

Evaluation

Information Retrieval

Indian Statistical Institute

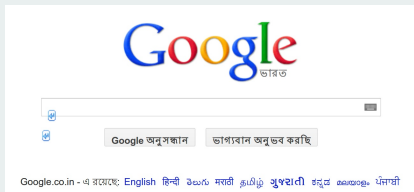
- 1 Preliminaries
- 2 Metrics
- 3 Evaluation forums
- 4 Significance tests (CMS Sec. 8.6)
- 5 Parameter tuning

- Which is better: Heap sort or Bubble sort?

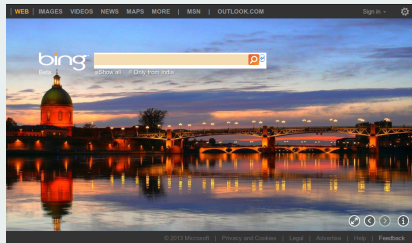
- Which is better: Heap sort or Bubble sort?

vs.

Which is better?



or



- IR is an *empirical* discipline.

- IR is an *empirical* discipline.
- Intuition can be wrong!
 - “sophisticated” techniques need not be the best
e.g. rule-based stemming vs. statistical stemming

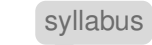
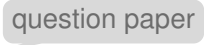

- IR is an *empirical* discipline.
- Intuition can be wrong!
 - “sophisticated” techniques need not be the best
e.g. rule-based stemming vs. statistical stemming
- Proposed techniques need to be validated and compared to existing techniques.

Benchmark data

- Document collection
- Query / topic collection
- Relevance judgments - information about which document is relevant to which query

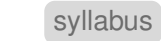
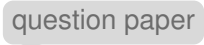

Cranfield method (CLEVERDON ET AL., 60S)

Benchmark data

- Document collection  syllabus
- Query / topic collection  question paper
- Relevance judgments - information about which document is relevant to which query  correct answers

Cranfield method (CLEVERDON ET AL., 60S)

Benchmark data

- Document collection syllabus
- Query / topic collection question paper
- Relevance judgments - information about which document is relevant to which query correct answers

Assumptions

- relevance of a document to a query is objectively discernible
- all relevant documents in the collection are known
- all relevant documents contribute equally to the performance measures
- relevance of a document is independent of the relevance of other documents

Outline

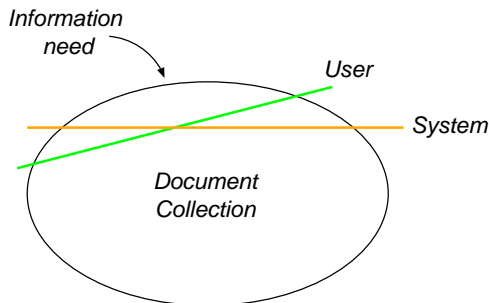
- 1 Preliminaries
- 2 Metrics**
- 3 Evaluation forums
- 4 Significance tests (CMS Sec. 8.6)
- 5 Parameter tuning

Background

- User has an information need.
- Information need is converted into a **query**.
- Documents are **relevant** or **non-relevant**.
- Ideal system retrieves all and only the relevant documents.

Background

- User has an information need.
- Information need is converted into a **query**.
- Documents are **relevant** or **non-relevant**.
- Ideal system retrieves all and only the relevant documents.



$$\begin{aligned}\text{Recall} &= \frac{\#(\text{relevant retrieved})}{\#(\text{relevant})} \\ &= \frac{\#(\text{true positives})}{\#(\text{true positives} + \text{false negatives})}\end{aligned}$$

$$\begin{aligned}\text{Precision} &= \frac{\#(\text{relevant retrieved})}{\#(\text{retrieved})} \\ &= \frac{\#(\text{true positives})}{\#(\text{true positives} + \text{false positives})}\end{aligned}$$

$$\begin{aligned}\mathbf{F} &= \frac{1}{\alpha/P + (1 - \alpha)/R} \\ &= \frac{(\beta^2 + 1)PR}{\beta^2 P + R}\end{aligned}$$

(Non-interpolated) average precision

Which is better?

