Evaluating IR Systems

INFORMATION RETRIEVAL AND RECOMMENDER SYSTEMS

Georgios Peikos

University of Milano-Bicocca, Milan, Italy

Department of Informatics, Systems, and Communication (DISCo)

Today's Lecture – Evaluating IR Systems

Types of evaluation

Offline Evaluation

Experimental Collections

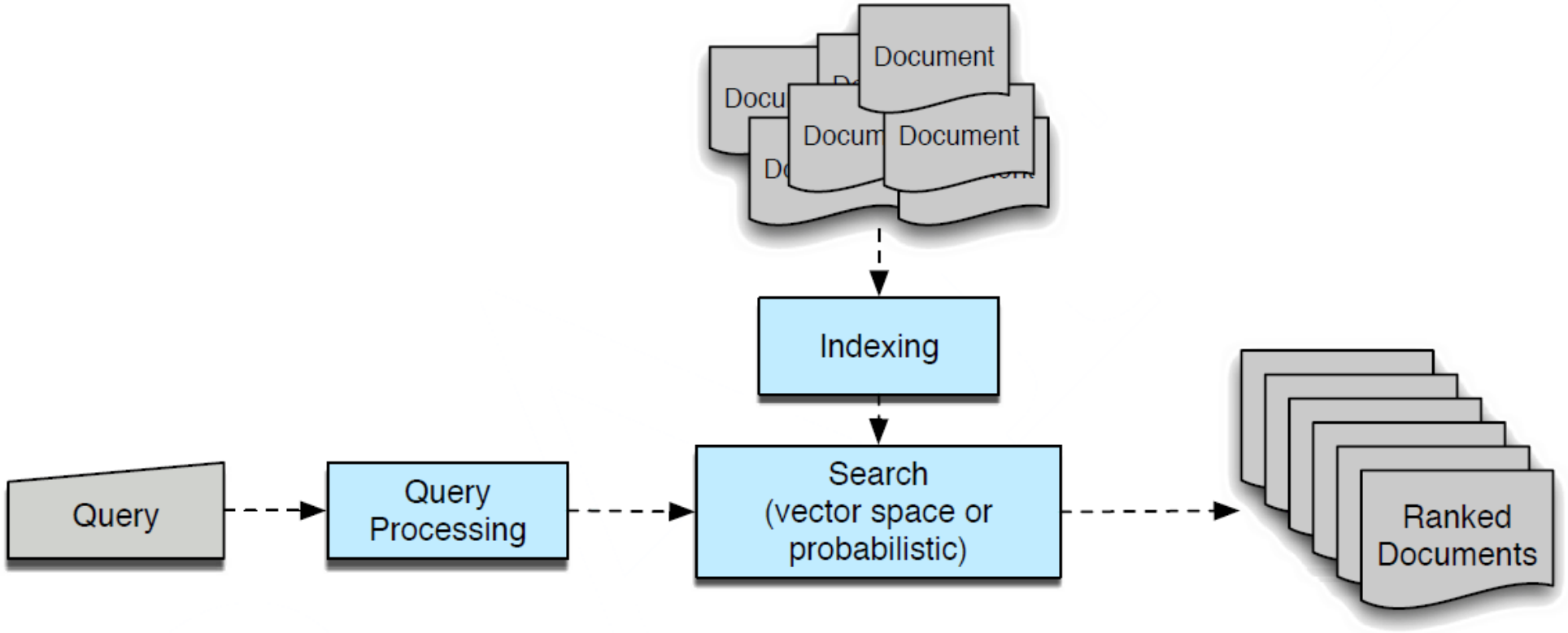
Evaluation Measures

Further Reading Resources

Evaluation Basics

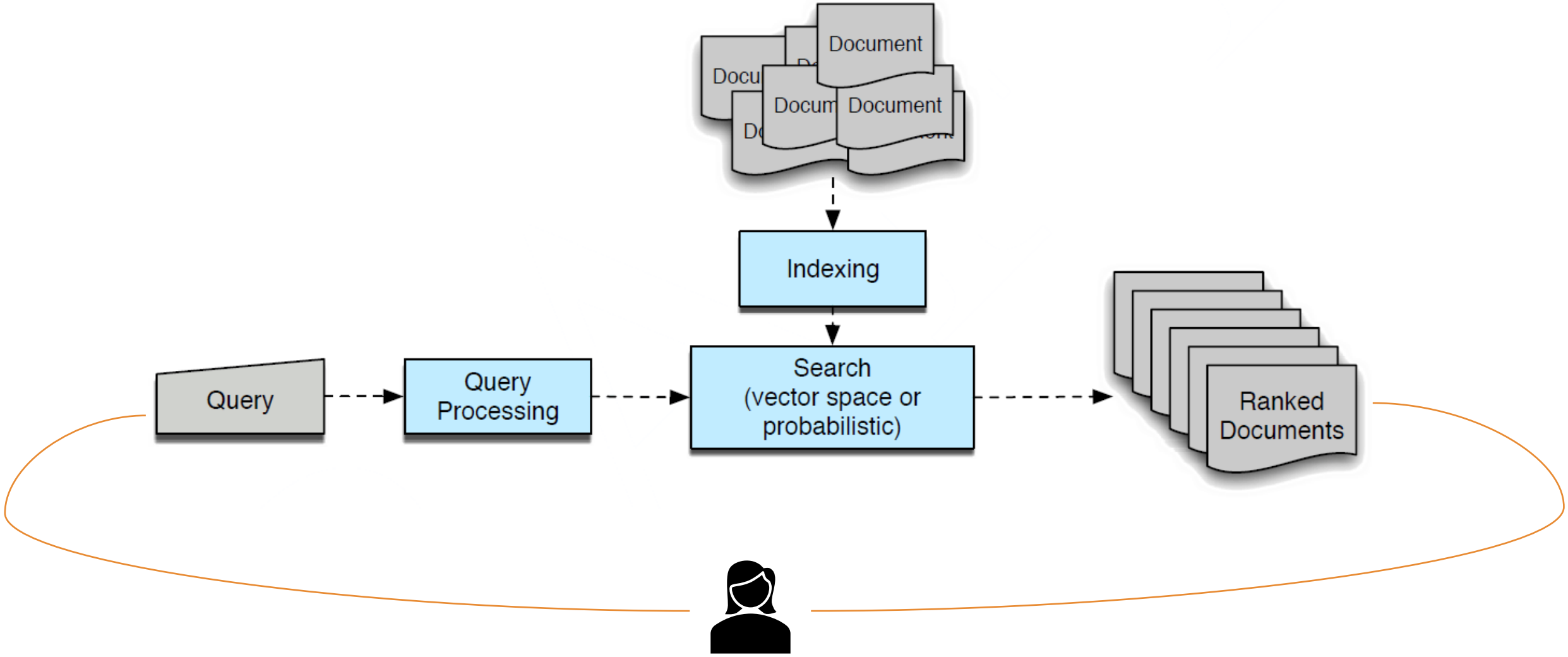
IRS Overview – Describe an IR System

IRS Overview – An IR System

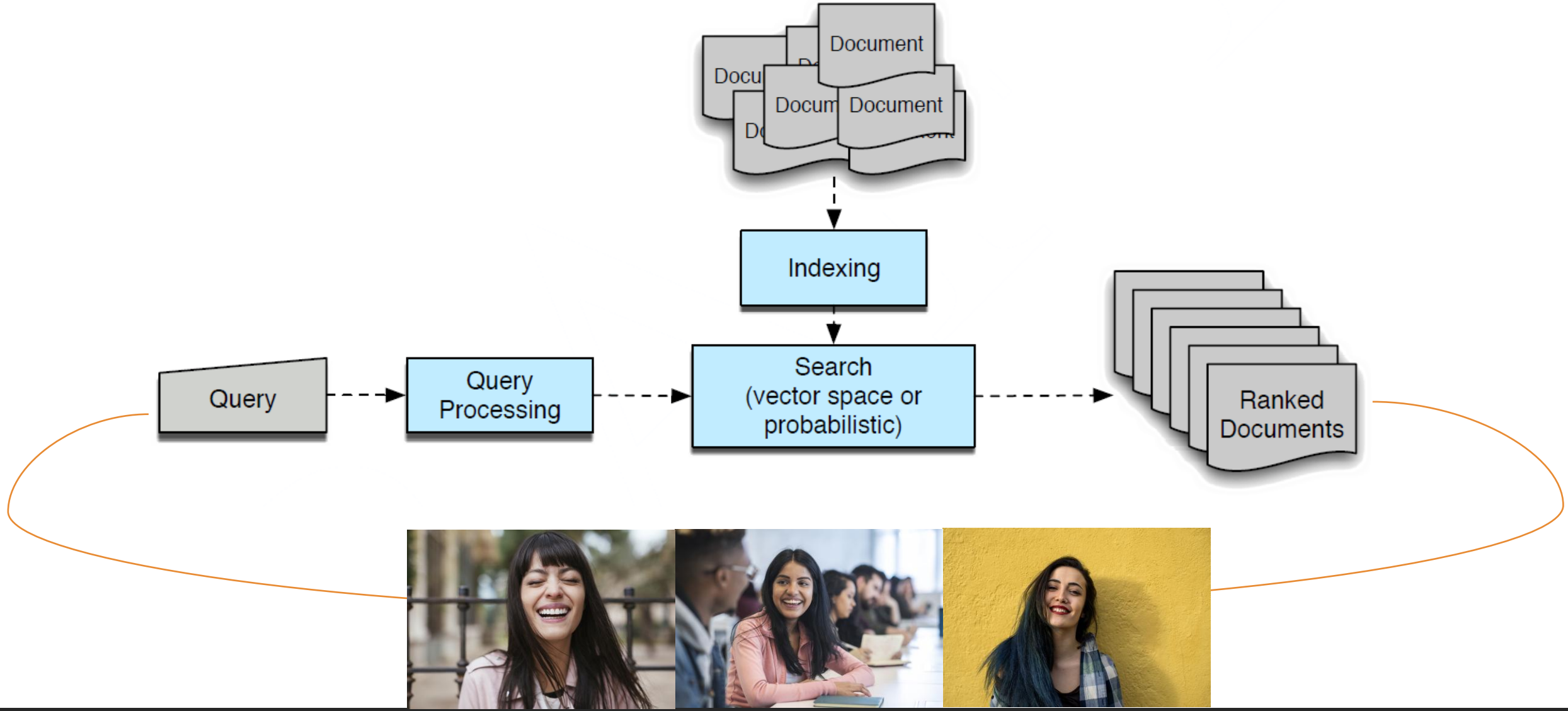


IRS Overview – What is its goal?

IRS Overview – Goal of an IRS



IRS Overview – Goal of an IRS



How do we make people happy?

By creating a **good** Information Retrieval System

How can we be sure that we created a good IRS?

IR Evaluation – Evaluate our Systems

IR Evaluation – Evaluate our Systems

Efficiency.

Algorithmic costs of IR systems

Computation Resources

Effectiveness.

