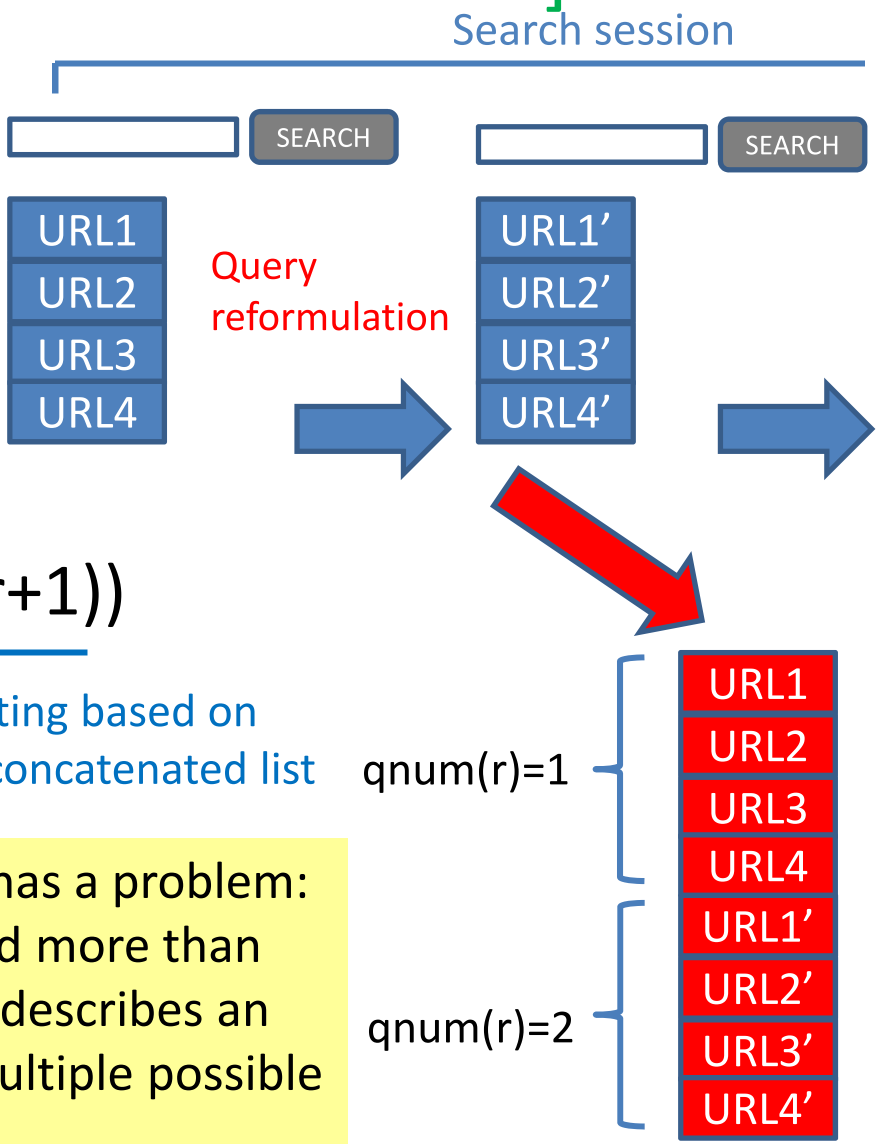
Extending DCG to multiple ranked lists: concatenate top l docs of m ranked lists in a session and compute

sDCG=

$$\sum_{r=1}^{m*l} g(r)/(\underline{\log_4(qnum(r)+3)}\,\underline{\log_2(r+1)})$$

Discounting based on number of query reformulations

Discounting based on rank in concatenated list



The original session DCG [Jarvelin+ECIR08] has a problem: documents in earlier lists may be discounted more than those in later lists. [Kanoulas+SIGIR11] also describes an evaluation method for sessions based on multiple possible browsing paths over multiple ranked lists.

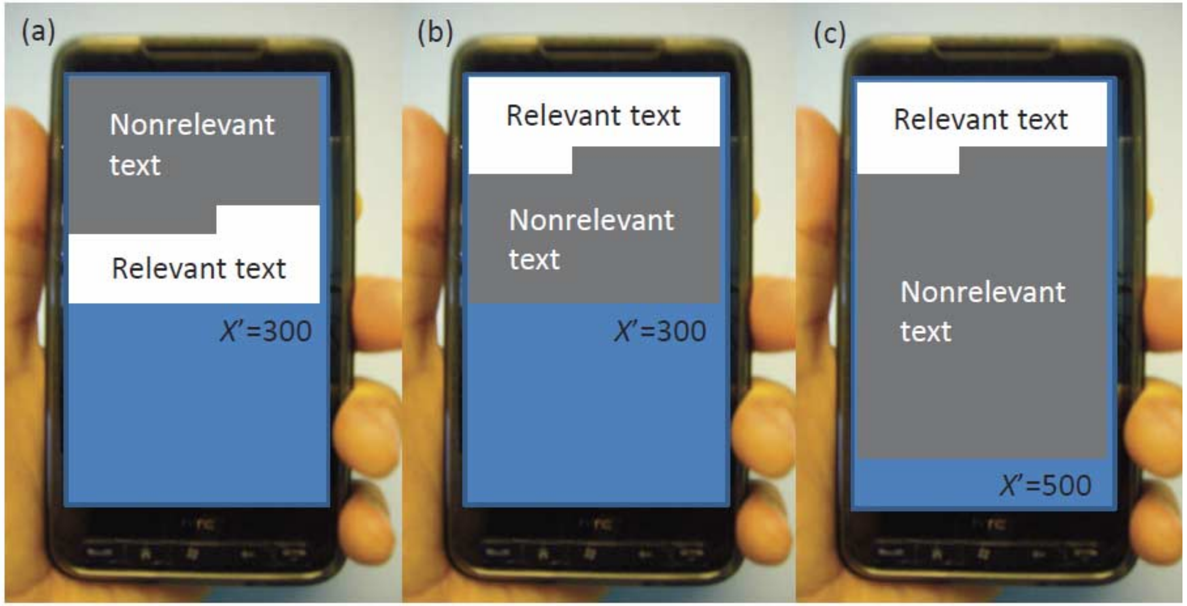
ROUGE, POURPRE

- Traditional IR evaluates a (ranked) list of documents, but text summarisation and question answering evaluate textual outputs.
- Instead of documents, nuggets and N-grams are used as the basic unit of evaluation.
- ROUGE [Lin ACL04ws] for summarisation is a recall/F-measure of automatically extracted word N-grams etc., based on gold standard summaries.
- POURPRE [Lin+IRJ06] for QA is an F-measure of answer nuggets, where nugget matching is done automatically using word N-grams.