1. Non-relevant

2. Non-relevant

3. Non-relevant

4. Relevant

5. Relevant

1. Relevant

2. Relevant

3. Non-relevant

4. Non-relevant

5. Non-relevant

Metrics for ranked results

(Non-interpolated) average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50

Metrics for ranked results

(Non-interpolated) average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

Metrics for ranked results

(Non-interpolated) average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

$$AvgP = \frac{1}{5} \left(1 + \frac{2}{3} + \frac{3}{6} \right)$$

(5 relevant docs. in all)

Metrics for ranked results

(Non-interpolated) average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

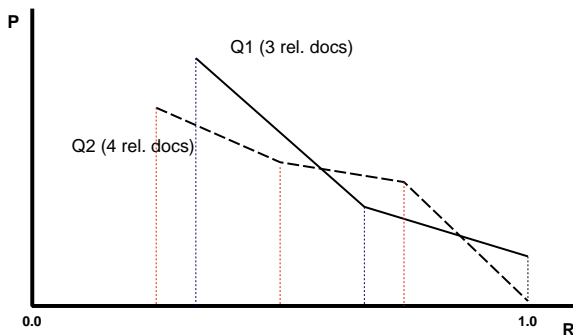
$$AvgP = \frac{1}{5} \left(1 + \frac{2}{3} + \frac{3}{6} \right)$$

(5 relevant docs. in all)

$$AvgP = \frac{1}{N_{Rel}} \sum_{d_i \in Rel} \frac{i}{Rank(d_i)}$$

Interpolated average precision at a given recall point

- Recall points correspond to $\frac{1}{N_{Rel}}$
- N_{Rel} different for different queries



- Interpolation required to compute averages across queries

Interpolated average precision

$$P_{int}(r) = \max_{r' \geq r} P(r')$$

Interpolated average precision

$$P_{int}(r) = \max_{r' \geq r} P(r')$$

11-pt interpolated average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

Metrics for ranked results

Interpolated average precision

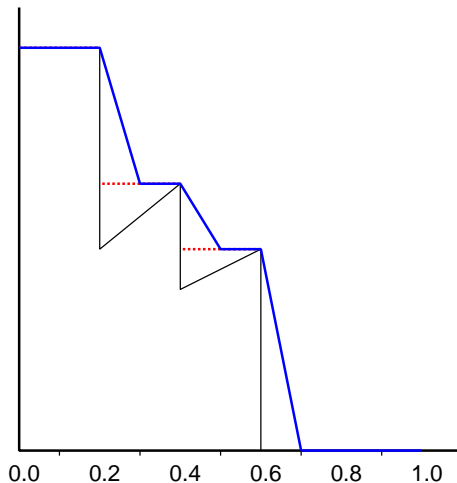
$$P_{int}(r) = \max_{r' \geq r} P(r')$$

11-pt interpolated average precision

Rank	Type	Recall	Precision
1	Relevant	0.2	1.00
2	Non-relevant		
3	Relevant	0.4	0.67
4	Non-relevant		
5	Non-relevant		
6	Relevant	0.6	0.50
∞	Relevant	0.8	0.00
∞	Relevant	1.0	0.00

<i>R</i>	Interp. <i>P</i>
0.0	1.00
0.1	1.00
0.2	1.00
0.3	0.67
0.4	0.67
0.5	0.50
0.6	0.50
0.7	0.00
0.8	0.00
0.9	0.00

11-pt interpolated average precision



Metrics for sub-document retrieval

Let p_r - document part retrieved at rank r

$rsize(p_r)$ - amount of relevant text contained by p_r

$size(p_r)$ - total number of characters contained by p_r

T_{rel} - total amount of relevant text for a given topic

$$P[r] = \frac{\sum_{i=1}^r rsize(p_i)}{\sum_{i=1}^r size(p_i)}$$

$$R[r] = \frac{1}{T_{rel}} \sum_{i=1}^r rsize(p_i)$$

- **Precision at k ($P@k$)** - precision after **k** documents have been retrieved
 - easy to interpret
 - not very stable / discriminatory
 - does not average well
- **R precision** - precision after N_{Rel} documents have been retrieved

