CS646: Information Retrieval

Evaluation

Hamed Zamani
University of Massachusetts Amherst

Some slides are borrowed from previous iterations of by James Allan and Jiepu Jiang.

Why Evaluation is Important?

- Without a proper evaluation methodology …
  - we cannot compare two IR models.
  - we cannot improve an IR system.
  - we cannot advance the state of the art.
  - we cannot measure how useful an IR system is.

What to measure?

- The ability of the system to present **all relevant** documents
- The ability of the system to **withhold non-relevant** documents
- The interval between the demand being made and the answer being given (**response time**)
- The physical form of the output (**presentation**)
- The effort, intellectual or physical, demanded of the user (**user's effort**).

And many more …

Cyril W. Cleverdon and Michael Keen. "Factors Affecting the Performance of Indexing Systems," Vol 2. ASLIB, Cranfield Research Project. (1966)

Evaluation Methodologies

- Offline:
  - Mostly using test collections
    - **IR collections with relevance judgments**
    - The data collection from users’ explicit feedback
    - The data collection from users’ implicit feedback
- Online:
  - User interactions with the system (explicit or implicit)

History

1950s - 1960s

Cyril W. Cleverdon established the test collection evaluation methodology.

Cyril W. Cleverdon
(SIGIR Salton Award, 1991)

History

1950s - 1960s

1970s - early 1990s

Pre-TREC period:

Initial development of test collections, mostly catalogue information about academic papers, later news articles.

Evaluation metrics primarily focused on recall.

History

1950s - 1960s → 1970s - early 1990s → 1990s - early 2000s

TREC ad hoc period:

Focusing on ad-hoc retrieval of news articles.

Measures primarily focused on recall.

Introducing some tasks beyond the standard ad-hoc retrieval (e.g., filtering tracks)

History

1950s - 1960s → 1970s - early 1990s → 1990s - early 2000s → 2000s - current

Post TREC ad hoc period:

Beyond news articles (e.g., web pages, blog posts, social media posts)

Beyond single query (e.g., session search, conversational search)

Diverse set of measures

Fair ranking

TREC: Text Retrieval Conference

- The major IR evaluation campaign established in 1992
- DARPA Funded NIST in 1990 to build a test collection.
- NIST proposed to distribute the dataset through TREC in 1991 (leader: Donna Harman)
- November 1992: The first TREC meeting
- URL: http://trec.nist.gov/
- Every participated team can submit a report describing their approach. Reports are published in the TREC proceedings without peer-review.

TREC General Format

- November: Tracks approved by TREC (each year's program consists of multiple tracks)
- Spring: The track materials (e.g., collection and query set) are released.
- August: Submission due for participants
- Fall: Evaluation done by NIST
- November: TREC annual meeting (conference)

Other Evaluation Campaigns

- **CLEF**: Conference and Labs of the Evaluation Forum (aka, the European version of TREC)
  - Before 2010: Cross Language Evaluation Forum
- **FIRE**: Forum for Information Retrieval Evaluation
- **NTCIR**: NII Testbeds and Community for Information access Research
- **INEX**: Initiative for the Evaluation of XML Retrieval

- Useful shared-tasks occasionally defined by WSDM Cup, CIKM Cup, RecSys Challenge, and KDD Cup.

TREC Ad hoc Retrieval Track

- Simulate an information analyst (high recall)
- Multi-field topic description (title, description, and narrative)
- News articles + Government documents
- Relevance criteria: “a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document)”
- Each submitted run returns 1000 documents for evaluation with various measures

Cranfield's Evaluation Methodology

- Specify a retrieval task
- Create a collection of documents
- Create a set of topics/queries appropriate for the retrieval task
- Create a set of relevance judgments (i.e., judgments about which document is relevant to which query)
- Define a set of measures
- Apply a method to the collection

Statistics of Some IR Collections

	Cranfield 2	TREC2	GOV2	ClueWeb 09
year	1962	1991	2004	2009
type	Scientific articles	News articles	.gov crawl	Common web crawl
# documents	1,400	742,611	25,205,179	1,040,809,705
size	1.6 MB	2,162 MB	426 GB	25 TB
# queries	225	100	150	200