S-measure, T-measure

[Sakai+CIKM11; Sakai+AIRS12b]

- Evaluating direct textual responses, not ranked lists of web pages
- Evaluate based on information units, not relevant documents
- Present important information first; minimise the user's reading effort



Unlike nugget precision/recall, S-measure (position-aware weighted recall) says (a)<(b). T-measure (a kind of precision) says (b)>(c). S# combines S and T.

LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary

Measuring agreement

- Cohen's kappa

For two raters who classify N items into C nominal categories

Observed

	Rater B			
Rater A		Yes	No	
	Yes	50	30	80
	No	10	10	20
		60	40	100

#Concordant=60

Chance expected

	Rater B			
Rater A		Yes	No	
	Yes	48	32	80
	No	12	8	20
		60	40	100

#Concordant=56

Cohen's kappa

$$= \frac{\text{Excess of observed concordant}}{\text{Chance expected nonconcordant}}$$

$$= (60-56)/(100-56)=0.09$$

range: [-1, 1]
1: complete agreement
0: completely due to chance

- Cohen's weighted kappa

For two raters who assign items into C ordinal categories e.g. relevance levels 1, 2 and 3 (|C|=3).

Considers relative concordances as well as absolute ones

- Fleiss' kappa

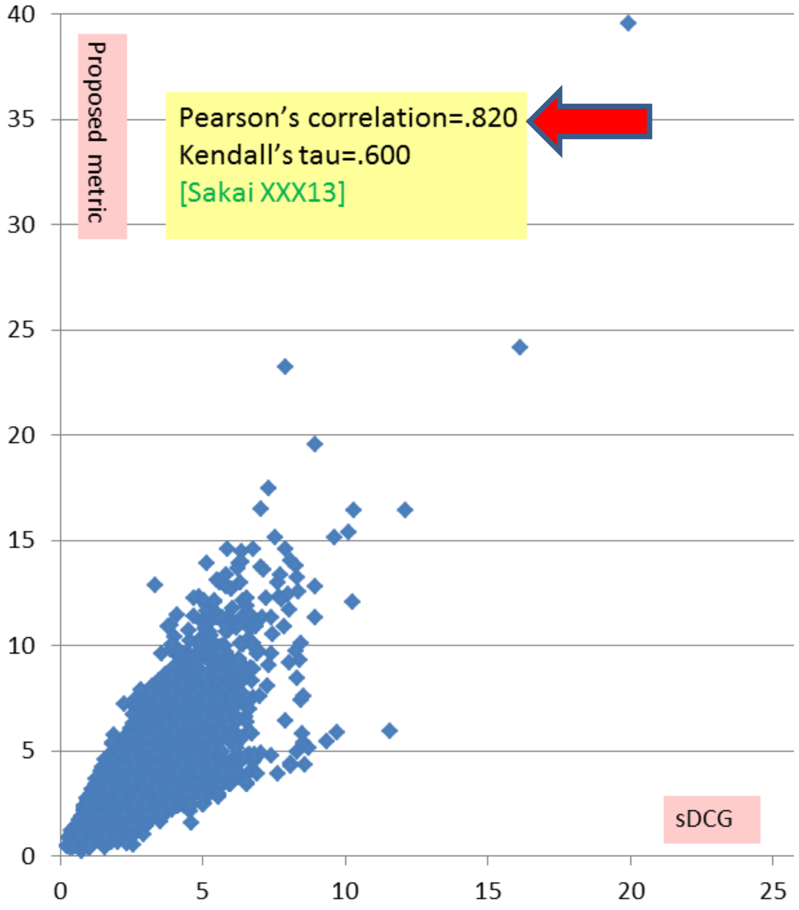
For three or more raters who classify items into C nominal categories

Pearson's correlation (Pearson product moment correlation)

- Degree of linear relationship between two variables (X,Y). Range: [-1, 1]
- $$\frac{\text{covariance}(X, Y)}{\text{stddev}(X) * \text{stddev}(Y)}$$
- For a sample, compute

$$\frac{N\,\Sigma XY - \Sigma X \Sigma Y}{\sqrt{(N\Sigma X^2 - (\Sigma X)^2)(N\Sigma Y^2 - (\Sigma Y)^2)}}$$

Shows that the values of the proposed metric correlate highly with sDCG



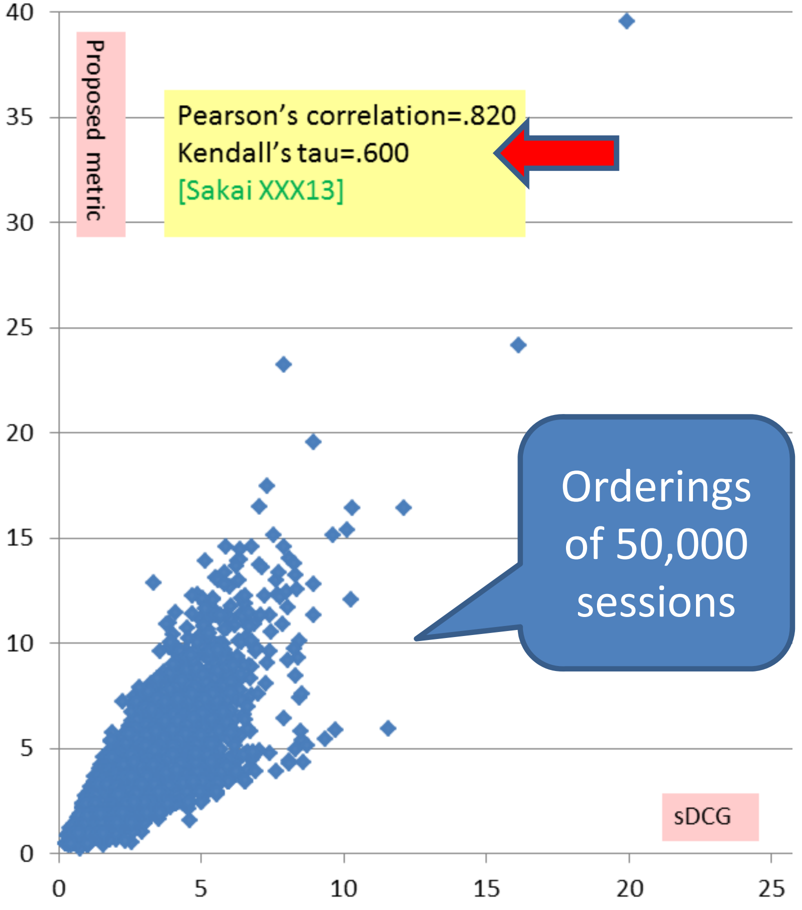
Pearson's correlation=.820
Kendall's tau=.600
[Sakai XXX13]

Kendall's τ rank correlation

- Similarity of the orderings of the data by X and Y (not absolute values)
- τ= (conc – disc)/all

all: all pairs of observations=N(N-1)/2
(xi,yi) and (xj,yj)
conc: concordant pairs
(xi>xj and yi>yj or xi<xj and yi<yj)
disc: discordant pairs
(xi>xj and yi<yj or xi<xj and yi>yj)

Range: [-1, 1]



Pearson's correlation=.820
Kendall's tau=.600
[Sakai XXX13]

Orderings of 50,000 sessions

Alternatives to Kendall's τ:
[Yilmaz+SIGIR08; Carterette SIGIR09; Webber+TOIS10]

LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing

- Standard significance tests
- Computer-based significance tests

5. Testing IR metrics

6. Lecture summary

Why do significance tests?

