بسم الله الرحمن الرحيم

دورة "استرجاع المعلومات" باللغة العربية ـ صيف ٢٠٢١

Information Retrieval – Summer 2021

6. Query Expansion

Tamer Elsayed

Qatar University

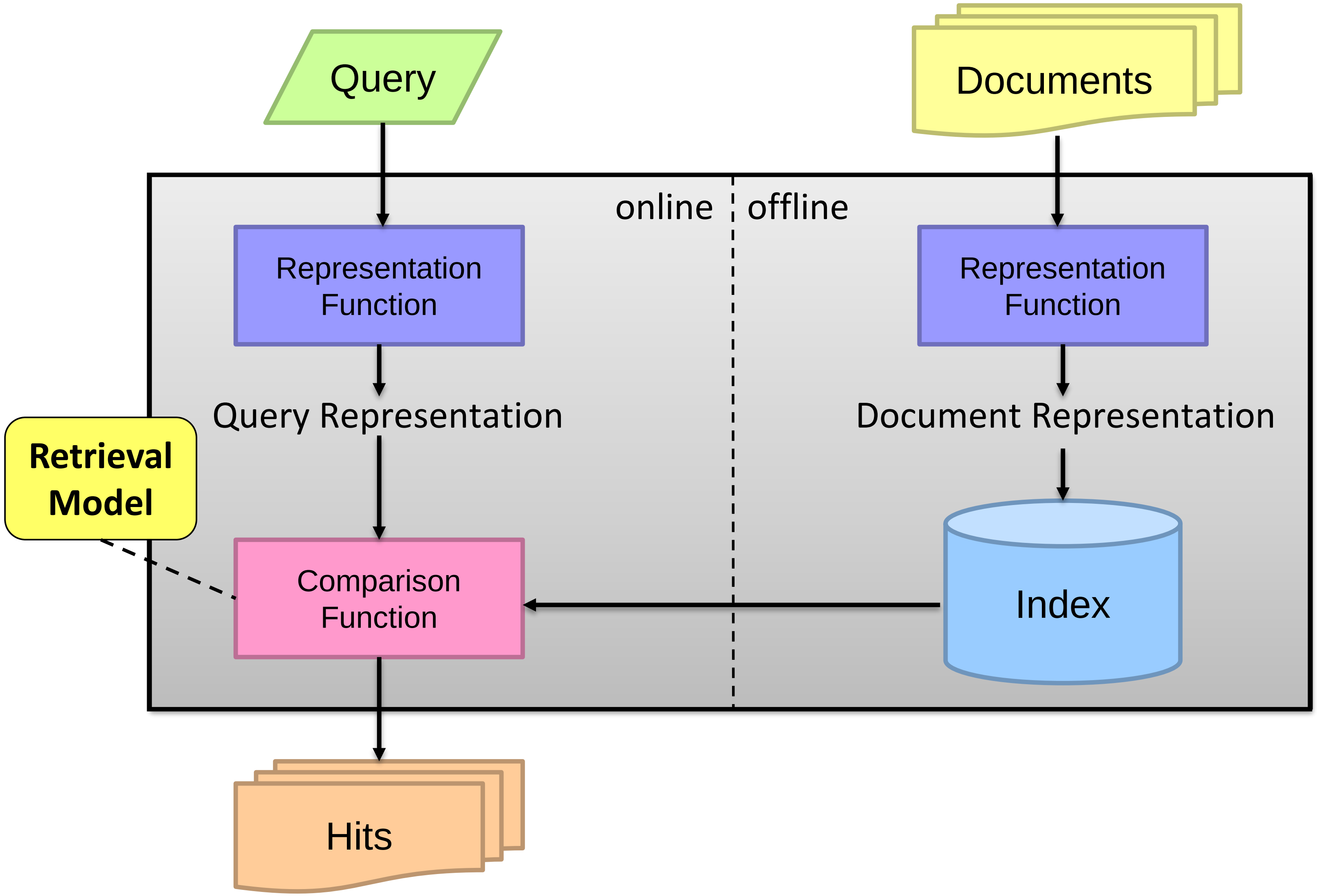
What is the application?