Relevance Judgments

Many types of relevance judgments:

- System or algorithmic relevance
- **Topical or subject relevance**
  - Aboutness
- Cognitive relevance or pertinence
  - Informativeness, novelty, information quality, …
- Situational relevance or utility
  - Usefulness in decision making, reduction of uncertainty, …
- Motivational or affective relevance
  - Satisfaction, success, accomplishment, …

Tefko Saracevic
(SIGIR Salton Award, 1997)

Tefko Saracevic. *"Relevance reconsidered."* In Proceedings of the Second Conference on Conceptions of Library and Information Science, 1996.

Relevance Judgments: Cost vs. Completeness

- To have accurate judgments, complete relevance judgments (for all query-document pairs) are required.
- Complete judgment for large-scale collections is almost impossible.
- Solutions:
  - Search-based
  - **Pooling**
  - Sampling

Relevance Judgments: Cost vs. Completeness

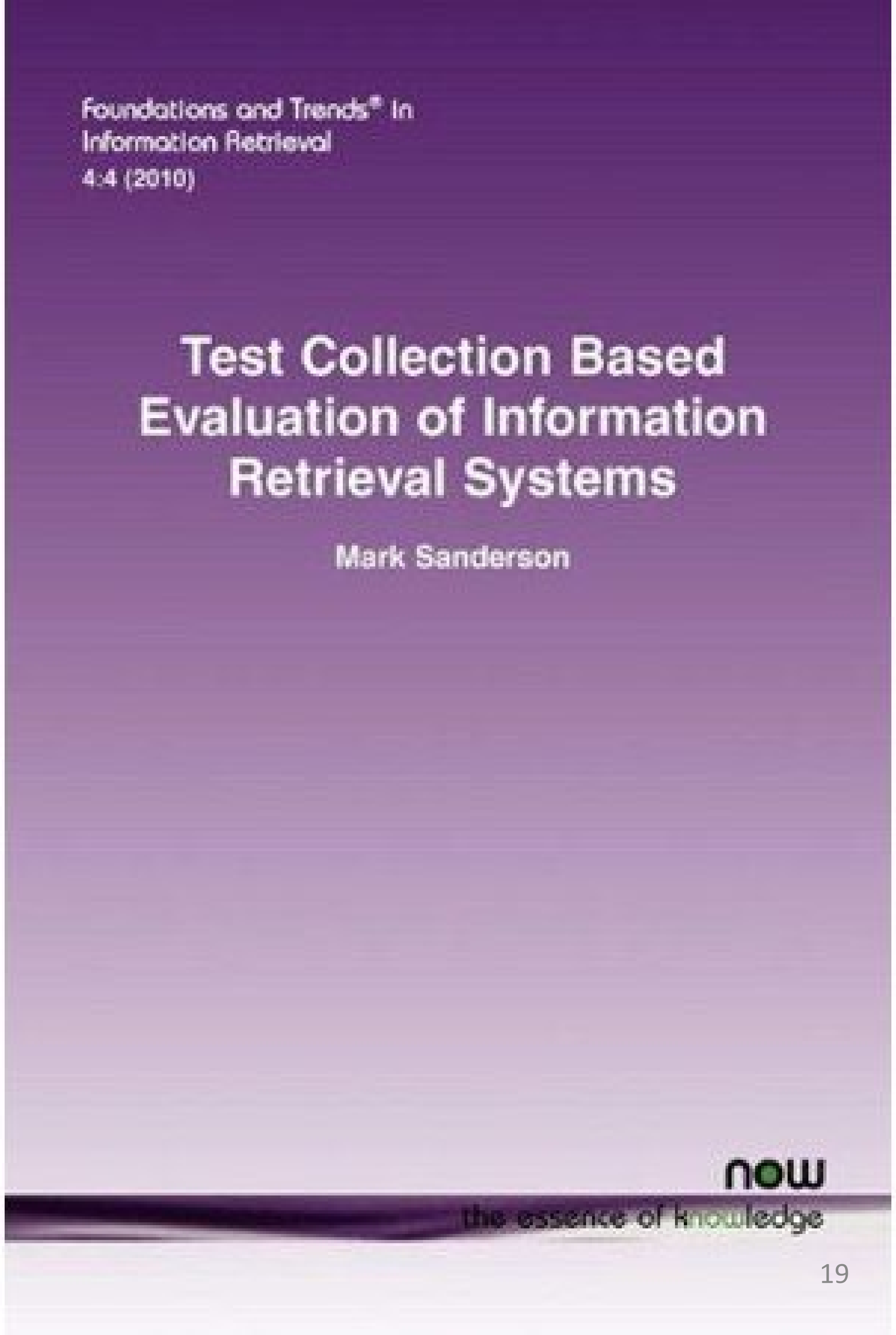
- Search-based:
  - Rather than read all documents, use manually guided search
  - Read retrieved documents until convinced that all relevant documents found.
- Pooling (Spark-Jones and Rijsbergen, 1975):
  - The most common technique used by TREC
  - Retrieve documents for each query by different retrieval techniques
  - Judge the union of the top n documents retrieved by each technique
- Sampling:
  - Possible to estimate size of true relevant set by sampling