- Useful for discussing whether the difference in effectiveness between Systems A and B is substantial or due to chance.
- Null hypothesis H_0: all systems are equivalent
- p-value: Pr(observed or more extreme data|H_0)
- Difference is statistically significant if p-value is less than the significance level α (α is just a threshold so report p-values)
- Statistical significance does not imply practical significance
- Statistical insignificance does not imply practical insignificance

	Accept H0	**Reject H0**
H0 true (equivalent)	correct	Type I error (α)
H0 false (different)	Type II error (β)	correct

(Student's) t-test

ANOVA (Analysis of Variance) can be used for more than two

- Paired test: one topic set, two systems X and Y (typical setting in IR experiments)
- Observed diffs $\mathbf{z}=(z_1,\dots,z_N)=(x_1-y_1,\dots,x_N-y_N)$
- Assumption: errors are normally distributed

(Even if not, central limit theorem says the distribution approaches normal as N grows large)

- H_0: $\mu=0$ (population mean of differences is zero)
- H_1(alternative hypothesis): $\mu\neq0$ (two-tailed)
- Under H_0, $t(\mathbf{z})= \bar{z}/(\bar{\sigma}/\sqrt{N})$ where $\bar{\sigma}=\sqrt{\Sigma_i(z_i-\bar{z})^2/(N-1)}$ follows Student's t distribution with N-1 degrees of freedom

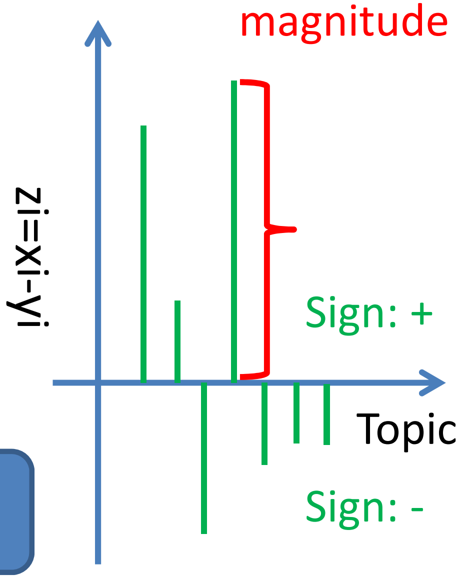
Paired nonparametric tests (fewer assumptions, less statistical power)

- Wilcoxon signed-rank test

Assumption: errors come from a continuous distribution symmetric about 0

- Rank zi's by magnitude;

Test statistic W= |Σ sign(zi)*rank(zi) |



- Sign test

Friedman test can be used for more than two systems

Assumption: errors come from a continuous distribution

- Only the sign of zi matters (ordinal scale)

Remove topics where Zi=0 (Reduce N)

Test statistic $|n^+ - n^-| / \sqrt{n^+ + n^-}$ follows standard normal distribution

n^+: number of topics where zi>0

n^- : number to topics where zi<0

On significance testing in the 20th-century IR literature

- [vanRijsbergen79] *"parametric tests are inappropriate because we do not know the form of the underlying distribution. […] One obvious failure is that the observations are not drawn from normally distributed populations."*

  "[…] the sign test […] can be used conservatively."
- [Hull SIGIR93] *"While the errors may not be normal, the t-test is relatively robust to many violations of normality. Only heavy skewness […] or large outliers […] will seriously compromise its validity."*

LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing

- Standard significance tests
- Computer-based significance tests

5. Testing IR metrics
6. Lecture summary

Why use computational power for significance testing?

- Standard significance tests were developed before the high-performance computer age. They rely on several assumptions (e.g. normality) on the underlying distributions, which often do not hold.
- Instead of making many assumptions, use the observed data and computational power to estimate the distributions!
- *“The use of the bootstrap either relieves the analyst from having to do complex mathematical derivations, or in some instances provides an answer where no analytical answer can be obtained.”* [Efron+93, p.394]

Bootstrap test for two systems
[Savoy IPM97; Sakai SIGIR06]

Two sample test also available

```
$\mathbf{z} = (z_1, \ldots, z_N)$ where $z_i =$ xi-yi ;
$t(\mathbf{z}) = \frac{\bar{z}}{\bar{\sigma}/\sqrt{N}}$ where $\bar{z}$ and $\bar{\sigma}$ are mean and standard deviation of $\mathbf{z}$;
$\mathbf{w} = (z_1 - \bar{z}, \ldots, z_N - \bar{z})$;
$count = 0$;
for $b = 1$ to $B$ do {
    $\mathbf{w}^{*b}$ = bootstrap sample of size $N$
        obtained by sampling with replacement from $\mathbf{w}$;
    $t(\mathbf{w}^{*b}) = \frac{\bar{w}^{*b}}{\bar{\sigma}^{*b}/\sqrt{N}}$ where $\bar{w}^{*b}$ and $\bar{\sigma}^{*b}$ are
        mean and standard deviation of $\mathbf{w}^{*b}$;
    if( $|t(\mathbf{w}^{*b})| \geq |t(\mathbf{z})|$ ) $count++$;
}
$ASL = count/B$;
```

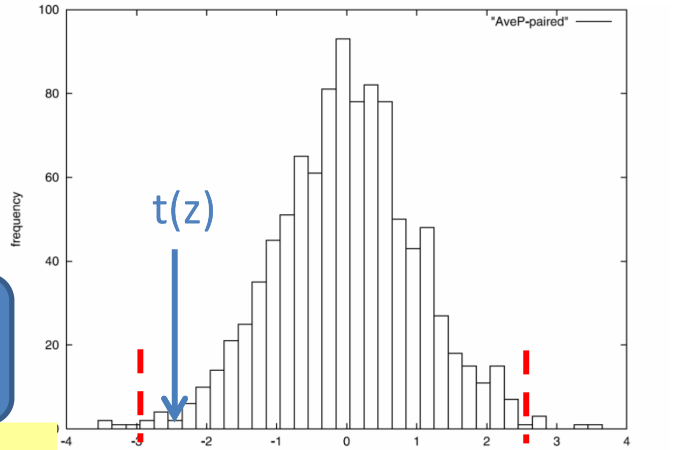
Difference for topic i

Studentised statistic of **z**

Shifted vector that obeys H0: population mean of the differences is zero

e.g. B=1000

i.e. p-value: how rare is this observation under H0?