Query → **Search** → *Ranked List of Search Results*

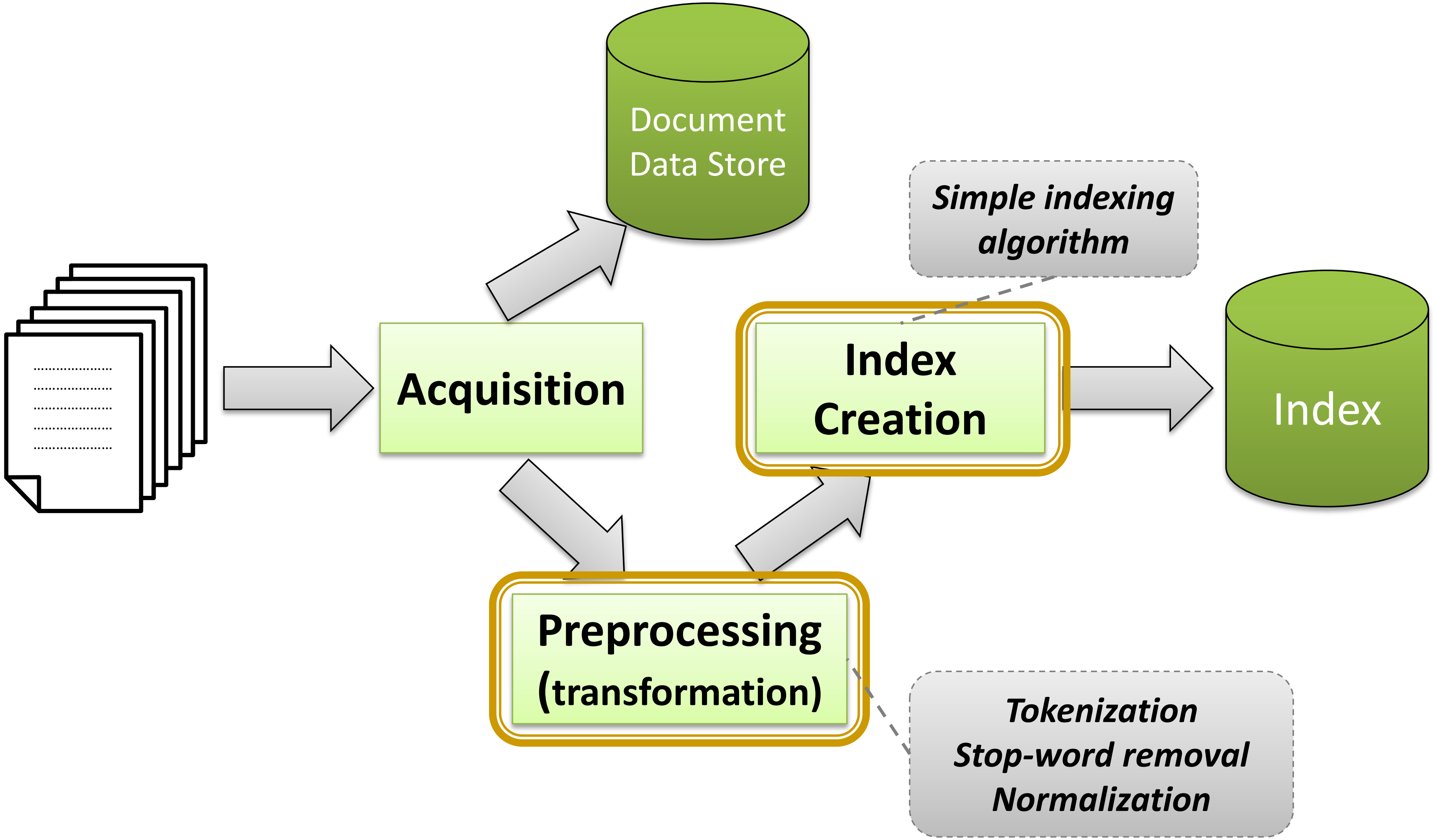
Ad-hoc Search

THE BIG PICTURE

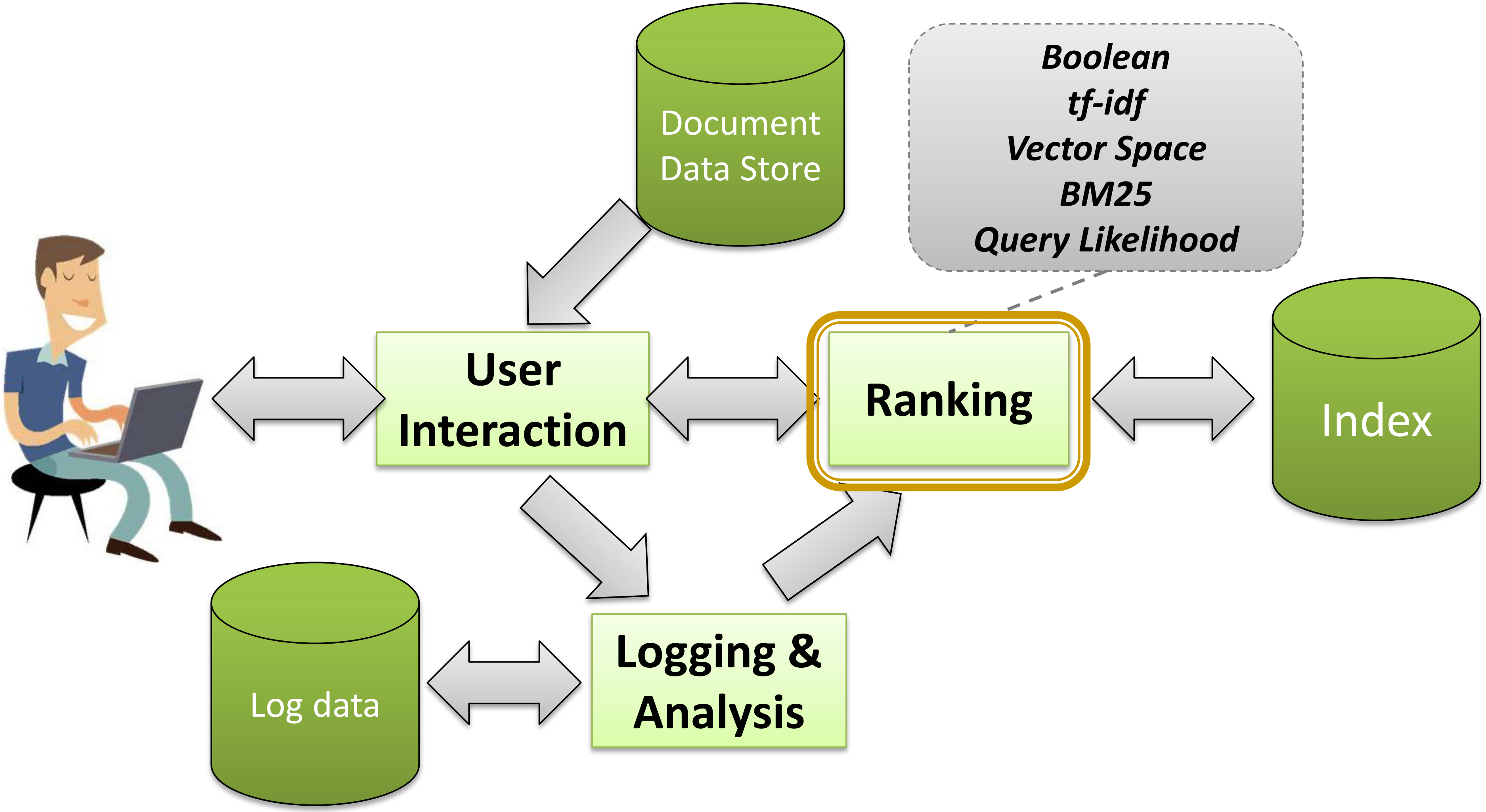
Inside the IR Black Box



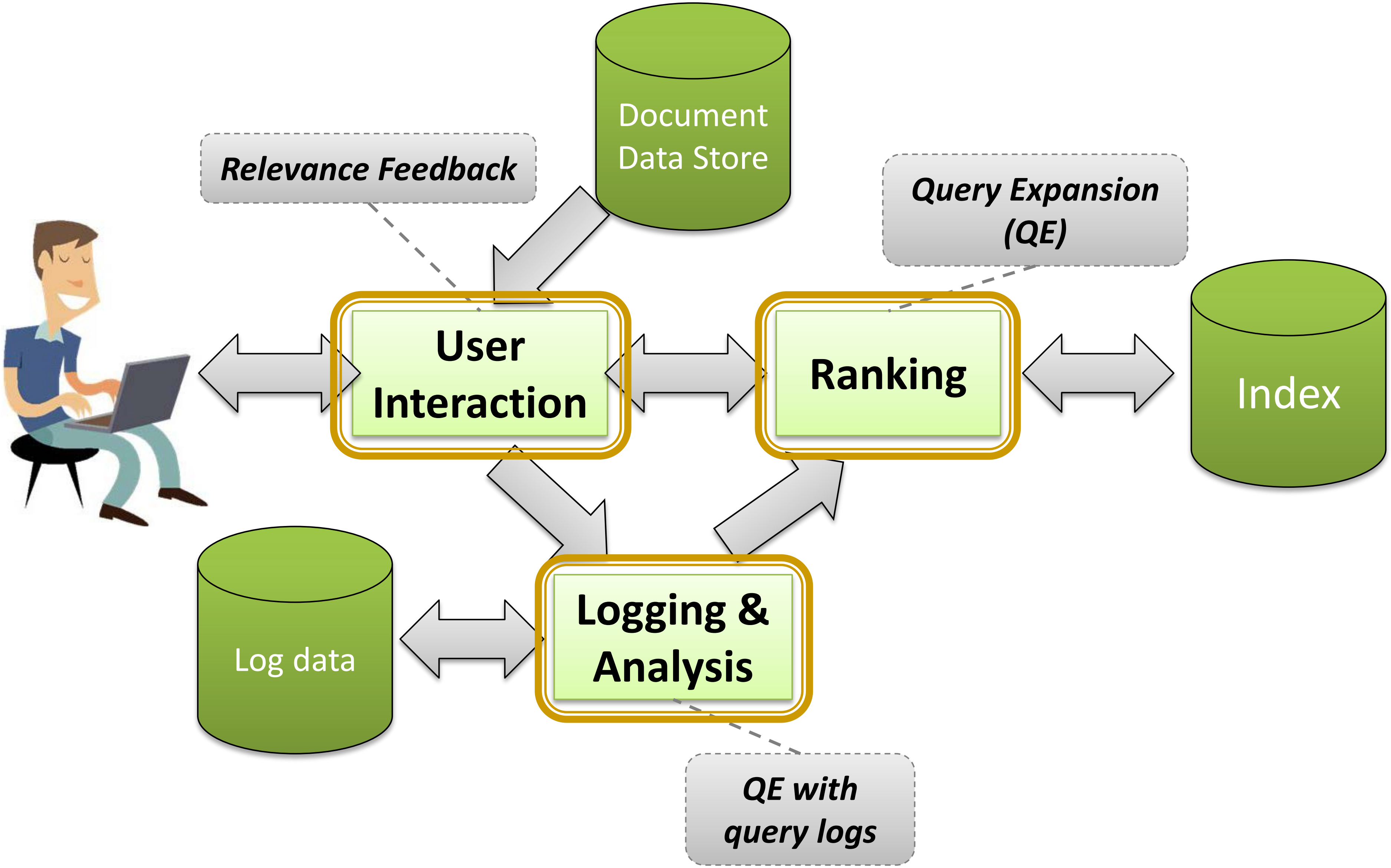
Indexing process (offline)



Search process (online)



Today



Document
Data Store
Relevance Feedback
Query Expansion
(QE)
User
Interaction
Ranking
Index
Logging &
Analysis
Log data
QE with
query logs

بسم الله الرحمن الرحيم

دورة "استرجاع المعلومات" باللغة العربية - صيف ٢٠٢١

Information Retrieval – Summer 2021

6. Query Expansion

Tamer Elsayed

Qatar University

Queries sometimes are <u>not</u> good representation of information needs

Query Expansion

Adding more words (related, relevant, or of the same meaning) to the query for better retrieval