Ability of the IRS to retrieve and rank information items properly.

IR Evaluation – Evaluate our Systems

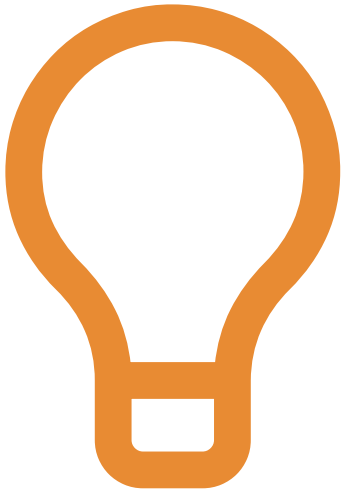
Efficiency can be express formally, for instance measuring the retrieval latency or report the system's computational complexity.

Effectiveness can be assessed only experimentally!

An IRS is effective if

it can **retrieve** as **many relevant** documents as possible

while **minimizing** the number of **non relevant** documents retrieved



We will see how to evaluate the **effectiveness** of an IR system!

Evaluating the effectiveness of IRS

IRS Effectiveness Evaluation – Online vs Offline

Online evaluation exploits the **behavior and interactions** of **real users**, while they engage with an IR system.

Online evaluation requires access to a large amount of log data from real users.

These evaluations are hard to **reproduce** as the evaluation outcome depends on the **users** and the **interpretation** of their implicit signals (clicks, highlighting text, etc.).

A/B testing can be one way to evaluate an IR system.

Offline IRS Effectiveness Evaluation

IRS Effectiveness Evaluation – Offline

We said that **IRS effectiveness** can be assessed **experimentally**!

We said that an effective search engine needs to retrieve as many **relevant** documents as possible while minimizing the number of **non relevant** documents retrieved.

IRS Effectiveness Evaluation – Offline

We said that IRS effectiveness can be assessed experimentally!

We said that an effective search engine needs to retrieve as many **relevant** documents as possible while minimizing the number of **non relevant documents** retrieved.

Who decides which documents are relevant or not to a query?

How can a document deemed to be relevant to a query?

How is relevance perceived?

IRS Effectiveness Evaluation – Offline (Cranfield paradigm)

We said that IRS effectiveness can be assessed experimentally!