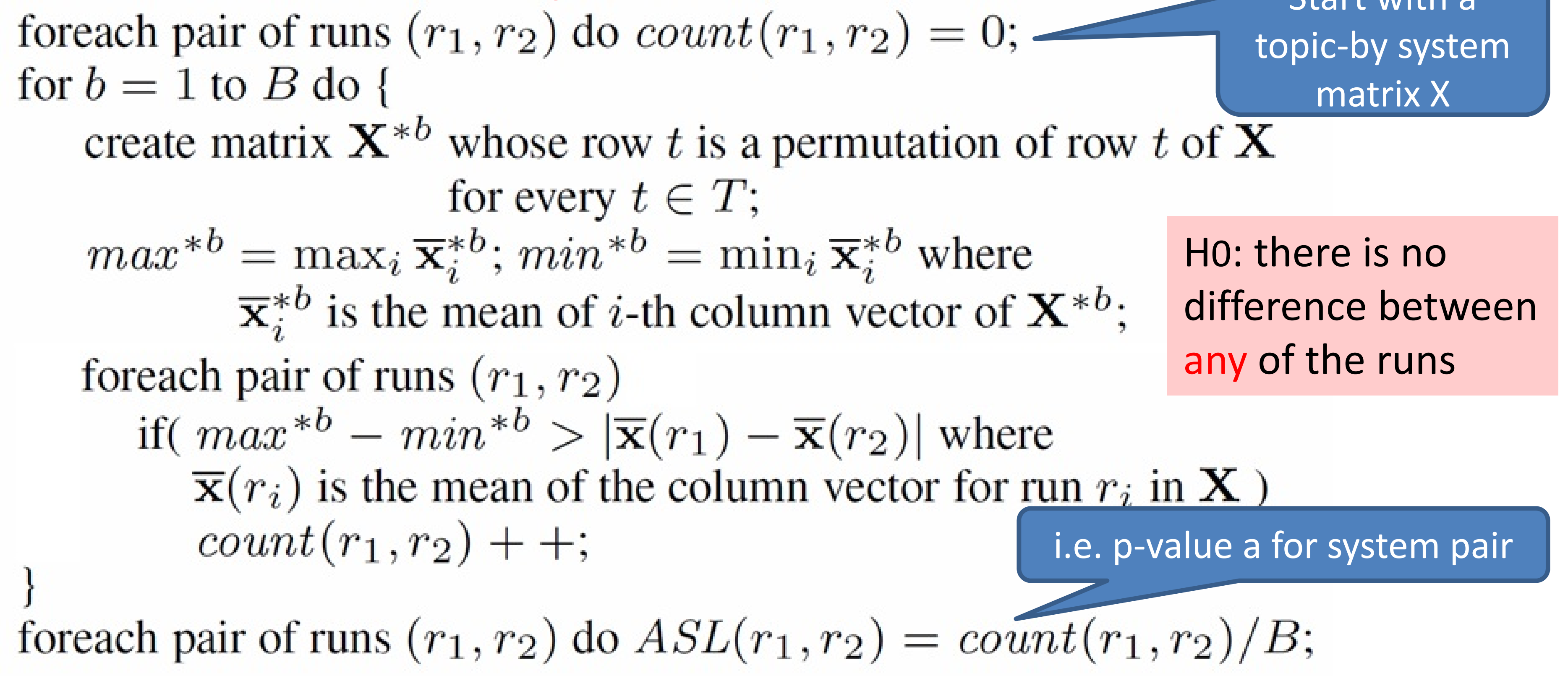


Histogram of $t(\mathbf{w}^{*b})$ for the difference in Mean Average Precision

See [Smucker+CIKM07] for randomisation test for two systems and comparison with classical and bootstrap tests

Randomised version of Tukey's Honestly Significantly Different (HSD) test for three or more systems [Carterette TOIS12]

If you have three or more systems but you are using pairwise tests, you may be jumping to wrong conclusions! Family-wise error rate=1-(1-α)$^{\text{nsystempairs}}$



```
foreach pair of runs $(r_1, r_2)$ do $count(r_1, r_2) = 0$;
for $b = 1$ to $B$ do {
    create matrix $\mathbf{X}^{*b}$ whose row $t$ is a permutation of row $t$ of $\mathbf{X}$
                    for every $t \in T$;
    $max^{*b} = \max_i \overline{\mathbf{x}}_i^{*b}$; $min^{*b} = \min_i \overline{\mathbf{x}}_i^{*b}$ where
            $\overline{\mathbf{x}}_i^{*b}$ is the mean of $i$-th column vector of $\mathbf{X}^{*b}$;
    foreach pair of runs $(r_1, r_2)$
        if( $max^{*b} - min^{*b} > |\overline{\mathbf{x}}(r_1) - \overline{\mathbf{x}}(r_2)|$ where
            $\overline{\mathbf{x}}(r_i)$ is the mean of the column vector for run $r_i$ in $\mathbf{X}$ )
            $count(r_1, r_2)++$;
}
foreach pair of runs $(r_1, r_2)$ do $ASL(r_1, r_2) = count(r_1, r_2)/B$;
```

Start with a topic-by system matrix X

H0: there is no difference between any of the runs

i.e. p-value a for system pair

Is significance testing useless? (from outside IR literature)

- [Johnson99] The insignificance of statistical significance testing

- [...] determining which outcomes of an experiment or survey are more

extreme than the observed one, so a P-value can be calculated, requires knowledge of the intentions of the investigator.

- If the null hypothesis truly is false (as most of those tested really are), then P can be made as small as one wishes, by getting a large enough sample.

- The famed quality guru W. Edwards Deming (1975) commented that the reason students have problems understanding hypothesis tests is that they may be trying to think.

- [Ioannidis05] Why most published research findings are false

- [...] most research questions are addressed by many teams, and it is misleading to emphasize the statistically significant findings of any single team. What matters is the totality of the evidence.

R: #true_relationships/#no_relationships among those tested in the field

- [...] instead of chasing statistical significance, we should improve our understanding of the range of R values —the pre-study odds— where research efforts operate

- Despite a large statistical literature for multiple testing corrections, usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding.