- All of these approaches provide incomplete relevance judgments.
- How should unjudged documents be treated?

Assessor Consistency

- Relevance judgments are subjective
  - Is inconsistency of assessors a concern?
- Studies mostly concluded that the inconsistency didn't affect relative ranking of systems
  - Lesk & Salton (1968): assessors mostly disagree on documents at lower ranks, but measures are more affected by top-ranked documents
  - Cleverdon (1970), Burgin (1992): similar conclusions
  - Harman (1994): 80% agreement between TREC assessors
  - Schultz (1967): Judgments on relative ranking of documents are more consistent

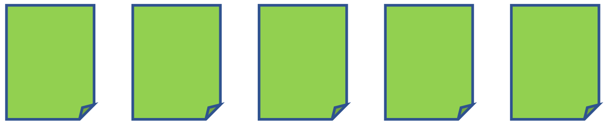
Want to Learn More?

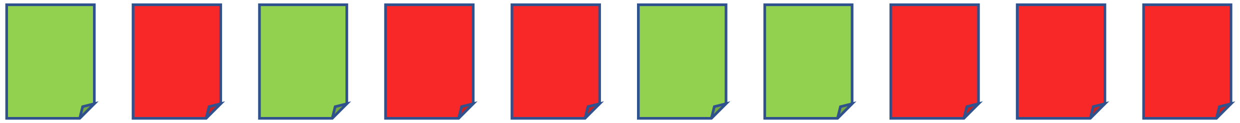


IR Evaluation Metrics

Precision

$$\text{precision} = \frac{|\text{relevant} \cap \text{retrieved}|}{|\text{retrieved}|} = \frac{\#\text{ of relevant retrieved documents}}{\#\text{ of retrieved documents}}$$

All relevant documents: 

Retrieval list: 

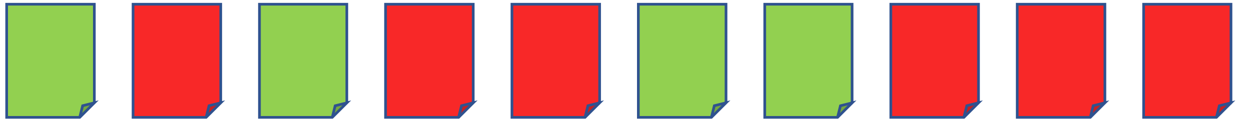
Precision = 0.40

Cleverdon and Keen, 1966

Recall

$$\text{recall} = \frac{|\text{relevant} \cap \text{retrieved}|}{|\text{relevant}|} = \frac{\#\text{ of relevant retrieved documents}}{\#\text{ of relevant documents}}$$

All relevant documents: 

Retrieval list: 

Recall = 0.80

Cleverdon and Keen, 1966

Precision and Recall

The precision-recall trade-off:

- Going to a deeper rank generally reduces precision and increases recall.

