Information Retrieval 1

Offline Evaluation

Ilya Markov
i.markov@uva.nl

University of Amsterdam

Offline evaluation

Evaluation

Document representation & matching

Learning to rank

IR—user interaction

How would you evaluate a search system?

Yandex

information retrieval Search

Web
Images
Video
Translate
More

Information retrieval - Wikipedia, the free encyclopedia
en.wikipedia.org › **Information retrieval**
Information retrieval (IR) is the activity of obtaining **information** resources relevant to an **information** need from a collection of **information** resources. Searches can be based on full-text or other content-based indexing.

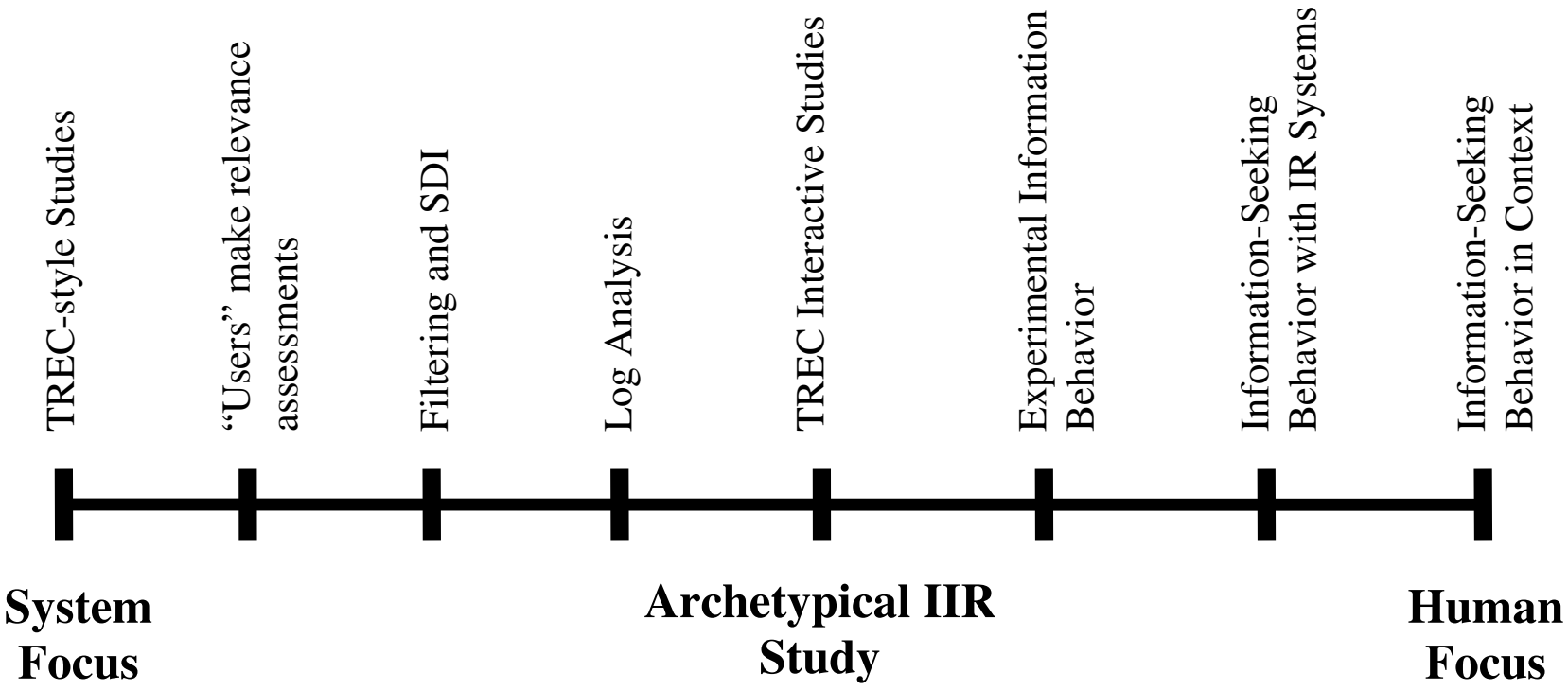
Information retrieval
en.academic.ru › dic.nsf/enwiki/9176
1950: The term "**information retrieval**" appears to have been coined by Calvin Mooers. 1951: Philip Bagley conducted the earliest experiment in computerized document **retrieval** in a...