We said that an effective search engine needs to retrieve as many **relevant** documents as possible while minimizing the number of **non relevant documents** retrieved.

Who decides which documents are relevant or not to a query?

How can a document deemed to be relevant to a query?

How is relevance perceived?

Considering all these questions, offline IRS effectiveness is being measured based on the Cranfield Paradigm.

IRS Effectiveness Evaluation – Cranfield Paradigm

The **Cranfield Paradigm** has been proposed by **Cyril W. Cleverdon** in mid 1960s

It leverages what we call **Experimental Collections** or **IR Benchmark Collections**

IRS Effectiveness Evaluation – Experimental Collections

Document Collection (corpus): A static set of documents that remains constant during evaluations, allowing different systems to be compared under the same conditions.

Queries: A predefined set of queries, representing the information needs of users.

Relevance Judgments (qrels): Assessments of whether specific documents are relevant or irrelevant to each query. These judgments are typically created by human annotators and provide a gold standard against which retrieval results can be measured.

As a result, this evaluation ensures comparability and repeatability (reproducibility) of the experiments.

Experimental Collections – Requirements of an ideal

Document Collection (corpus): Should be more than **10,000**.

Queries: Usually, 50, but they should be more than **250**.

Relevance Judgments (qrels): Ideally, all documents for all queries; commonly **30k-75k query-document pairs**.

Usually binary or 3-graded

Still an open research **which documents** should be **annotated**, as we can not annotate all of them.

We do document pooling to identify documents that should be annotated.

Experimental Collections – Assumptions

The **relevance** of a document to a user is considered binary (relevant / not relevant) or 3-graded (highly relevant, somehow relevant, irrelevant)

The **relevance** of a document is **independent** of the relevance of other documents. **What do you think about this?**

IRS Effectiveness Evaluation – Relevance Judgements

Relevance Judgments (qrels): Assessments of whether specific documents are relevant or irrelevant to each query.

These judgments are typically created by human annotators and provide a gold standard against which retrieval results can be measured.

So, which documents they annotate?

IRS Effectiveness Evaluation – Relevance Judgements

Relevance Judgments (qrels): Assessments of whether specific documents are relevant or irrelevant to each query.

These judgments are typically created by human annotators and provide a gold standard against which retrieval results can be measured.

So, which documents they annotate?

- **Open research question**, the documents are selected based on document pooling (see additional resources).
- Not pooled/not assessed documents are typically assumed to **be not relevant**.

IRS Effectiveness Evaluation – Relevance Judgements

Relevance Judgments (qrels): Assessments of whether specific documents are relevant or irrelevant to each query.

These judgments are typically created by human annotators and provide a gold standard against which retrieval results can be measured.

So, which documents they annotate?

- Open research question, the documents are selected based on document pooling (see additional resources).
- Not pooled/not assessed documents are typically assumed to be not relevant.

How they decide what makes a document relevant to a query?

IRS Effectiveness Evaluation – Relevance Judgements

Relevance Judgments (qrels): Assessments of whether specific documents are relevant or irrelevant to each query.

These judgments are typically created by human annotators and provide a gold standard against which retrieval results can be measured.

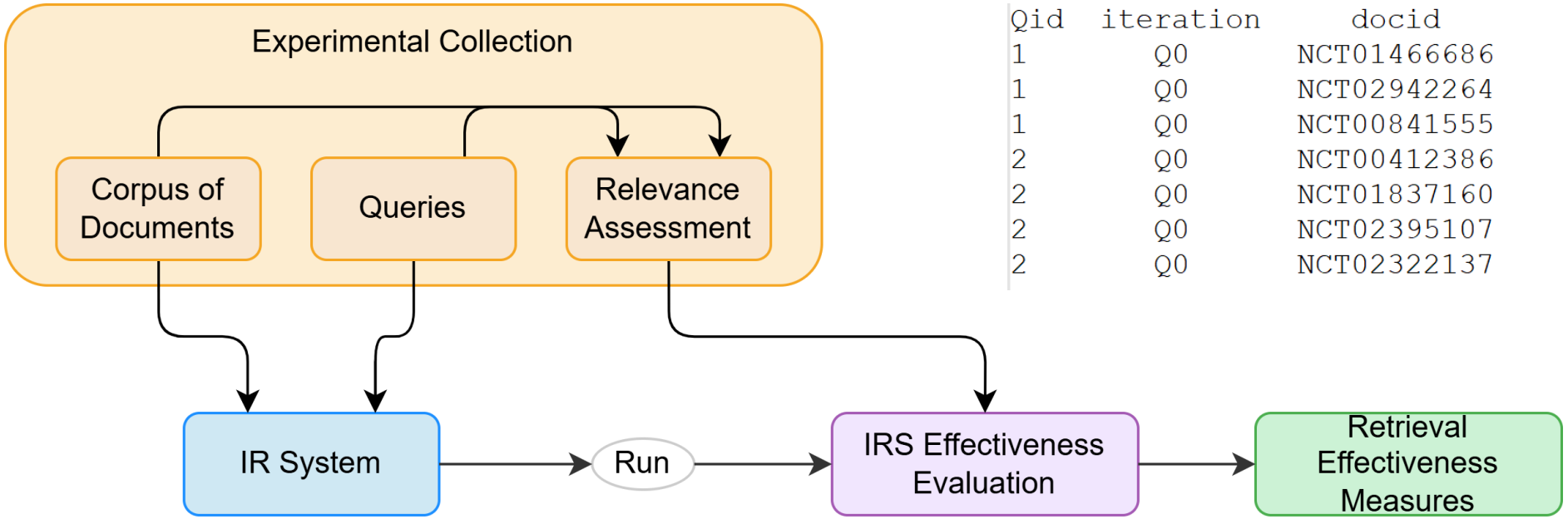
So, which documents they annotate?