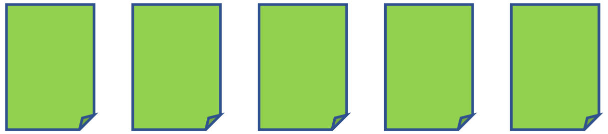
Which one is more important?

F-measure

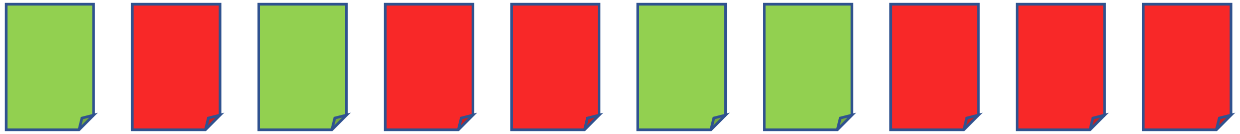
The harmonic mean of precision (P) and recall (R).

$$\text{F}-\text{measure} = \frac{1}{\beta\left(\frac{1}{P}\right) + (1-\beta)\left(\frac{1}{R}\right)}, \qquad \text{F1} = \frac{1}{\frac{1}{2}\left(\frac{1}{P}\right) + \frac{1}{2}\left(\frac{1}{R}\right)} = \frac{2PR}{P+R}$$

All relevant documents:



Retrieval list:

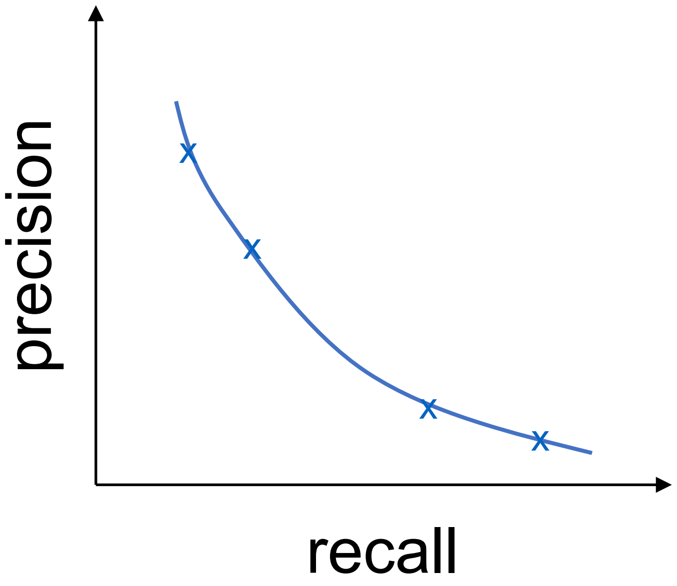


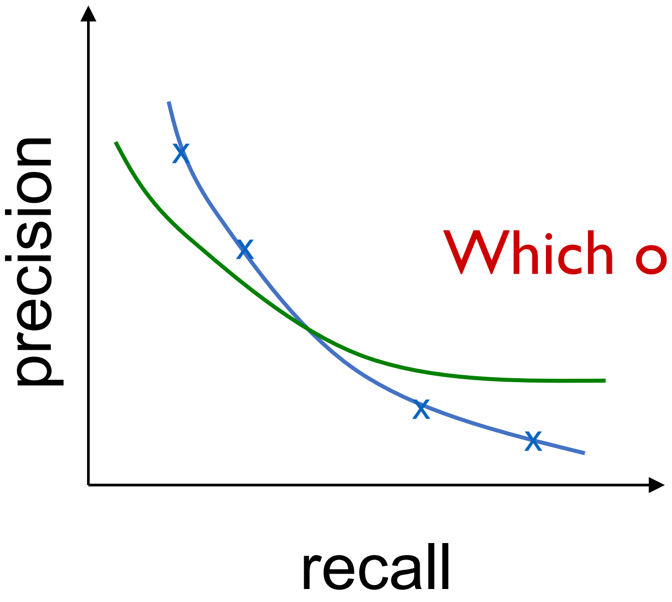
F1 = $\frac{2*0.4*0.8}{0.4+0.8} = 0.53$

Rijsbergen, 1979

Precision-Recall Curve

- Compute precision at every recall point
- Plot the precision-recall (PR) curve





Which one is better?