Information retrieval - Psychology Wiki - Wikia
psychology.wikia.com › wiki/**Information_retrieval**
Assessment | Biopsychology | Comparative | Cognitive | Developmental | Language | Individual differences | Personality | Philosophy | Social | Methods | Statistics | Clinical | Educational | Industrial | Professional items | World psychology |.

Introduction to **Information Retrieval**
nlp.stanford.edu › IR-book/
A list of **information retrieval** resources is also available. Introduction to **Information Retrieval**: Table of Contents.

Information retrieval - Wikiquote
en.wikiquote.org › wiki/**Information_retrieval**
Information retrieval is the activity of obtaining **information** resources relevant to an **information** need from a collection of **information** resources, and the part of **information** science, which studies of these activity.

Taxonomy of evaluation approaches



Diane Kelly, "Methods for Evaluating Interactive Information Retrieval Systems with Users"

Outline

1. Test collections
2. Metrics

Outline

1. Test collections
   - Components of test collections
   - Evaluation campaigns
2. Metrics

Outline

1. Test collections
   - Components of test collections
   - Evaluation campaigns

What components should a test collection comprise?

- Test documents
- Test queries
- Ground truth

Test documents

Use a document collection
that is representative for the application
in terms of the number, size, and type.

Test queries

- Where can we get test queries?
  - Example queries from potential users
  - Query log
- How many queries should we get?
  - The more the better
  - At least 50

Ground truth

wars in netherlands in 17th century

Anglo-Dutch Wars - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Anglo-**Dutch_Wars**
The Anglo-**Dutch wars** (**Dutch**: Engels–Nederlandse Oorlogen or Engelse Zeeoorlogen) were ... A view of the **Dutch** factory at Ambon, early to mid-**17th century**.

First Anglo-Dutch War - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/First_Anglo-**Dutch_War**
1654, depicts the final battle of the First Anglo-**Dutch War**. ... By the middle of the **17th century** the **Dutch** had built by far the largest mercantile fleet in Europe, ...

1652-1674 Anglo-Dutch Wars - Rijksmuseum
https://www.rijksmuseum.nl/en/...**dutch**.../1652-1674-anglo-**dutch-wars**
In the **17th century**, England fought three **wars** with the Republic in a little over twenty years. Rivalry between the two mercantile nations and European power ...

The Anglo-Dutch wars - Het Geheugen van Nederland
www.geheugenvannederland.nl/?/en/collecties/nederland_engeland/...
(**Dutch**-English (Naval) **Wars**). Three of them were fought in the **seventeenth century**, one in the eighteenth. Trade conflicts and naval supremacy were at stake in ...

Anglo-Dutch Wars | European history | Britannica.com
www.britannica.com/topic/Anglo-**Dutch-Wars**
Jul 4, 2014 - Anglo-**Dutch Wars**, also called **Dutch Wars**, **Dutch** Engelse Oorlogen, (English **Wars**), the four **17th**- and 18th-**century** naval conflicts between ...

Relevance judgements

- Where can we get relevance judgements?
  - Users
  - Independent judges
  - Crowdsourcing
- How many relevance judgements should we get?
  - The more the better
  - More judged queries, fewer judgements per query
  - Multiple judges
- Graded relevance
  - 4 – perfect
  - 3 – excellent
  - 2 – good
  - 1 – fair
  - 0 – bad

Depth-k pooling

- Impossible to obtain judgments for all documents
- Depth-k pooling
  1. consider multiple search systems (by participants)
  2. consider top-k results from each system
  3. remove duplicates
  4. present documents to judges in a random order
- Produces a large number of judgments for each query
- Still incomplete

Multiple assessors

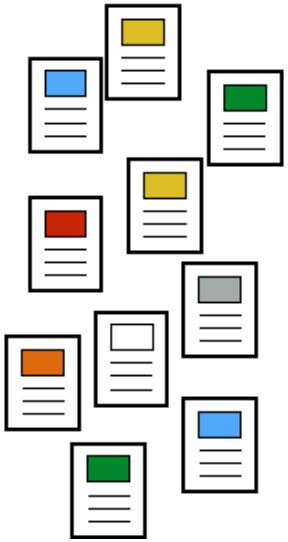
- Inter-assessor agreement, Cohen's kappa coefficient