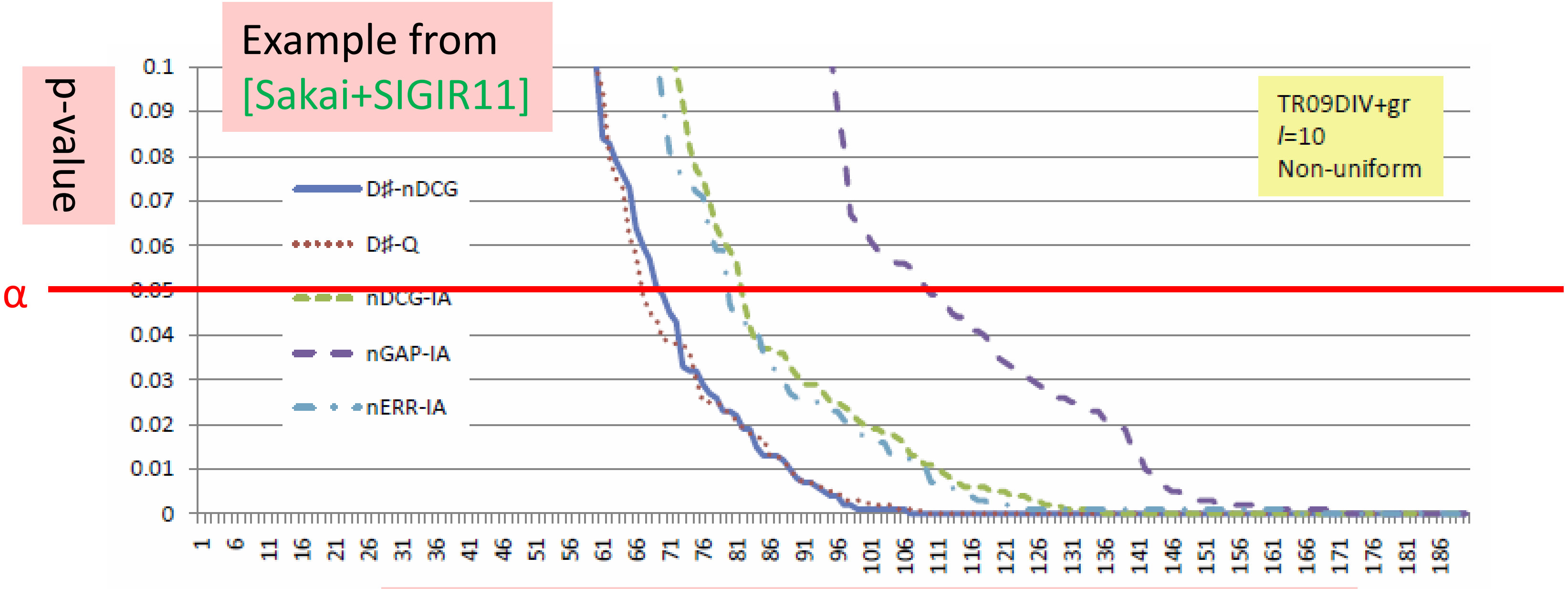
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary

Discriminative power

[Sakai SIGIR06; Sakai SIGIR07]

A method for comparing the robustness to topic variance: given a test collection, how many significantly different system pairs can be obtained?



Example from [Sakai+SIGIR11]

20 runs: 20*19/2= 190 run pairs sorted by p-value

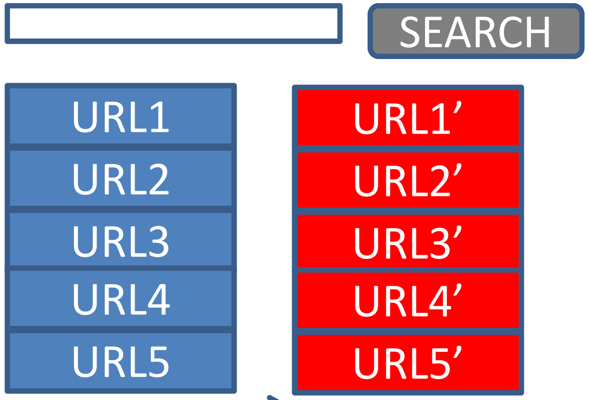
Discriminative power results are consistent with the swap method [Voorhees+SIGIR02] results but the latter needs to split the topic set in half. Discriminative power is now more widely used e.g. [Robertson+SIGIR10; Clarke+WSDM11; Smucker SIGIR12]

Comments on discriminative power [Sakai WWW12]

- Metrics with low discriminative power are not useful because they can’t give you conclusive results.
- It does not tell you whether the metric is measuring what you want to measure or not.
- Q: If a metric *knows* one list from Google and the other is from Bing, and says Bing is better no matter what the query is, isn’t discriminative power 100% and useless? [Sanderson FnTIR10]
- A: No, that’s cheating. A metric is a function of (a) the system output and (b) the gold standard. It doesn’t know which one is Google!

Side-by-side test

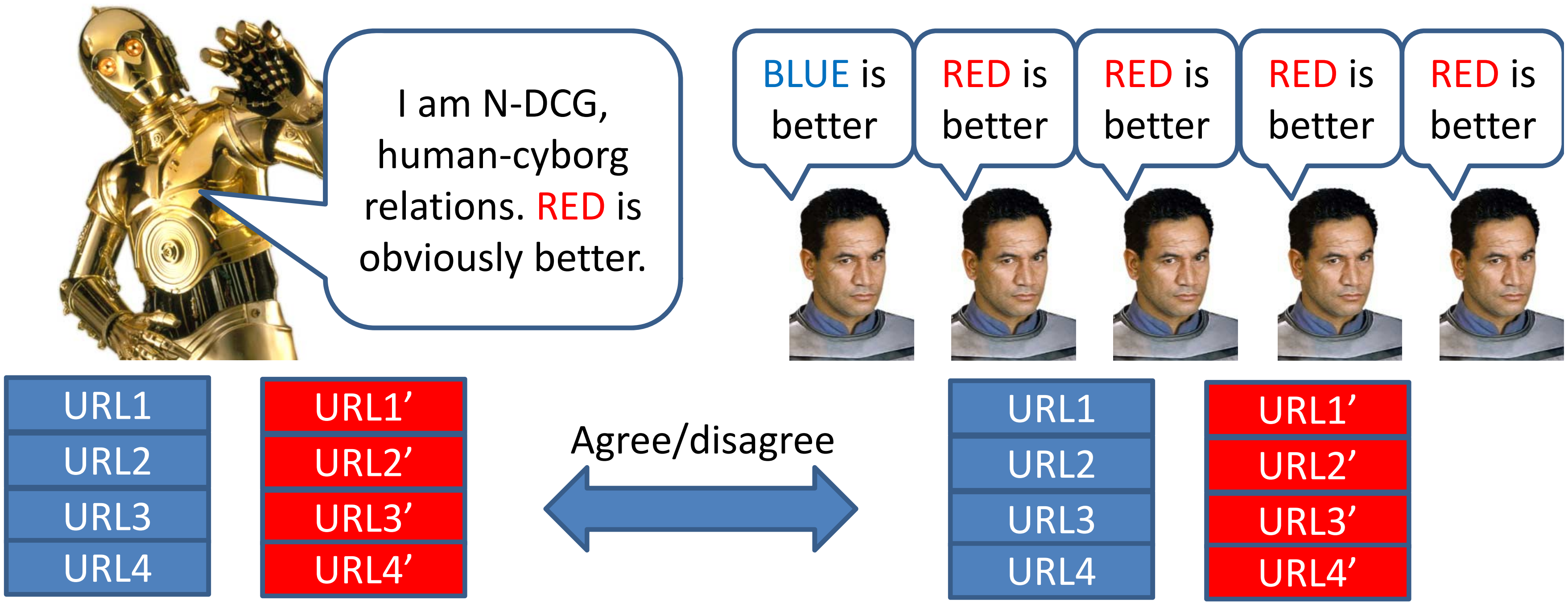
Microsoft’s campaign in 2012: blind comparison of Google’s and Bing’s ranked lists