Average Precision (AP)

The most common metric in TREC.

$$AP = \frac{1}{|R|} \sum_{i=1}^{n} \mathrm{Prec}(i).\, \mathrm{Relevance}(i)$$

$|R|$: Total number of relevant documents

n: Length of the rank list

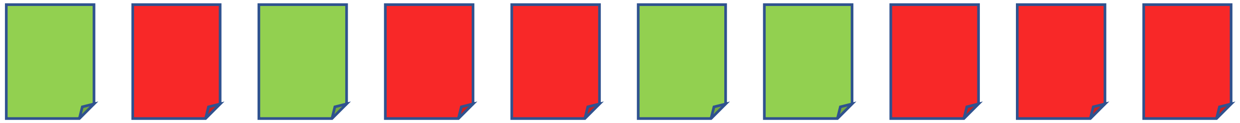
$\mathrm{Prec}(i)$: Precision of the top i documents

$\mathrm{Relevance}(i)$: 1 if document i is relevant, otherwise 0.

Average Precision (AP)

All relevant documents:

Retrieval list:



AP = 1/5 (1+2/3+3/6+4/7) = 0.5476

Reciprocal Rank (RR)

$$RR = \frac{1}{r}$$

where r is the rank of the first relevant retrieved document.

- If no relevant document retrieved, then $RR = 0$.

- Only considered the rank of the first relevant answer
- Suitable for some applications, such as:
  - Web search (why?)
  - Email search (why?)

Normalized Discounted Cumulative Gain (nDCG)

- Useful for graded relevance judgments
- Cumulative Gain at rank n:
  - Let the relevance labels for the top n documents be $R_1, R_2, \ldots, R_n$
  - $CG = g(R_1) + g(R_2) + \cdots + g(R_n)$
  - $g(R_i) = R_i$ or $g(R_i) = 2^{R_i} - 1$
- Discounted Cumulative Gain at rank n:
  - $DCG = \frac{g(R_1)}{\log_2(2)} + \frac{g(R_2)}{\log_2(3)} + \frac{g(R_3)}{\log_2(4)} + \cdots + \frac{g(R_n)}{\log_2(n+1)}$
- Normalized Discounted Cumulative Gain at rank n:
  - Dividing DCG by the ideal DCG

Kalervo Järvelin and Jaana Kekäläinen. *"Cumulated gain-based evaluation of IR techniques."* TOIS 2002.

nDCG Example

All relevant documents: 3 3 2 1 1

Retrieval list: 3 0 2 0 0 1 3 0 0 0

Aggregation over Queries

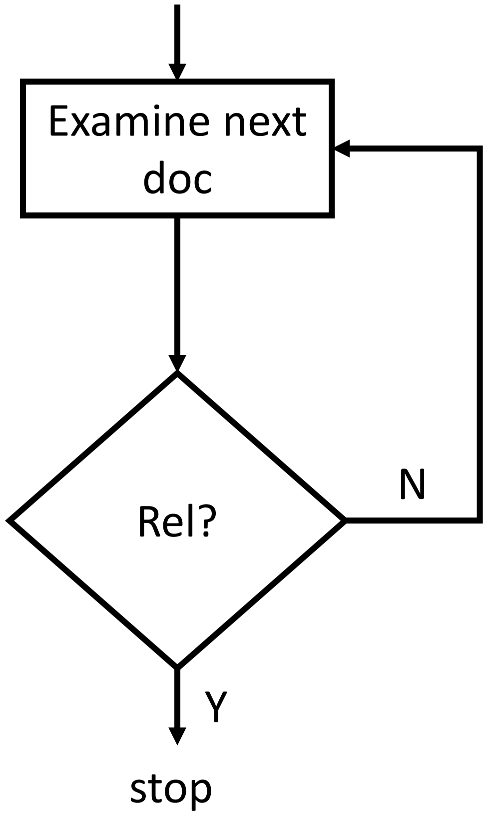
- Arithmetic mean is a common technique to summarize the retrieval performance for a set of queries
- precision, recall, AP, RR, nDCG -> precision, recall, **MAP, MRR**, nDCG
- Interpolation for precision-recall curve
- Alternative:
  - Geometric mean for AP: gMAP
- Information loss?