Today's Roadmap

- Thesaurus-based methods
- Query Logs
- Relevance Feedback

THESAURUS-BASED METHODS

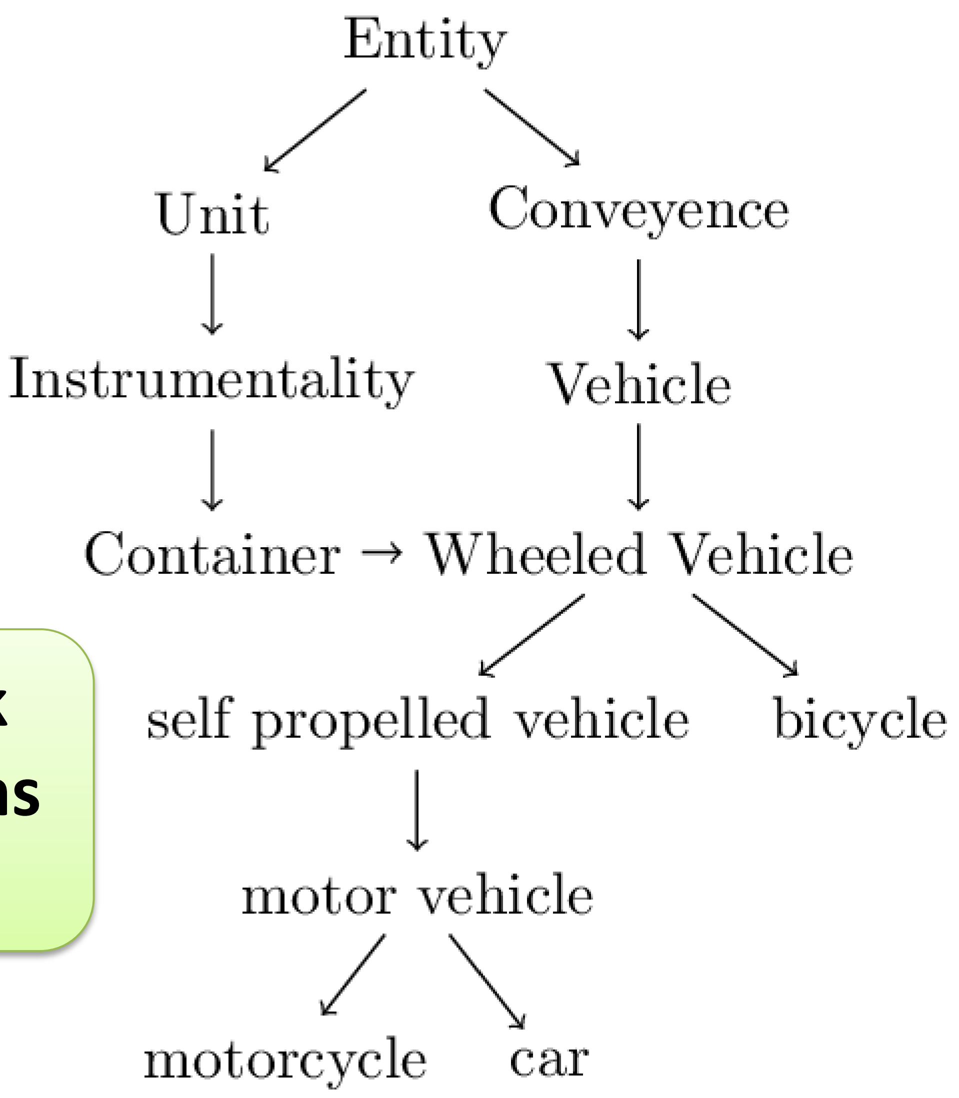
Thesaurus

- A data structure that shows words in groups of synonyms and related concepts.

- Manually built
- Automatically-built

Manually-built Thesaurus

- e.g., WordNet



Expand query by top k synonyms/related terms of each query term

problem?

Automatic Thesaurus

Main Idea

Words co-occurring in same documents/paragraphs are likely to be (in some sense) similar or related in meaning.

Using Term Vectors

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

Each term is now represented by a real-valued vector of weights $\in R^{|C|}$

Compute similarity between terms

Query Expansion

Each term is represented by a vector

Compute similarity between term vectors

offline

online

Expand query by top k similar terms of each query term

Using Term Association Measures

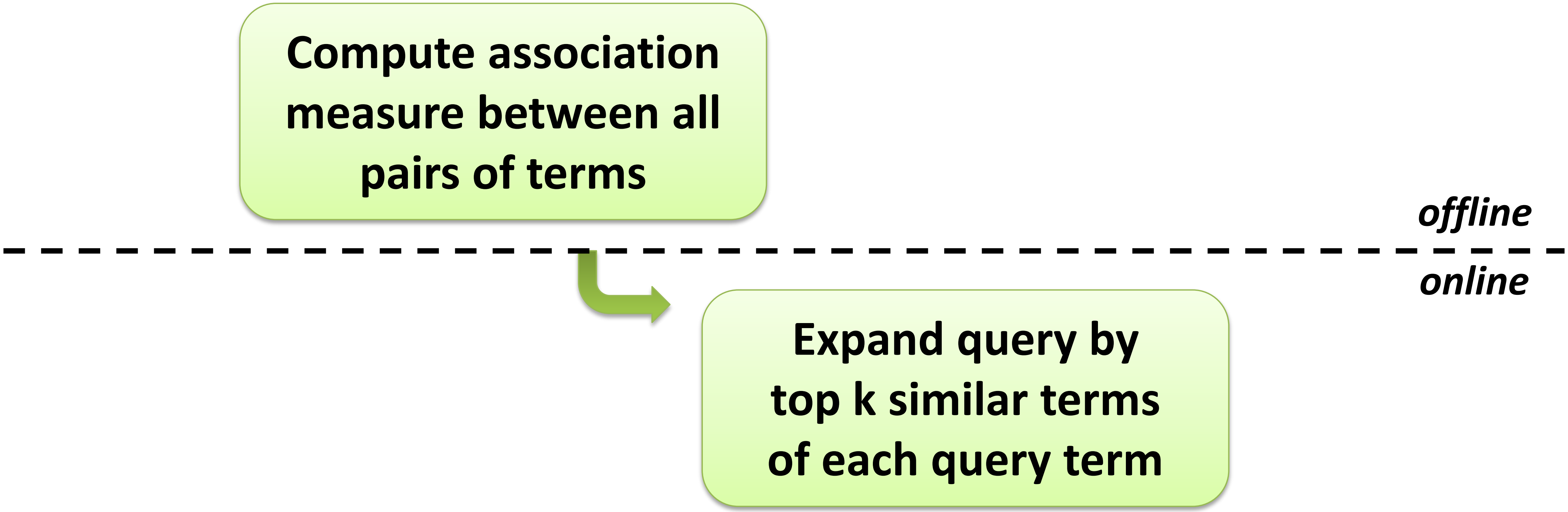
Examples:

- *Dice's Coefficient* $\frac{2.n_{ab}}{n_a+n_b} \stackrel{rank}{=} \frac{n_{ab}}{n_a+n_b}$

- *Mutual Information*

$$\log \frac{P(a,b)}{P(a)P(b)} = \log N.\frac{n_{ab}}{n_a.n_b} \stackrel{rank}{=} \frac{n_{ab}}{n_a.n_b}$$

Query Expansion



17

17

Ad-hoc search is the typical task we do on Google.

