Evaluation

Information Retrieval

Indian Statistical Institute

Outline

Motivation

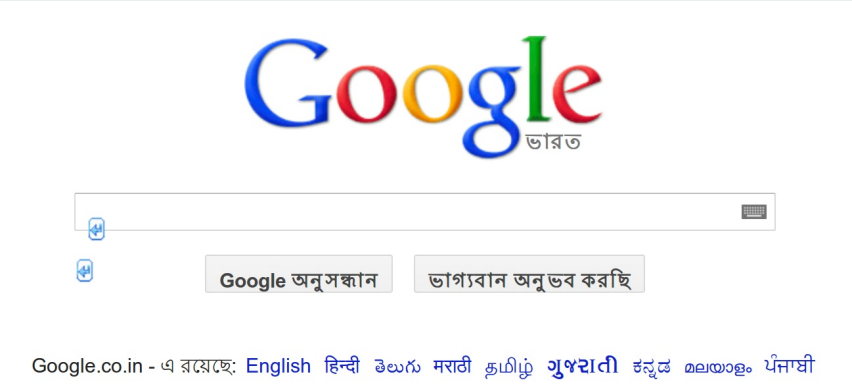
- Which is better: Heap sort or Bubble sort?

Motivation

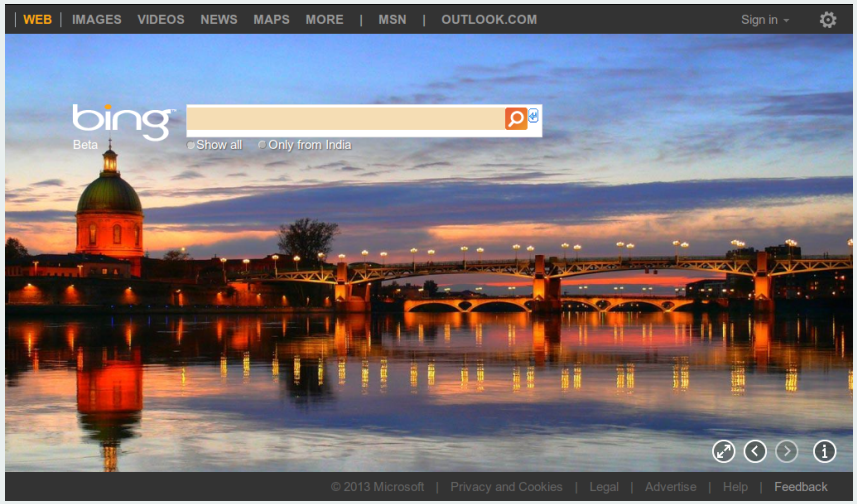
- Which is better: Heap sort or Bubble sort?

vs.

Which is better?



or



Motivation

- IR is an *empirical* discipline.

Motivation

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

Motivation

- 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.

Cranfield method (CLEVERDON ET AL., 60S)