Which is better? Left or right?

Predictive power [Sanderson+SIGIR10]

Is a metric "right?" Let's ask people!



I am N-DCG, human-cyborg relations. RED is obviously better.

BLUE is better

RED is better

RED is better

RED is better

RED is better

URL1
URL2
URL3
URL4

URL1'
URL2'
URL3'
URL4'

Agree/disagree

URL1
URL2
URL3
URL4

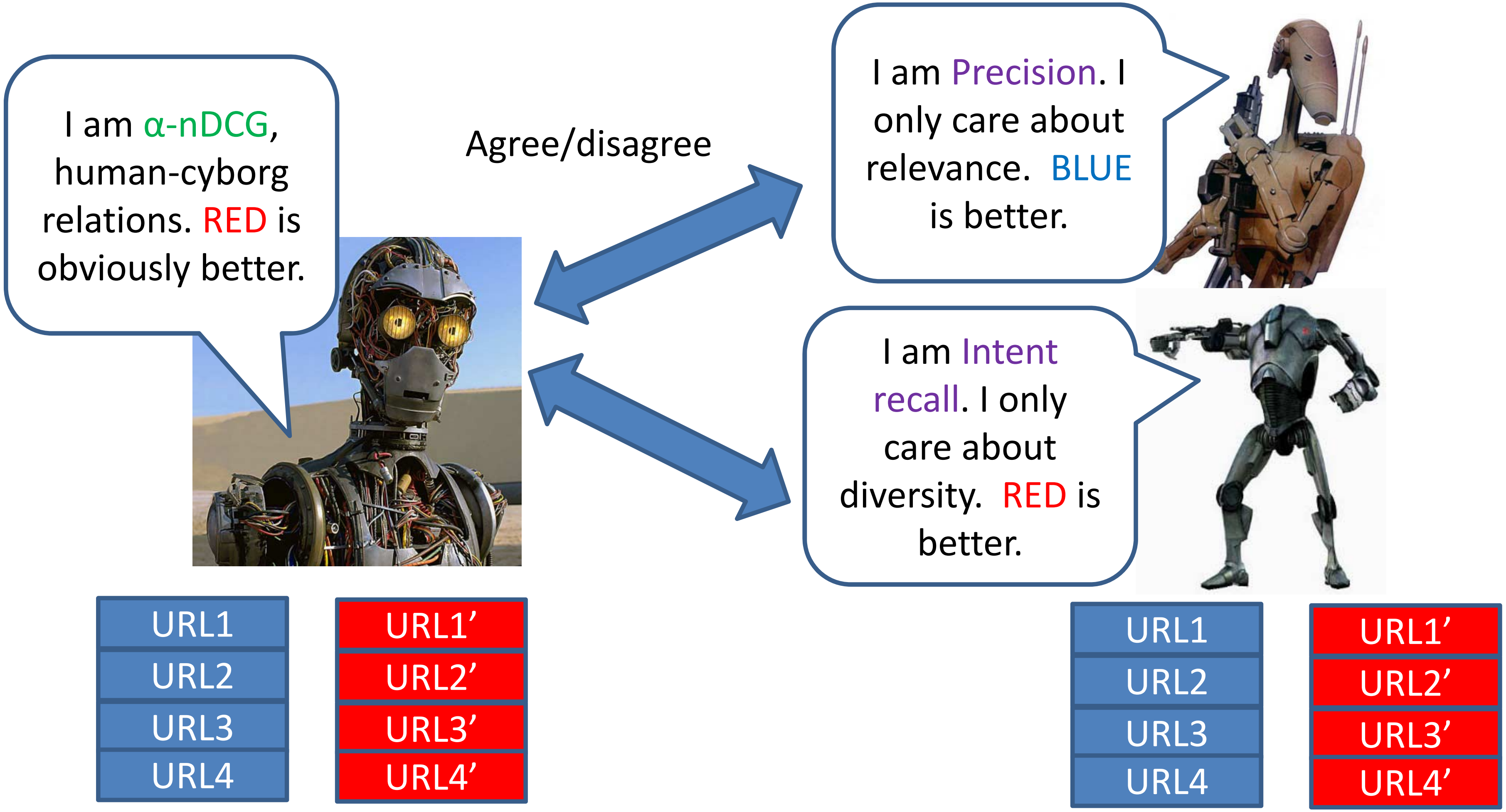
URL1'
URL2'
URL3'
URL4'

- Difficult to apply directly to diversified search metrics (each diversified list is intended for a population of users having different intents)
- Mechanical Turkers are not real users; need screening

Concordance test (a.k.a. intuitiveness test)

[Sakai WWW12; Sakai+IRJ13]

Is a diversity metric “right?” Let’s ask simpler metrics!



I am α-nDCG, human-cyborg relations. RED is obviously better.

Agree/disagree

I am Precision. I only care about relevance. BLUE is better.

I am Intent recall. I only care about diversity. RED is better.