- Yes
- No

Terms that co-occur in very few paragraphs, but co-occur in many documents, will have ... assuming the measure is based on paragraph context.

- Yes
- No

Example: “world cup”

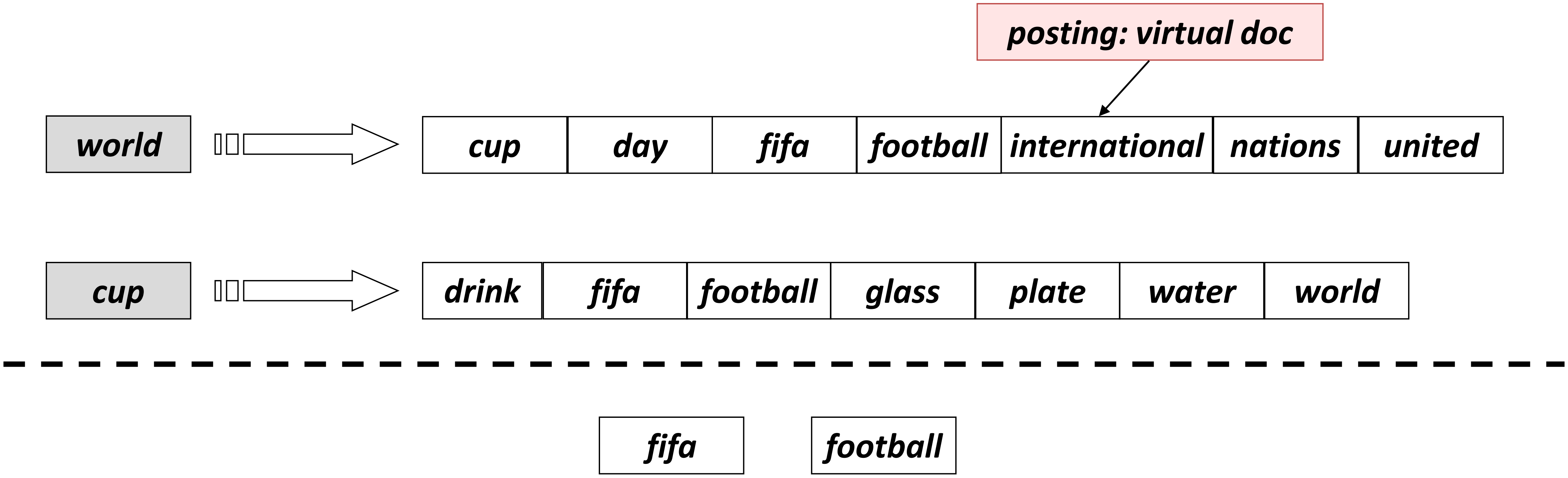
world	*cup*	*day*	*fifa*	*football*	*international*	*nations*	*united*
cup	*drink*	*fifa*	*football*	*glass*	*plate*	*water*	*world*

problem?

Using Context Vectors

- **Context vectors:** Represent words by other words that co-occur with them.
  - e.g., top 35 most strongly associated words.
- Use them as “virtual documents” in an inverted index.
- Rank words for a query by ranking virtual documents

Example: “world cup”



Query Expansion

Compute association measure between all pairs of terms

Build the "context" index

offline

---

online

given the query,
rank virtual docs (terms)

Expand the query with top k terms

Using Query logs

- Best source of information about queries and related terms
  - short pieces of text and click data

- Compute association measure between all pairs of terms from the ***query logs***.

18

18

Query logs will usually give much better expansion terms.

- Yes
- No

In context vectors, postings actually represent "terms".

- Yes
- No

Today's Roadmap

- Thesaurus-based methods
- Query Logs
- Relevance Feedback

Relevance Feedback

- From user perspective: it may be difficult to formulate a good query when you don’t know the collection well, BUT easier to judge particular documents.