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

- Expected chance agreement $P(E)$
- Values
  - > 0.8 – high
  - $0.67 - 0.8$ – acceptable
  - < 0.67 – low
- For more than two assessors, average pair-wise coefficients

Components of test collections

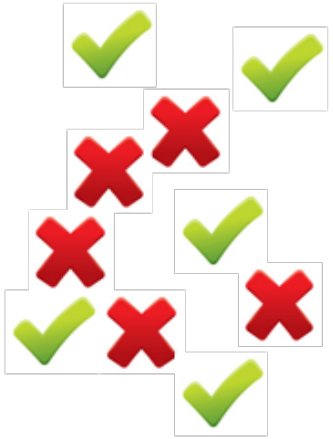
Documents



Queries

search engine evaluation
Amsterdam
web search
University of Amsterdam
information studies

Judgements



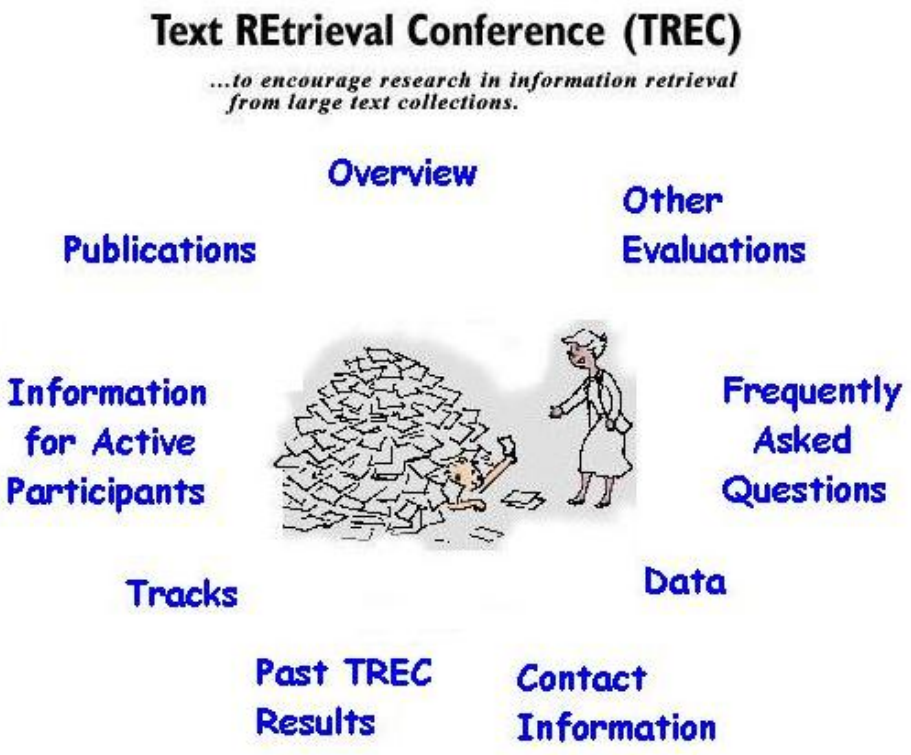
Outline

1. Test collections
   - Components of test collections
   - Evaluation campaigns

Evaluation campaigns

- Text REtrieval Conference (TREC)
  - US National Institute of Standard and Technology (NIST)
  - `http://trec.nist.gov`
- Cross-Language Education and Function (CLEF)
  - Mainly European
  - `http://www.clef-campaign.org`
- NII Test Collections for IR (NTCIR)
  - National Institute of Informatics of Japan (NII)
  - `http://research.nii.ac.jp/ntcir/index-en.html`

Text REtrieval Conference (TREC)



http://trec.nist.gov

TREC greatest hits

Track	Dataset	Year	Documents	Queries
Ad hoc track	TREC 1–8	1994–1999	1,89 million	450
Web track	WT10G	2000–2001	1,692,096	100
	ClueWeb09	2009–2012	1,040,809,705	200
	ClueWeb12	2013–2014	733,019,372	100
Terabyte track	GOV2	2004–2006	25,205,179	150