Idea:

- Highly relevant documents are more valuable than marginally relevant documents
- Documents ranked low are less valuable

Idea:

- Highly relevant documents are more valuable than marginally relevant documents
- Documents ranked low are less valuable

$$Gain \in \{0, 1, 2, 3\}$$

$$G = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots \rangle$$

$$CG[i] = \sum_{j=1}^i G[j]$$

$$DCG[i] = \begin{array}{ll} CG[i] & \text{if } i < b \\ DCG[i - 1] + G[i] / \log_b i & \text{if } i \geq b \end{array}$$

$$DCG[i] = \begin{cases} CG[i] & \text{if } i < b \\ DCG[i-1] + G[i]/\log_b i & \text{if } i \geq b \end{cases}$$

$$\mathbf{Ideal} \ G = \langle 3, 3, \dots, 3, 2, \dots, 2, 1, \dots, 1, 0, \dots \rangle$$

$$nDCG[i] = \frac{DCG[i]}{\mathbf{Ideal} \ DCG[i]}$$

Mean Reciprocal Rank

- Useful for *known-item* searches with a single target
- Let r_i — rank at which the “answer” for query i is retrieved.
Then reciprocal rank = $1/r_i$

$$\text{Mean reciprocal rank (MRR)} = \sum_{i=1}^n \frac{1}{r_i}$$

- Relevance of a document to a query is objectively discernible.
- All relevant documents in the collection are known.
- All relevant documents contribute equally to the performance measures.
- Relevance of a document is independent of the relevance of other documents.

- Judges / assessors may not agree about relevance.

Example (MANNING ET AL.)

	Yes ₁	No ₁	Total ₂
Yes ₂	300	20	320
No ₂	10	70	80
Total ₁	310	90	400

$$P(A) = (300 + 70)/400 = 370/400 = 0.925$$

$$P(\text{nrel}) = (80 + 90)/(400 + 400) = 0.2125$$

$$P(\text{rel}) = (320 + 310)/(400 + 400) = 0.7878$$

$$P(E) = P(\text{non-rel})^2 + P(\text{rel})^2 = 0.665$$

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.925 - 0.665}{1 - 0.665} = 0.776$$

- Rules of thumb:

$\kappa > 0.8$ — good agreement

$0.67 \leq \kappa \leq 0.8$ — fair agreement

$\kappa < 0.67$ — poor agreement

- Exhaustive relevance judgments may be infeasible.
- Pool top results obtained by various systems and assess the pool.
- *Unjudged documents are assumed to be non-relevant.*

- Exhaustive relevance judgments may be infeasible.
- Pool top results obtained by various systems and assess the pool.
- *Unjudged documents are assumed to be non-relevant.*
- A wide variety of models, retrieval algorithms is important.
- **Manual interactive retrieval** is a must.

- Exhaustive relevance judgments may be infeasible.
- Pool top results obtained by various systems and assess the pool.
- *Unjudged documents are assumed to be non-relevant.*
- A wide variety of models, retrieval algorithms is important.
- **Manual interactive retrieval** is a must.

Can unbiased, incomplete relevance judgments be used to reliably compare the relative effectiveness of different retrieval strategies?

- Based on number of times judged nonrelevant documents are retrieved before relevant documents

Let R - set of relevant documents for a topic

N - set of first $|R|$ (or $10 + |R|$) judged non-rel docs retrieved

$$bpref = \frac{1}{|R|} \sum_{r \in R} \left(1 - \frac{|n \in N \text{ and } n \text{ ranked higher than } r|}{|R|} \right)$$

$$bpref10 = \frac{1}{|R|} \sum_{r \in R} \left(1 - \frac{|n \in N \text{ and } n \text{ ranked higher than } r|}{10 + |R|} \right)$$

- With complete judgments:
system rankings generated based on MAP and bpref10 are highly correlated
- When judgments are incomplete:
system rankings generated based on bpref10 are more stable