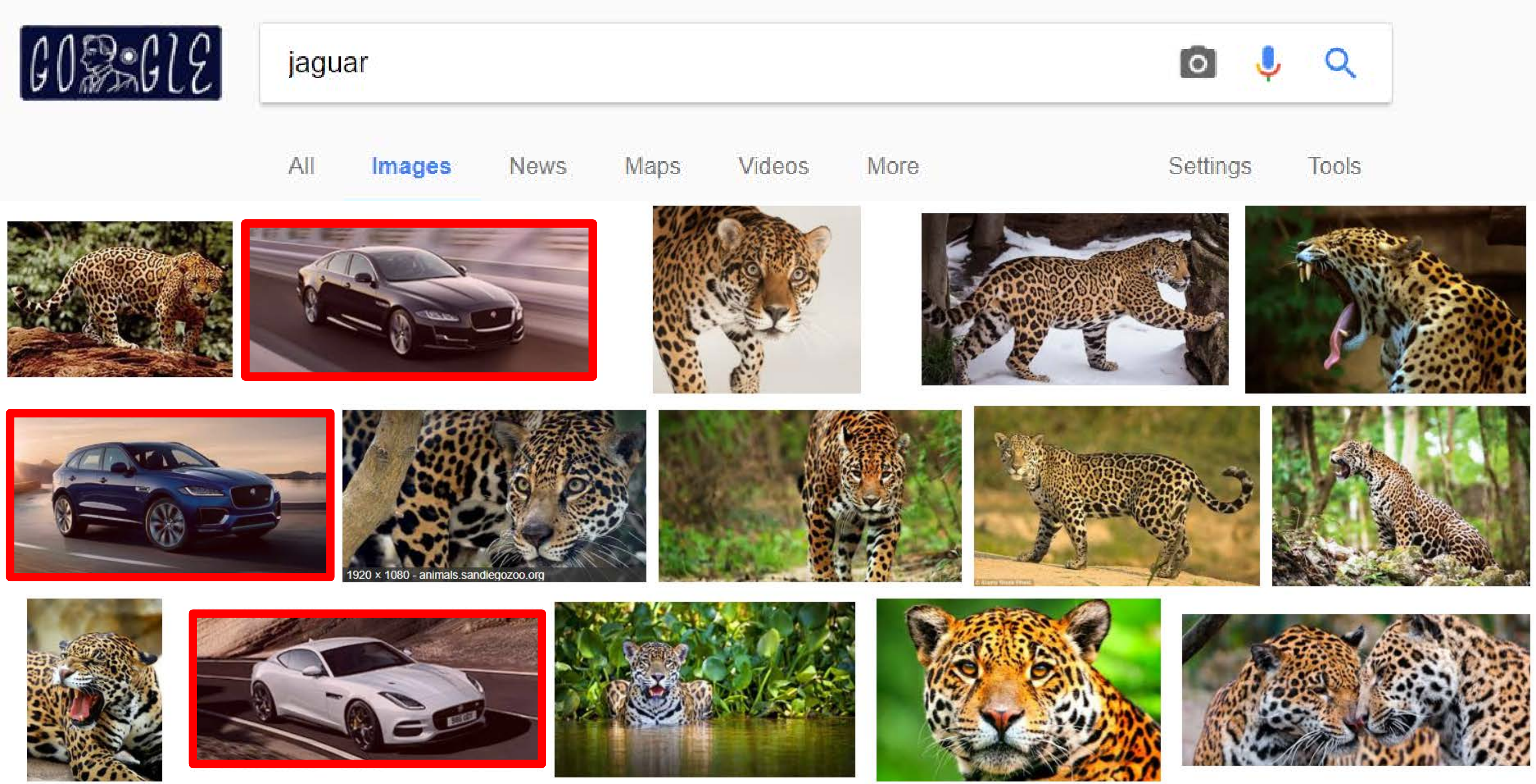
Idea: let user give feedback to the IR system about samples of what is relevant and what is not.

Relevance Feedback Process

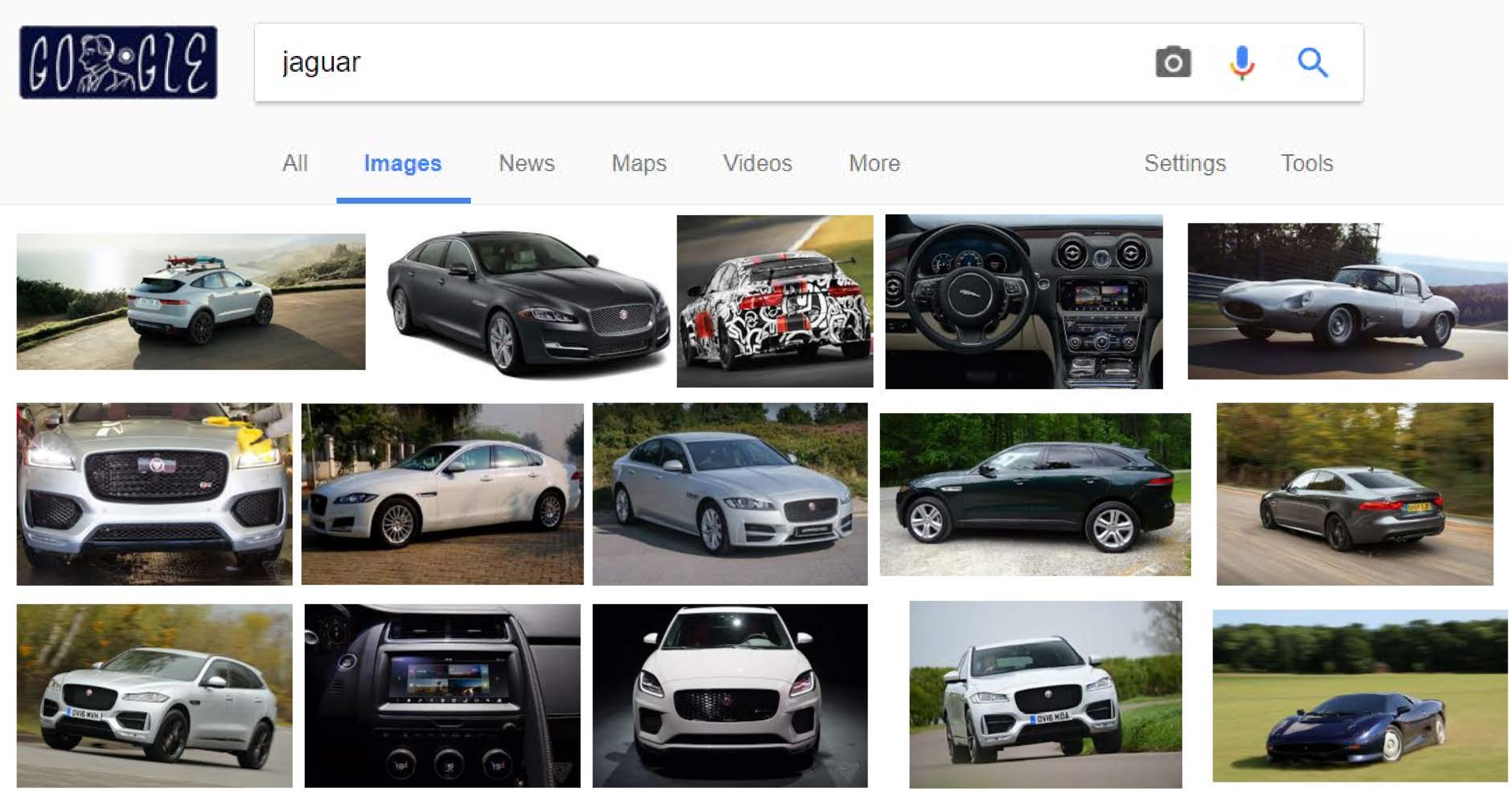
1. User issues a (short, simple) query
2. The system retrieves (initial) results and show to user
3. The user marks some results as relevant or non-relevant.
4. The system computes a *better representation of the information need* based on feedback.
5. The system retrieves new results based on the new representation and show to user.