URL1
URL2
URL3
URL4

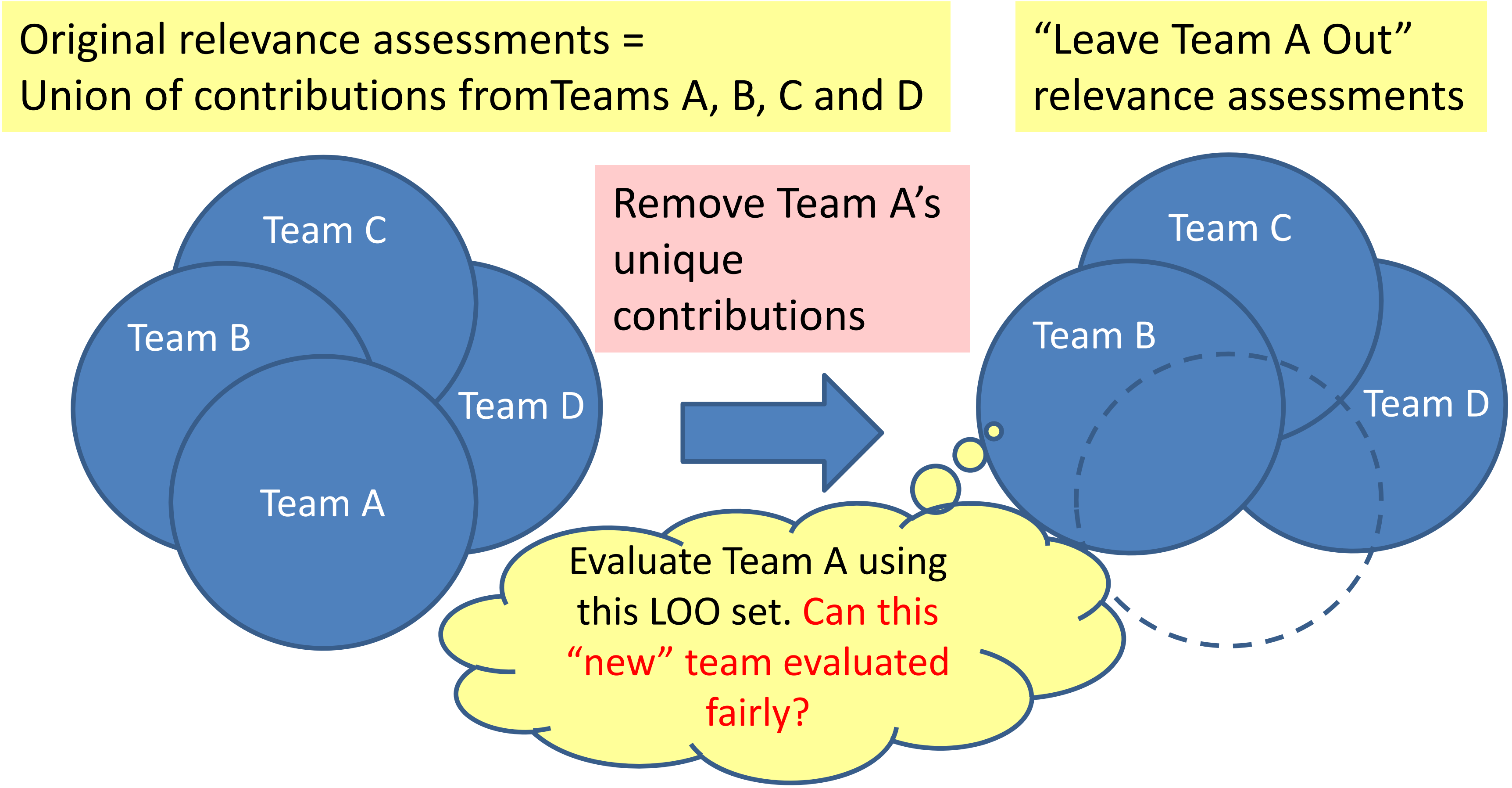
URL1’
URL2’
URL3’
URL4’

URL1
URL2
URL3
URL4

URL1’
URL2’
URL3’
URL4’

Leave-One-Out Test [Zobel SIGIR98]

Used for testing whether new systems can be evaluated fairly with a pooling-based test collection and an evaluation metric



Original relevance assessments = Union of contributions fromTeams A, B, C and D

"Leave Team A Out" relevance assessments

Remove Team A's unique contributions

Evaluate Team A using this LOO set. Can this "new" team evaluated fairly?

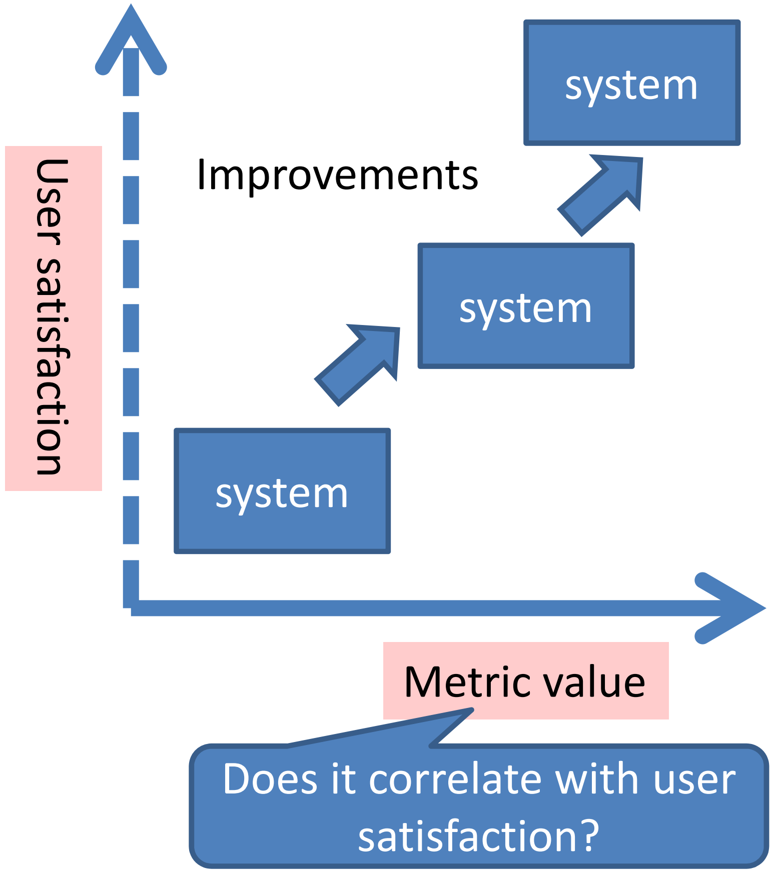
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary

Summary: using metrics correctly

- Understand and use the right metrics to evaluate your task.
- Several methods exist for discussing which metrics are “good.”
- Do significance testing with proper baselines.
- But statistical significance does not imply practical significance; statistical insignificance does not imply practical insignificance.
- Use multiple metrics/test collections and look for consistency.

“If you cannot measure it, you cannot improve it.”



Does it correlate with user satisfaction?