Components of test collections

- Test documents
- Test queries
- Ground truth

Test document

⟨DOC⟩
⟨DOCNO⟩ GX000-22-11749547 ⟨/DOCNO⟩
⟨TEXT⟩
Document text
⟨/TEXT⟩
⟨/DOC⟩

Test query

⟨TOP⟩
⟨NUM⟩ Number: 701
⟨TITLE⟩ U.S. oil industry history
⟨DESC⟩ Description: Describe the history of the U.S. oil industry
⟨NARR⟩ Narrative: Relevant documents will include those on historical exploration and drilling as well as history of regulatory bodies. Relevant are history of the oil industry in various states, even if drilling began in 1950 or later.
⟨/TOP⟩

Ground truth

```
701 0 GX000-22-11749547 0
701 0 GX000-25-2008761 1
701 0 GX000-27-14827260 0
701 0 GX000-41-2972136 0
701 0 GX000-43-8149041 2
701 0 GX000-45-2286833 0
701 0 GX000-55-12164304 0
701 0 GX000-55-3407826 2
701 0 GX000-67-12045787 2
701 0 GX000-72-8784276 2
```

Outline

1 Test collections

2 Metrics
- Unranked evaluation
- Ranked evaluation
- User-oriented evaluation

Outline

2. Metrics
   - Unranked evaluation
   - Ranked evaluation
   - User-oriented evaluation

Precision and recall

- **Precision** is the fraction of retrieved items that are relevant

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

- **Recall** is the fraction of relevant items that are retrieved

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

Manning et al., "Introduction to Information Retrieval"

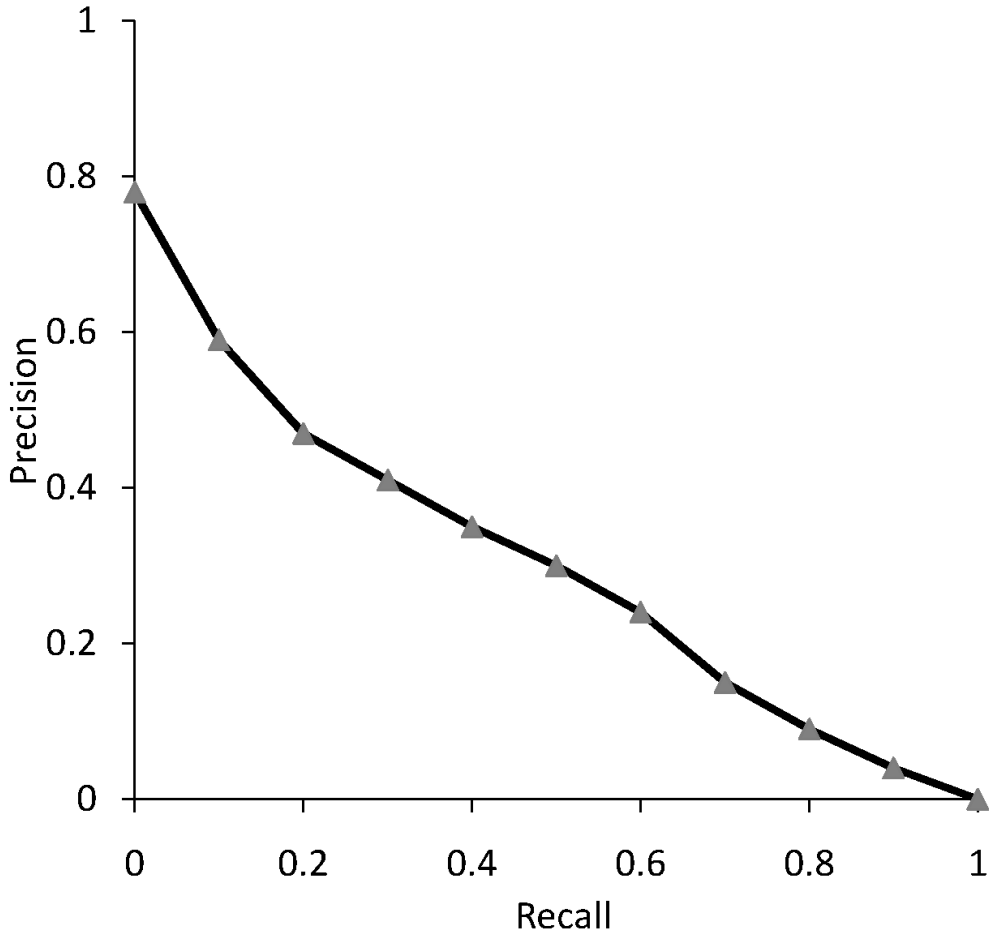
Precision and recall

	Relevant	Non-relevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$

Manning et al., "Introduction to Information Retrieval"