- Relevance feedback can go through one or more iterations.

Example 1: Image Search



Example 1: Image Search



Example 2: Text Search

- Initial query: ***New space satellite applications***
- ***Initial Results***
  1. NASA Hasn’t Scrapped Imaging Spectrometer
  2. NASA Scratches Environment Gear From Satellite Plan
  3. Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
  4. A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
  5. Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
  6. Report Provides Support for the Critics Of Using Big Satellites to Study Climate
  7. Arianespace Receives Satellite Launch Pact From Telesat Canada
  8. Telecommunications Tale of Two Companies
- User then marks relevant documents with “+”.
- System learns new terms

Expanded Query after Rel. Feedback

2.074 *new*

30.816 *satellite*

5.991 **nasa**

4.196 **launch**

3.516 instrument

3.004 bundespost

2.790 rocket

2.003 broadcast

0.836 oil

15.106 *space*

5.660 *application*

5.196 **eos**

3.972 **aster**

3.446 arianespace

2.806 ss

2.053 scientist

1.172 earth

0.646 measure

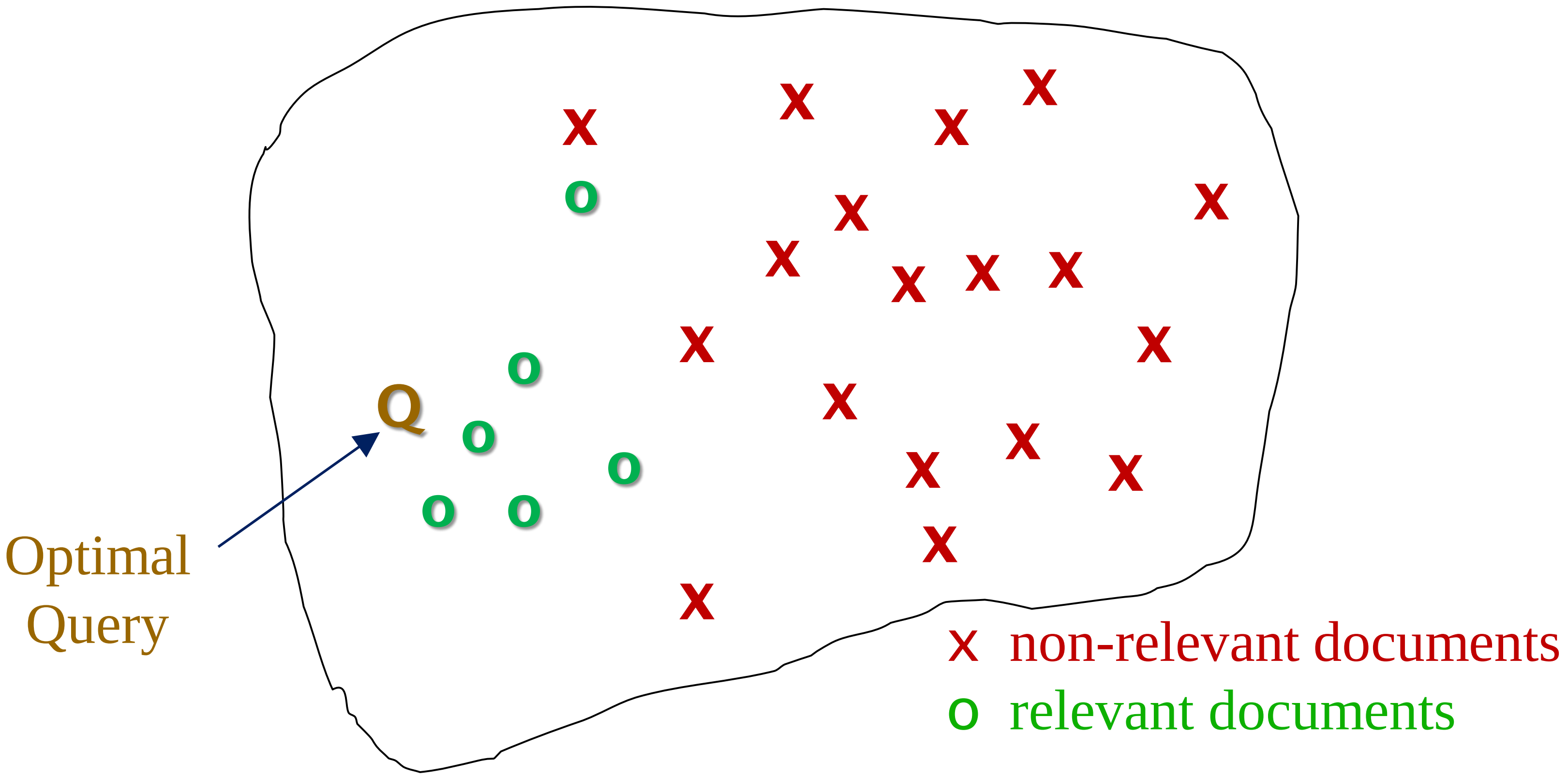
Results for Expanded Query

1. NASA Scratches Environment Gear From Satellite Plan
2. NASA Hasn't Scrapped Imaging Spectrometer
3. When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
4. NASA Uses 'Warm' Superconductors For Fast Circuit
5. Telecommunications Tale of Two Companies
6. Soviets May Adapt Parts of SS-20 Missile For Commercial Use
7. Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
8. Rescue of Satellite By Space Agency To Cost $90 Million

Hopefully better results!

ROCCHIO ALGORITHM IN VSM

Theoretical Optimal Query



Key Concept: Centroid

- Recall that, in VSM, we represent documents as points in a high-dimensional space
- The <u>centroid</u> is the center of mass of a set of points

$$\vec{\mu}(C) = \frac{1}{|C|} \sum_{\vec{d} \epsilon C} \vec{d}$$

where C is a set of documents.