Further reading 1/2

- [Agrawal+WSDM09] Agrawal et al.: Diversifying search results, WSDM 2009.
- [Armstrong+CIKM09] Armstrong et al.: Improvements that don't add up: ad-hoc retrieval results since 1998, CIKM 2009.
- [Aslam+CIKM07] Aslam and Yilmaz: Inferring document relevance from incomplete information, CIKM 2007.
- [Buckley+SIGIR04] Buckley and Voorhees: Retrieval evaluation with incomplete information, SIGIR 2004.
- [Burges+ICML05] Burges et al.: Learning to rank using gradient descent, ICML 2005.
- [Buttcher+SIGIR07] Buttcher et al.: Reliable information retrieval evaluation with incomplete and biased judgments, SIGIR 2007.
- [Carterette SIGIR07] Carterette: Robust test collections for retrieval evaluation, SIGIR 2007.
- [Carterette SIGIR09] Carterette: On rank correlation and the distance between rankings, SIGIR 2009.
- [Carterette TOIS12] Carterette: Multiple testing in statistical analysis of systems-based information retrieval experiments, ACM TOIS, 2012.
- [Chapelle+CIKM09] Chapelle et al.: Expected reciprocal rank for graded relevance, CIKM 2009.
- [Chapelle+IRJ11] Chapelle et al.: Intent-based diversification of web search results: metrics and algorithms, Information Retrieval, 2011.
- [Chinchor MUC92] Chinchor: MUC-4 evaluation metrics, MUC-4, 1992.
- [Clarke+SIGIR08] Clarke et al.: Novelty and diversity in information retrieval evaluation, SIGIR 2008.
- [Clarke+WSDM11] Clarke et al.: A comparative analysis of cascade measures for novelty and diversity, WSDM 2011.
- [Efron+93] Efron and Tibshirani: An introduction to the bootstrap, Chapman & Hall/CRC, 1993.
- [Hull SIGIR93] Hull: Using statistical testing in the evaluation of retrieval experiments, SIGIR 1993.
- [Ioannidis05] Ioannidis: Why most published research findings are false, PLoS Med, 2005.
- [Jarvelin+TOIS02] Jarvelin and Kekalainen: Cumulated gain-based evaluation of IR techniques, ACM TOIS, 2002.
- [Jarvelin+ECIR08] Jarvelin et al.: Discounted Cumulated Gain based Evaluation of Multiple-Query IR Sessions, ECIR 2008.
- [Johnson99] Johnson: The insignificance of statistical significance testing, Journal of Wildlife Management, 1999.
- [Kanoulas+SIGIR11] Kanoulas et al.: Evaluating multi-query sessions, SIGIR 2011.
- [Korfhage97] Korfhage: Information Storage and Retrieval, Chapter 8, Wiley, 1997.
- [Moffat+TOIS08] Moffat and Zobel: Rank-Biased Precision for Measurement of Retrieval Effectiveness, ACM TOIS, 2008.
- [Lin ACL04ws] Lin: ROUGE: a package for automatic evaluation of summaries, ACL 2004 Workshop on Text Summarization Branches Out.
- [Lin+IRJ06] Lin and Demner-Fushman: Methods for automatically evaluating answers to complex questions, Information Retrieval, 2006.
- [Pollack AD68] Pollack: Measures for the comparison of information retrieval systems, American Documentation, 1968.
- [Robertson CIKM06] Robertson: On GMAP, CIKM 2006.
- [Robertson SIGIR08] Robertson: A new interpretation of average precision, SIGIR 2008 (poster).

Further reading 2/2

- [Robertson+SIGIR10] Robertson et al.: Extending average precision to graded relevance judgments, SIGIR 2010.
- [Sakai AIRS06] Sakai: Bootstrap-based comparisons of IR metrics for finding one relevant document, AIRS 2006.
- [Sakai SIGIR06] Sakai: Evaluating evaluation metrics based on the bootstrap, SIGIR 2006.
- [Sakai IPM07] Sakai: On the reliability of information retrieval metrics based on graded relevance, Information Processing and Management, 2007.
- [Sakai SIGIR07] Sakai: Alternatives to bpref, SIGIR 2007.
- [Sakai+EVIA08] Sakai and Robertson: Modelling A User Population for Designing Information Retrieval Metrics, EVIA 2008.g
- [Sakai CIKM08] Sakai: Comparing Metrics across TREC and NTCIR: The Robustness to System Bias, CIKM 2008.
- [Sakai+IRJ08] Sakai and Kando: On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments, Information Retrieval, 2008.
- [Sakai+SIGIR11] Sakai and Song: Evaluating diversified search results using per-intent graded relevance, SIGIR 2011.
- [Sakai+CIKM11] Sakai, Kato and Song: Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access, CIKM 2011.
- [Sakai+AIRS12a] Sakai et al.: The reusability of a diversified search test collection, AIRS 2012.
- [Sakai+AIRS12b] Sakai and Kato: One click one revisited: enhancing evaluation based on information units, AIRS 2012.
- [Sakai WWW12] Sakai: Evaluation with informational and navigational intents, WWW 2012.
- [Sakai+IRJ13] Sakai and Song: Diversified Search Evaluation: Lessons from the NTCIR-9 INTENT Task, Information Retrieval, 2013.
- [Sanderson FnTIR10] Sanderson: Test collection based evaluation of information retrieval systems, Foundations and Trends in Information Retrieval, 2010.
- [Sanderson+SIGIR10] Sanderson et al.: Do user preferences and evaluation measures line up? SIGIR 2010.
- [Savoy IPM97] Savoy: Statistical inference in retrieval effectiveness evaluation, Information Processing and Management, 1997.
- [Smucker+CIKM07] Smucker et al.: A comparison of statistical significance test for information retrieval evaluation, CIKM 2007.
- [Smucker SIGIR12] Smucker and Clarke: Time-based calibration of effectiveness measures, SIGIR 2012.
- [vanRijsbergen79] van Rijsbergen: Information Retrieval, Chapter 7, Butterworths, 1979.
- [Voorhees+SIGIR02] Voorhees and Buckley: The effect of topic set size on retrieval experiment error, SIGIR 2002.
- [Webber+SIGIR09] Webber and Park: Score adjustment for correction of pooling bias, SIGIR 2009.
- [Webber+TOIS10] Webber et al.: A similarity measure for indefinite rankings, ACM TOIS, 2010.
- [Yilmaz+CIKM06] Yilmaz and Aslam: Estimating average precision with incomplete and imperfect judgments, CIKM 2006.
- [Yilmaz+SIGIR08] Yilmaz et al.: A new rank correlation coefficient for information retrieval, SIGIR 2008.
- [Zhai+SIGIR03] Zhai et al.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, SIGIR 2003.
- [Zobel SIGIR98] Zobel: How reliable are the results of large-scale information retrieval experiments? SIGIR 1998.