- Open research question, the documents are selected based on document pooling (see additional resources).
- Not pooled/not assessed documents are typically assumed to be not relevant.

How they decide what makes a document relevant to a query?

- The annotators are given **clear instructions**. Often, there is a test annotation, where they discuss are resolve disagreements.
- Even then, the inter-assessor agreement can be low (relevance is subjective).

IRS Effectiveness Evaluation – How?



Relevance Assessments

Qid	iteration	docid	relevance
1	Q0	NCT01466686	1
1	Q0	NCT02942264	2
1	Q0	NCT00841555	1
2	Q0	NCT00412386	0
2	Q0	NCT01837160	1
2	Q0	NCT02395107	2
2	Q0	NCT02322137	0

Qid	iteration	docid	rank	score	runid
1	Q0	NCT01466686	1	0.860	Experiment_name
1	Q0	NCT02942264	2	0.840	Experiment_name
1	Q0	NCT00841555	3	0.810	Experiment_name
2	Q0	NCT00412386	1	0.932	Experiment_name
2	Q0	NCT01837160	2	0.863	Experiment_name
2	Q0	NCT02395107	3	0.860	Experiment_name
..					
2	Q0	NCT02322137	1000	0.060	Experiment_name

Output of an IR System

Retrieval Effectiveness Measures

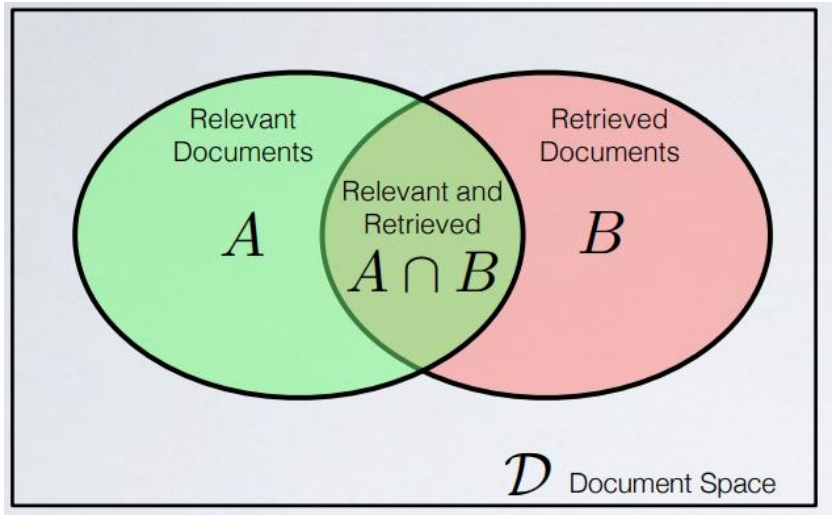
Measures – Precision and Recall (set-based)

Precision is the proportion of retrieved documents that are actually relevant.

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents}}$$

Recall is the proportion of relevant documents that are actually retrieved.