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

syllabus

question paper

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

Evaluation metrics

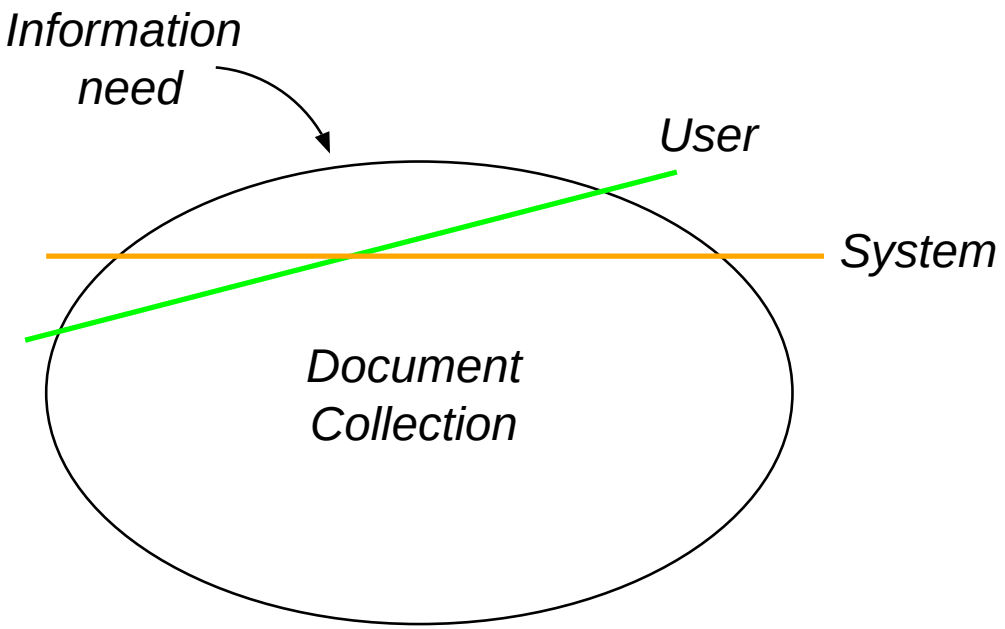
Background

- 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.

Evaluation metrics

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.



Set-based metrics

$$\textbf{Recall} = \frac{\#(\text{relevant retrieved})}{\#(\text{relevant})} = \frac{\#(\text{true positives})}{\#(\text{true positives + false negatives})}$$

$$\textbf{Precision} = \frac{\#(\text{relevant retrieved})}{\#(\text{retrieved})} = \frac{\#(\text{true positives})}{\#(\text{true positives + false positives})}$$

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

Metrics for ranked results

(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

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

(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

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

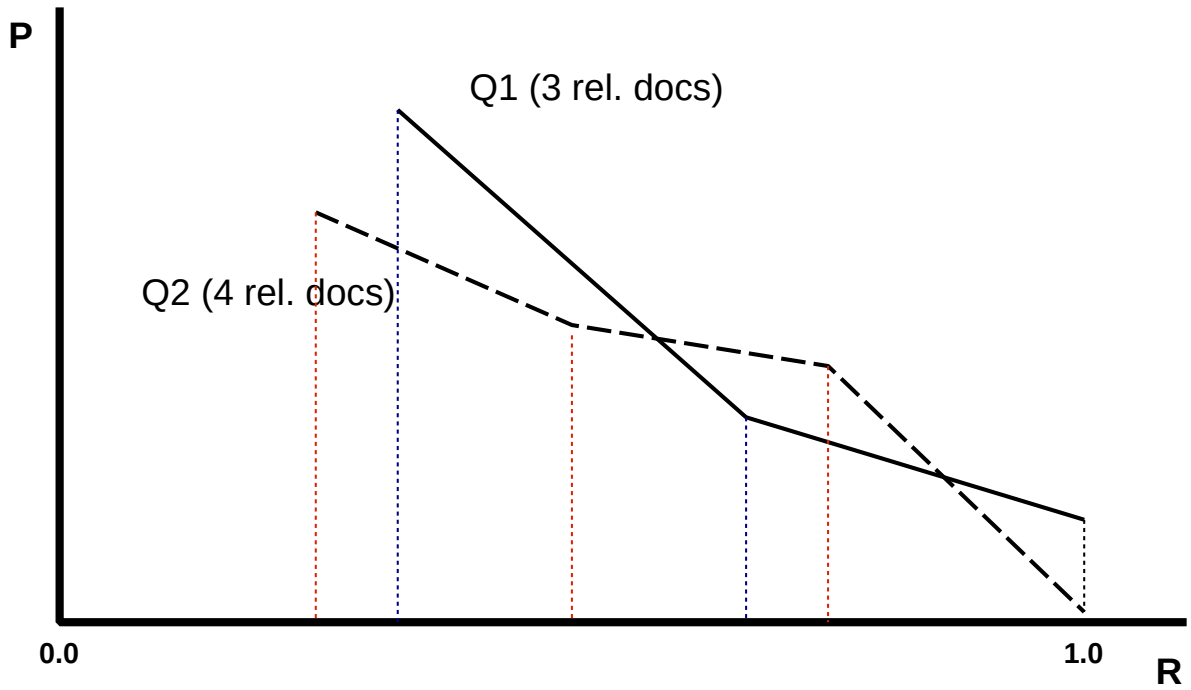
(5 relevant docs. in all)