Precision-recall curve



Manning et al., "Introduction to Information Retrieval"

F-measure

- F-measure

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

where $\beta^2 = \frac{1-\alpha}{\alpha}$

- F1-measure $(\alpha = 0.5, \beta^2 = 1)$

$$F_1 = \frac{2PR}{P+R}$$

Manning et al., "Introduction to Information Retrieval"

Any problems with the metrics so far?

The ranking of items is not taken into account

Outline

2 Metrics
- Unranked evaluation
- Ranked evaluation
- User-oriented evaluation

Precision and recall

- Precision at rank k

$$P@k = \frac{\#(\text{relevant items at } k)}{k}$$

- Recall at rank k

$$R@k = \frac{\#(\text{relevant items at } k)}{\#(\text{relevant items})}$$

Other common metrics

- Reciprocal rank

$$RR = \frac{1}{\text{rank of first relevant item}}$$

- Average precision (AP)

$$AP = \frac{\sum_{d \in rel} P@k_d}{\#(\text{relevant items})}$$

- Average over multiple queries
  - mean $P@k$
  - mean $R@k$
  - MRR
  - MAP

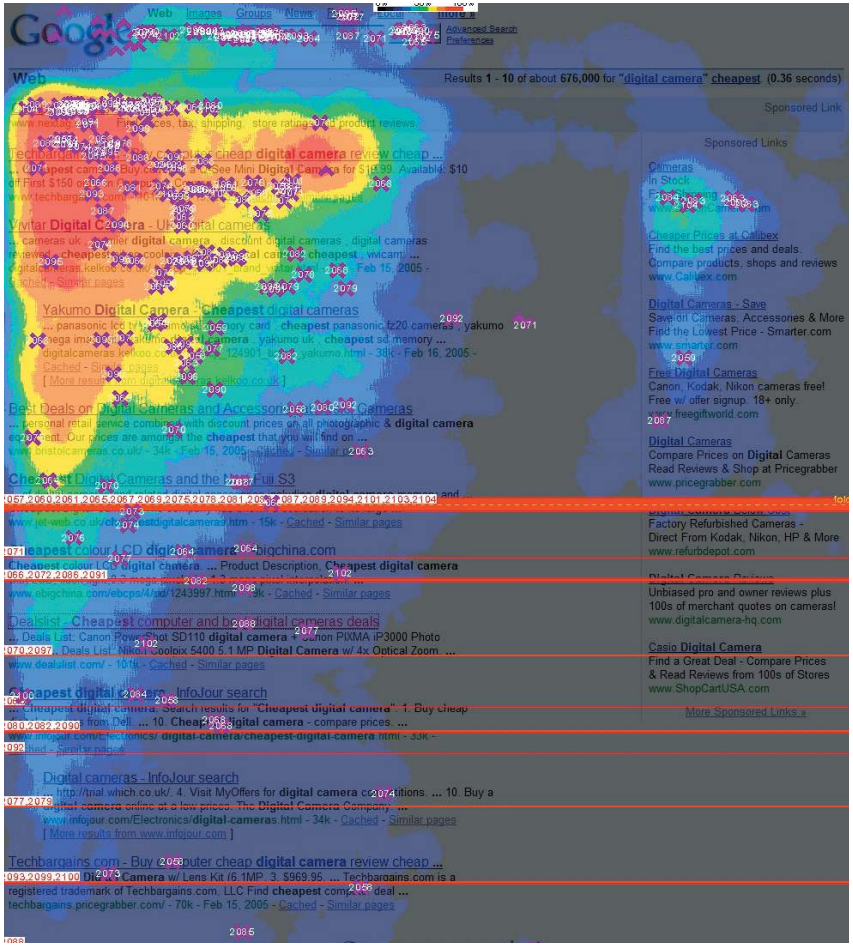
Any problems with the metrics so far?