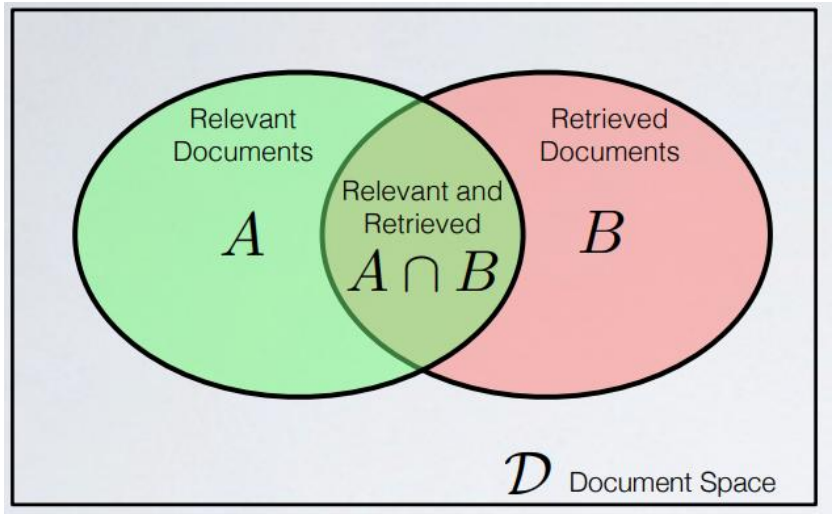
$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$



Measures – Precision and Recall (set-based)

Precision is the proportion of retrieved documents that are actually relevant.

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents}}$$



Recall is the proportion of relevant documents that are actually retrieved.

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

F-measure is the harmonic mean of Precision and Recall, summarizing them into a single score.

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

IR Evaluation – Precision@K (Rank-Based)

Select a rank threshold K

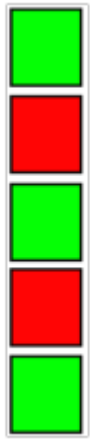
Compute % relevant in top K

Ignores documents ranked lower than K

Prec@3 of 2/3

Prec@4 of 2/4

Prec@5 of 3/5



Similarly for the **Recall@K**, as we know the total relevant documents per query.

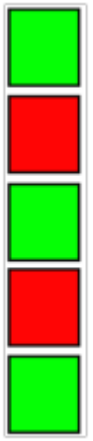
IR Evaluation – Mean Average Precision

Consider rank position of each relevant document

K1,K3, K5,…,KR

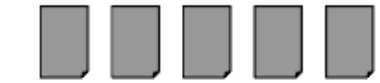
Compute Precision@K for each *K1,K2,…,KR*

Average precision = **average of P@K**

The Average Precision of  is $\frac{1}{3} \cdot \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$

MAP is the Average Precision across **all queries in the collection**.

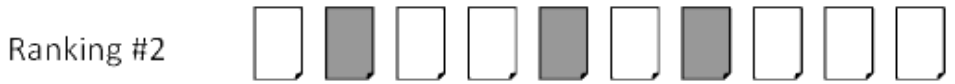
IR Evaluation – Mean Average Precision



= relevant documents for query 1

Ranking #1										
Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

= relevant documents for query 2



Ranking #2										
Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

IR Evaluation – Discounted Cumulated Gain

As we said, relevance can be assessed also in a graded scale. To measure IRS effectiveness based on graded relevance, one can leverage the Discounted Cumulative Gain.

It relied on a utility accumulation model, where the utility provided by a document is discounted proportionally to the rank position at which that document is retrieved.

The underlying idea is that a relevant document retrieved at the top of the ranking can be more useful to the user than, the same document retrieved at the bottom of the ranking.

IR Evaluation – Normalized Discounted Cumulative Gain

From Cumulative Gain (CG), one can derive Discounted Cumulative Gain (DCG) by introducing a logarithmic discounting element to account for the position of each information item, acknowledging that items retrieved earlier are more valuable to the user. The formula transitions from:

$$\text{CG@}k = \sum_{i=1}^{k} \text{rel}_i \text{ to DCG@}k = \sum_{i=1}^{k} \frac{\text{rel}_i}{\log_2(i+1)},$$

Normalized Discounted Cumulative Gain (nDCG) further refines DCG by normalizing it against a perfect ranking to ensure the values lie between 0 and 1, making comparisons across queries and systems fairer. The formula is:

$$\text{nDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k},$$

by relevance in descending order.


Questions?

IR Evaluation – Resources

Evaluation of IR Systems: https://www.dei.unipd.it/~ferro/papers/2024/IR-Book2024-FM.pdf

Tool for IR Evaluation: https://github.com/GiorgosPeikos/ASPIRE

ASPIRE is available online: https://aspire-ir-eval.streamlit.app/