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

Metrics for ranked results

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

Metrics for ranked results

Interpolated average precision

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

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

Metrics for ranked results

Interpolated average precision

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

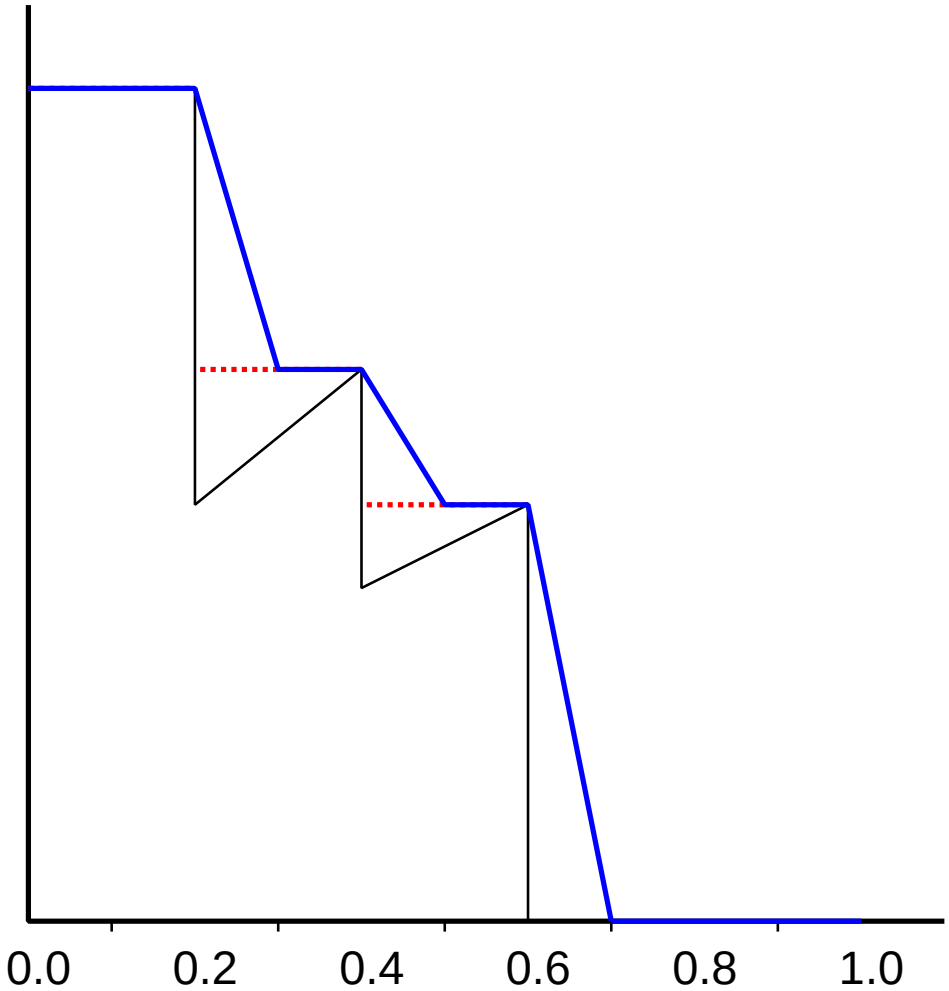
11-pt interpolated average precision

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

R	Interp. P
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

Metrics for ranked results

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

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

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

Metrics for ranked results

- **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

Cumulated Gain

Idea:

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

Cumulated Gain

Idea:

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

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

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

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

(n)DCG

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

(n)DCG

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

$$\textbf{Ideal } G = \langle 3, 3, \ldots, 3, 2, \ldots, 2, 1, \ldots, 1, 0, \ldots \rangle$$

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

Outline

TREC

`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

TREC data

- 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

CLEF

`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

CLEF tasks

- 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

NTCIR

`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

FIRE

- 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

FIRE: goals

- 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

FIRE: 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