- 1 Preliminaries
- 2 Metrics
- 3 Evaluation forums**
- 4 Significance tests (CMS Sec. 8.6)
- 5 Parameter tuning

`http://trec.nist.gov`

- Organized by NIST every year since 1992
- Typical tasks
 - adhoc
 - user enters a search topic for a one-time information need
 - document collection is static
 - routing/filtering
 - user's information need is persistent
 - document collection is a stream of incoming documents
 - question answering

■ Documents

■ Genres:

- news (AP, LA Times, WSJ, SJMN, Financial Times, FBIS)
- govt. documents (Federal Register, Congressional Records)
- technical articles (Ziff Davis, DOE abstracts)

- Size: 0.8 million documents – 1.7 million web pages
(cf. Google indexes several billion pages)

■ Topics

- title
- description
- narrative

<http://www.clef-campaign.org/>

- CLIR track at TREC-6 (1997), CLEF started in 2000
- Objectives:
 - to provide an infrastructure for the testing and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts
 - to construct test-suites of reusable data that can be employed by system developers for benchmarking purposes
 - to create an R&D community in the cross-language information retrieval (CLIR) sector

- Monolingual retrieval
- Bilingual retrieval
 - queries in language X
 - document collection in language Y
- Multi-lingual retrieval
 - queries in language X
 - multilingual collection of documents (e.g. English, French, German, Italian)
 - results include documents from various collections and languages in a single list
- Other tasks: spoken document retrieval, image retrieval

<http://research.nii.ac.jp/ntcir>

- Started in late 1997
- Held every 1.5 years at NII, Japan
- Focus on East Asian languages (Chinese, Japanese, Korean)
- Tasks
 - cross-lingual retrieval
 - patent retrieval
 - geographic IR
 - opinion analysis

- Forum for Information Retrieval Evaluation
<http://www.isical.ac.in/~fire>
- Evaluation component of a DIT-sponsored, consortium mode project
- Assigned task: create a portal where
 1. a user will be able to give a query in one Indian language;
 2. s/he will be able to access documents available in the language of the query, Hindi (if the query language is not Hindi), and English,
 3. all presented to the user in the language of the query.
- Languages: Bangla, Hindi, Marathi, Punjabi, Tamil, Telugu

- To encourage research in South Asian language Information Access technologies by providing reusable large-scale test collections for ILIR experiments
- To provide a common evaluation infrastructure for comparing the performance of different IR systems
- To explore new Information Retrieval / Access tasks that arise as our information needs evolve, and new needs emerge
- To investigate evaluation methods for Information Access techniques and methods for constructing a reusable large-scale data set for ILIR experiments.
- To build language resources for IR and related language processing tasks

- Ad-hoc monolingual retrieval
 - Bengali, Hindi Marathi and English
- Ad-hoc cross-lingual document retrieval
 - documents in Bengali, Hindi, Marathi, and English
 - queries in Bengali, Hindi, Marathi, Tamil, Telugu , Gujarati and English
 - Roman transliterations of Bengali and Hindi topics
- MET: Morpheme Extraction Task (MET)
- RISOT: Retrieval from Indic Script OCR'd Text
- SMS-based FAQ Retrieval
- Older tracks:
 - Retrieval and classification from mailing lists and forums
 - Ad-hoc Wikipedia-entity retrieval from news documents

- 1 Preliminaries
- 2 Metrics
- 3 Evaluation forums
- 4 Significance tests (CMS Sec. 8.6)**
- 5 Parameter tuning

- Objective: compare A (your algorithm / system) with B (existing standard algorithm / system)
- Method
 1. Run A and B on one or more test collections $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$
 2. Compute standard evaluation metrics (M) from search results.
 3. Compare $M_1^{(A)}, M_1^{(B)}, \dots, M_k^{(A)}, M_k^{(B)}$.