Limitations

- Aggregation over queries hides many information.
- Most metrics only consider binary relevance judgments.
- Redundancy, novelty, and diversity are not considered.
- User-independent

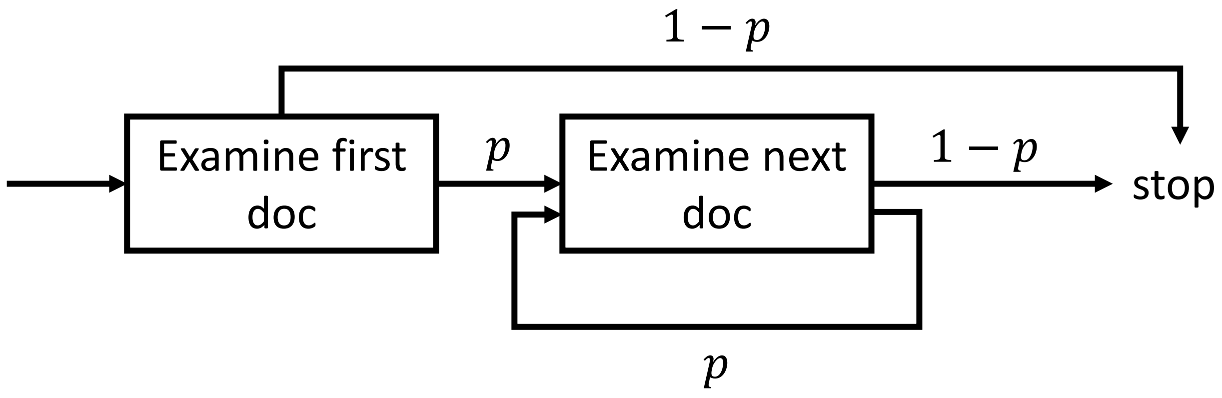
User Browsing Models (UBMs)

UBM for Reciprocal Rank



RR is an effort-based metric

UBM for Modeling Patient and Impatient Users



- Larger p can be associated to more patient users.

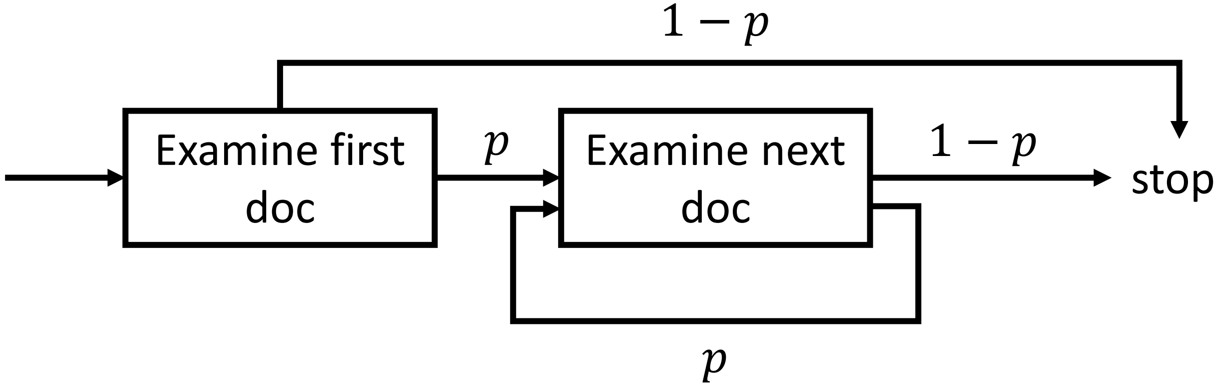
Alistair Moffat and Justin Zobel. *“Rank-biased precision for measurement of retrieval effectiveness.”* TOIS 2008.