User search behavior is not taken into account

Outline

2 Metrics
- Unranked evaluation
- Ranked evaluation
- User-oriented evaluation

User search behavior



Hotchkiss et al. "An In Depth Look at Interactions with Google using Eye Tracking Methodology"

Discounted cumulative gain (DCG)

- Graded relevance $R_k \in \{0, 1, 2, 3, 4\}$
- Cumulative gain

$$CG = \sum_{k=1}^{N} (2^{R_k} - 1)$$

- Gain is discounted by rank

$$D(k) = \frac{1}{\log(k+1)}$$

- Discounted cumulative gain

$$DCG = \sum_{k=1}^{N} \frac{2^{R_k} - 1}{\log(k+1)}$$

- Normalized DCG

$$NDCG = \frac{DCG}{DCG_{ideal}}$$

Rank-biased precision (RBP)

- View next item with probability θ
- Stop with probability $1 - \theta$
- Probability of looking at rank k

$$P(\text{look at } k) = \theta^{k-1}$$

- Average number of examined items

$$\begin{aligned}\text{Avg. exam} &= \sum_{k=1}^{\infty} k \cdot P(\text{look at } k) \cdot P(\text{stop at } k) \\ &= \sum_{k=1}^{\infty} k \cdot \theta^{k-1} \cdot (1-\theta) \\ &= \frac{1}{1-\theta}\end{aligned}$$

Rank-biased precision (RBP)

- Utility at rank k

$$U@k = P(\text{look at } k) \cdot R_k = \theta^{k-1} \cdot R_k$$

- Average utility of all results

$$RBP = \frac{\sum_{k=1}^{N} U@k}{\text{Avg. exam}} = (1-\theta) \cdot \sum_{k=1}^{N} \theta^{k-1} \cdot R_k$$

- θ is usually close to 1

Expected reciprocal rank (ERR)

- Reciprocal rank

$$RR = \frac{1}{\text{rank of first relevant item}}$$

- If an item is relevant (R_k) then stop
- Otherwise ($1 - R_k$), continue with probability θ
- Probability of looking at rank k

$$P(\text{look at } k) = \prod_{i=1}^{k-1} ((1 - R_i) \cdot \theta)$$

- Probability of reciprocal rank $= \frac{1}{k}$

$$P(RR = \frac{1}{k}) = R_k \cdot \prod_{i=1}^{k-1} ((1 - R_i) \cdot \theta)$$

Expected reciprocal rank (ERR)

- Expected reciprocal rank

$$ERR = \sum_{k=1}^{N} \frac{1}{k} \cdot P(RR = \frac{1}{k})$$
$$= \sum_{k=1}^{N} \frac{1}{k} \cdot \theta^{k-1} \cdot R_k \cdot \prod_{i=1}^{k-1} (1 - R_i)$$

- θ is usually close to 1

Offline evaluation summary

- Test collection
  - Test documents
  - Test queries
  - Ground truth
- Metrics
  - Unranked
  - Ranked
  - User-oriented

Materials

- Croft et al., Chapter 8
- Manning et al., Chapter 8
- Evangelos Kanoulas
  A Short Survey on Online and Offline Methods for Search Quality Evaluation
  Proceedings of RuSSIR, 2015

Materials

- DCG

  Kalervo Järvelin, Jaana Kekäläinen
  Cumulated gain-based evaluation of IR techniques
  ACM Transactions on Information Systems, 2002

- RBP

  Alistair Moffat, Justin Zobel
  Rank-biased precision for measurement of retrieval effectiveness
  ACM Transactions on Information Systems , 2008

- ERR

  Olivier Chapelle, Donald Metlzer, Ya Zhang, Pierre Grinspan
  Expected reciprocal rank for graded relevance
  Proceedings of CIKM, 2009

- Evaluation of metrics

  Aleksandr Chuklin, Pavel Serdyukov, Maarten de Rijke
  Click model-based information retrieval metrics
  Proceedings of SIGIR, 2013