Basic idea

Objective

Determine whether the observed difference between M_A, M_B is *meaningful* / *important* / *significant* (as opposed to *random* / *by chance*)

Objective

Determine whether the observed difference between M_A, M_B is *meaningful* / *important* / *significant* (as opposed to *random* / *by chance*)

Examples:

1. C_1 contains 10 queries; A beats B on 7 of them.
2. C_2 contains 100 queries; A beats B on 80 of them.
3. C_3 contains 1000 queries; A beats B on 950 of them.

Objective

Determine whether the observed difference between M_A, M_B is *meaningful* / *important* / *significant* (as opposed to *random* / *by chance*)

Examples:

1. \mathcal{C}_1 contains 10 queries; A beats B on 7 of them.
2. \mathcal{C}_2 contains 100 queries; A beats B on 80 of them.
3. \mathcal{C}_3 contains 1000 queries; A beats B on 950 of them.

How confident can we be that A is better than B ?

Objective

Determine whether the observed difference between M_A, M_B is *meaningful* / *important* / *significant* (as opposed to *random* / *by chance*)

Examples:

1. C_1 contains 10 queries; A beats B on 7 of them.
2. C_2 contains 100 queries; A beats B on 80 of them.
3. C_3 contains 1000 queries; A beats B on 950 of them.

How confident can we be that A is better than B ?

Significance tests allow us to quantify this confidence.

Null hypothesis (H_0) *There is no significant difference between A and B.*

Alternative hypothesis (H_1) A is significantly better than B.

Type I error Null hypothesis is rejected when it is in fact true.

Type II error Null hypothesis is accepted when it is in fact false.

Power Probability that the test will reject the null hypothesis correctly
(= $1 - P[\text{Type II error}]$)

Given: A particular test collection with n queries,
a particular evaluation metric, m , and
a significance level α .

Input: $\langle m_1^{(A)}, m_1^{(B)} \rangle, \langle m_2^{(A)}, m_2^{(B)} \rangle, \dots, \langle m_n^{(A)}, m_n^{(B)} \rangle$.

1. Compute a *test statistic*, $s = f(m^{(A)} - m^{(B)})$.
2. Compute the p -value, i.e., *probability that s would have the observed value (or a more extreme value) in a random sample if H_0 were true*.
3. If $p\text{-value} \leq \alpha$, H_0 is rejected in favor of H_1 . Otherwise, H_0 is not rejected.

Sign test (non-parametric)

$$\mathbf{H_0} : P(B > A) = P(A > B) = \frac{1}{2}$$

Test statistic = number of queries for which A is better than B .

Wilcoxon signed-rank test (non-parametric)

Assumption: Differences can be ranked, but their magnitudes are not important.

H_0 : sum of the positive ranks will be the same as the sum of the negative ranks

$$w = \sum_{i=1}^n R_i$$

where R_i - signed rank

Method

1. Order the differences $(m^{(A)} - m^{(B)})$ in increasing order of their absolute values.
2. Assign ranks to the differences.
3. Ranks are given the sign of the original difference.

Student's t test (parametric)

Assumption: Differences are normally distributed.

H_0 : Mean of the distribution of differences is zero.

$$t = \frac{\overline{m^{(A)}} - \overline{m^{(B)}}}{\sigma_{m^{(A)} - m^{(B)}}} \times \sqrt{N}$$

- For sign test, Wilcoxon test, are all differences considered non-zero?
- One-tailed (✓) vs. two-tailed tests (✗)

- 1 Preliminaries
- 2 Metrics
- 3 Evaluation forums
- 4 Significance tests (CMS Sec. 8.6)
- 5 Parameter tuning**

- *An Introduction to Information Retrieval*. Manning, Raghavan, Schutze, 2008 (Chapter 8).
<https://nlp.stanford.edu/IR-book/>
- *Test Collection Based Evaluation of Information Retrieval Systems*. Mark Sanderson. Foundations and Trends in Information Retrieval: Vol. 4: No. 4, pp. 247–375, 2010.
<http://dx.doi.org/10.1561/15000000009>
- *Information Retrieval Evaluation*. Donna Harman. Synthesis Lectures on Information Concepts, Retrieval, and Services, 2011.
<https://doi.org/10.2200/S00368ED1V01Y201105ICR019>