A Simple **Expected Utility** Function

$$\text{Expected Utility} = \frac{1}{N}\sum_{k=1} P(E_k = 1).\, R_k$$

- N*:* Expected number of examined documents
- $P(E_k = 1)$: examination probability of the k^{th} document in the ranked list by the user
- R_k: relevance label

UBM for Modeling Patient and Impatient Users



$$\text{Expected Utility} = \frac{1}{N}\sum_{k=1} P(E_k = 1).\, R_k$$

$$= \frac{1}{N}\sum_{k=1} p^{k-1}.\, R_k$$

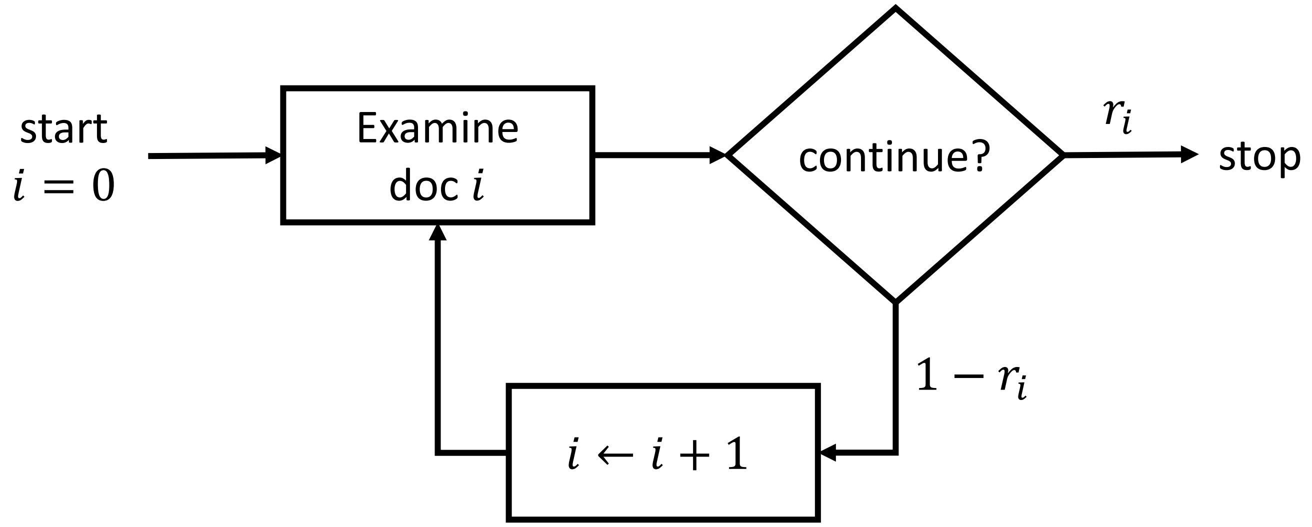
$$N = \sum_{k=1}^{\infty} k.\, p^{k-1}.\, (1-p) = \frac{1}{1-p}$$

Ranked-biased Precision: $(1-p)\sum_{k=1} p^{k-1}.\, R_k$

Alistair Moffat and Justin Zobel. *"Rank-biased precision for measurement of retrieval effectiveness."* TOIS 2008.

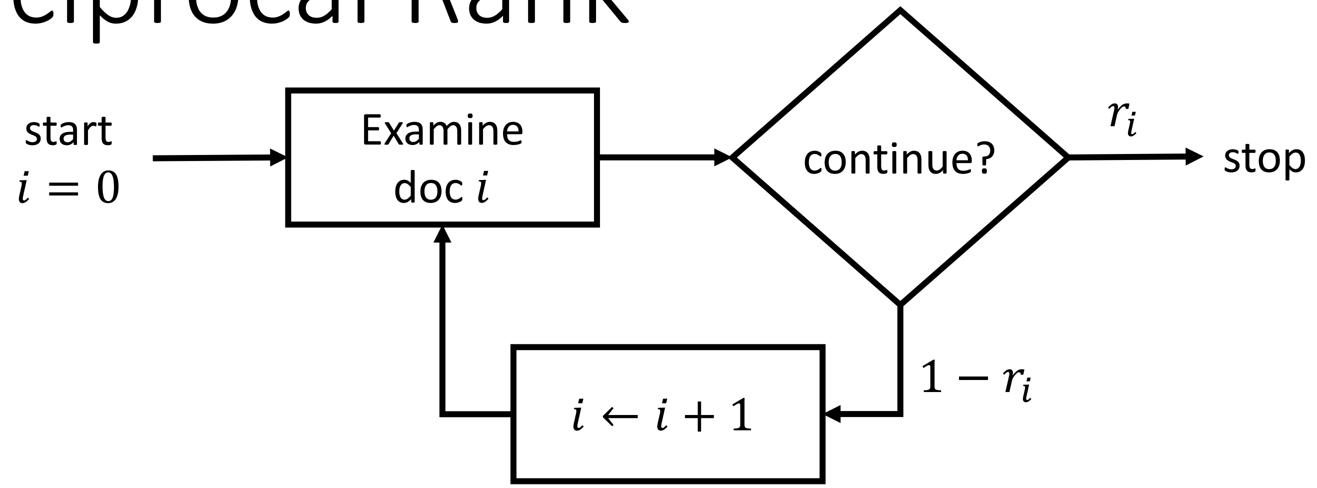
UBM for Expected Reciprocal Rank

- Reciprocal Rank does not support graded relevance.
- Metrics like NDCG and RBP are utility based (not effort based)



Olivier Chapelle et al. *"Expected Reciprocal Rank for Graded Relevance."* CIKM 2009.

UBM for Expected Reciprocal Rank



Expected Reciprocal Rank:

$$\sum_{k=1} \frac{1}{k} \prod_{i=1}^{k-1} (1 - r_i) r_k$$

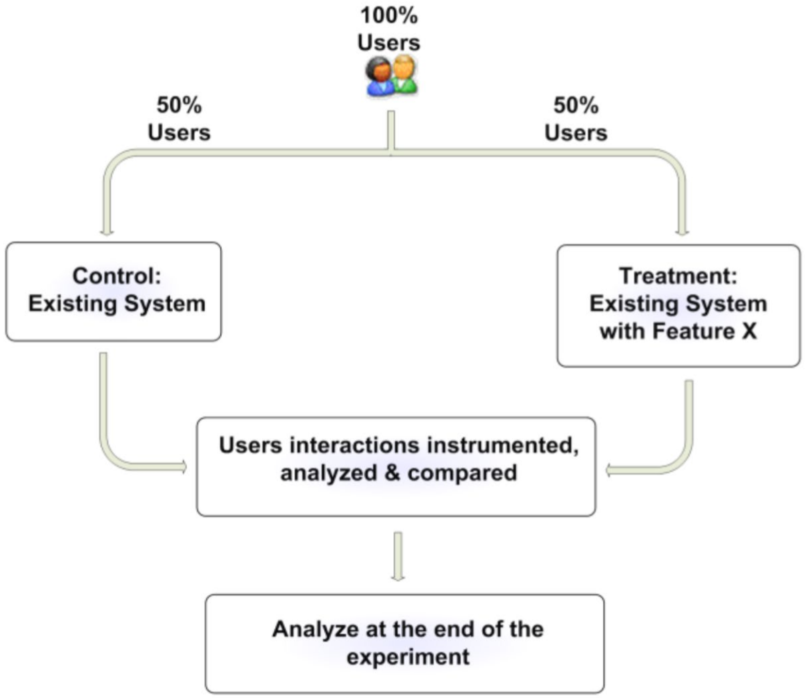
$$r_i = \frac{2^{g_i} - 1}{2^{G-1}}, \qquad g_i \in \{0, 1, \cdots, G-1\}$$

Olivier Chapelle et al. *"Expected Reciprocal Rank for Graded Relevance."* CIKM 2009.

A/B Testing

High-Level Overview of A/B Testing

- What to measure?



Summary

- Definition of Relevance
- Evaluation Methodologies in IR
- Test Collection Creation
- IR Metrics
  - P@k, R@k, F1, RR, AP, NDCG, …
- User Browsing Models
- A/B Testing

Questions?