Rocchio Algorithm: Theory

- Rocchio seeks the query $\overrightarrow{q_{opt}}$ that maximizes:

$$\vec{q}_{opt} = \underset{\vec{q}}{\operatorname{argmax}}[sim(\vec{q}, C_r) - sim(\vec{q}, C_{nr})]$$

- For cosine similarity:

$$\vec{q}_{opt} = \frac{1}{|C_r|}\sum_{\overrightarrow{d_j}\in C_r}\overrightarrow{d_j} - \frac{1}{|C_{nr}|}\sum_{\overrightarrow{d_j}\in C_{nr}}\overrightarrow{d_j}$$

$$\vec{q}_{opt} = \vec{\mu}(C_r) - \vec{\mu}(C_{nr})$$

Challenge: we don't know the truly relevant docs

Rocchio Algorithm: in Practice

- Only small set of docs are known to be REL or non-REL

$$\vec{q}_m = \alpha\, \vec{q}_0 + \beta \frac{1}{|Dr|} \sum_{\vec{d}_j \in Dr} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in Dnr} \vec{d}_j$$

$\vec{q}_0$ = original query vector

D_r = set of known relevant doc vectors

D_{nr} = set of known non-relevant doc vectors

$\vec{q}_m$ = modified query vector

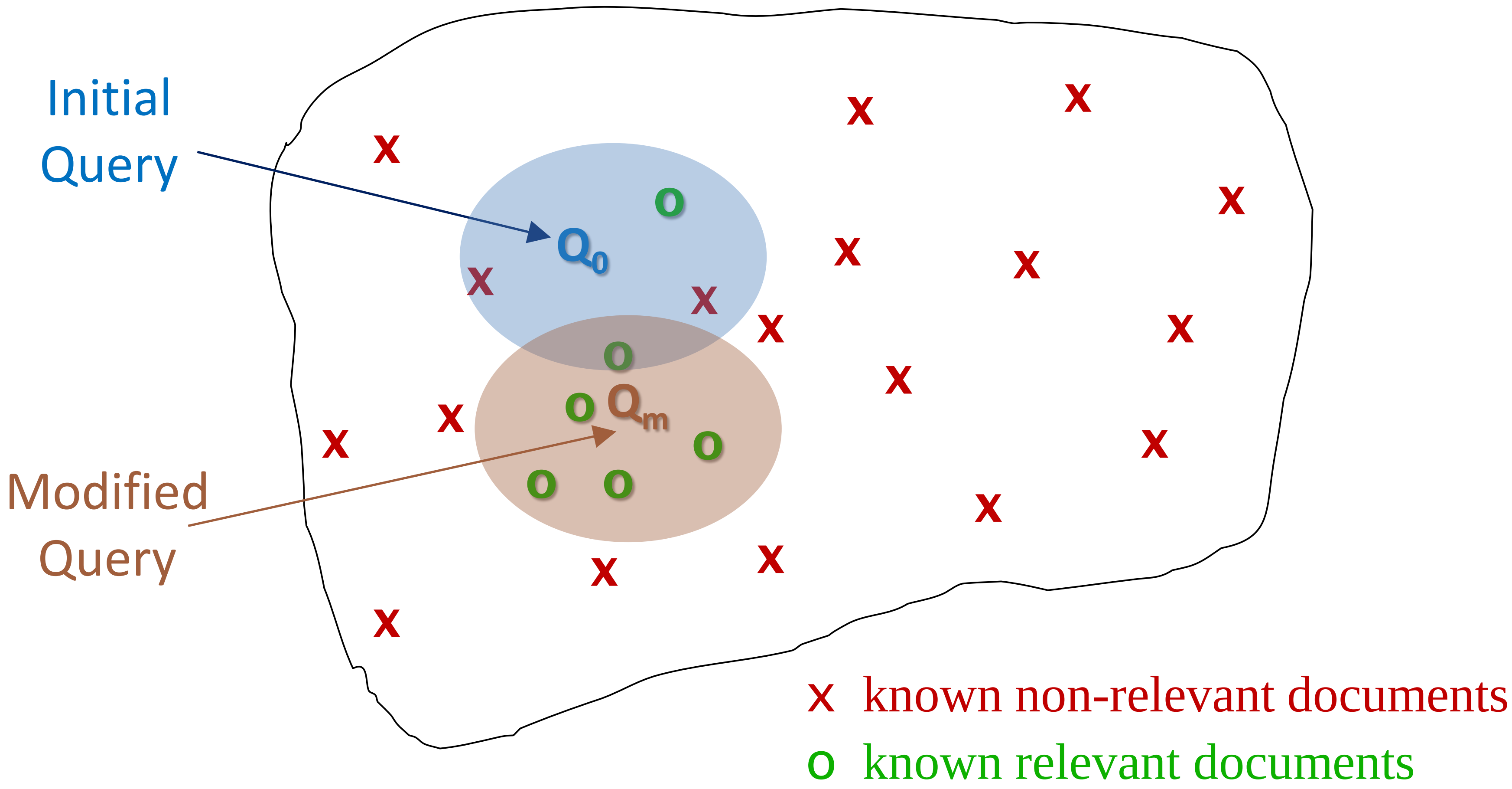
α = original query weight

β = positive feedback weight

γ = negative feedback weight

New query moves toward relevant documents and away from non-relevant documents

Effect of Relevance Feedback on Query



Notes about setting weights: α, β, γ

- Values of β, γ compared to α are set high when many judged documents are available.
- In practice, +ve feedback is more valuable than -ve feedback (usually, set β>γ)
  - many systems only allow positive feedback (γ=0).
  - Or, use only highest-ranked negative document.
- When γ>0, some weights in query vector can go -ve.
  - negative term weights are ignored (set to 0)

Effect of Relevance Feedback on Retrieval

- Relevance feedback can improve recall and precision
- In practice, relevance feedback is most useful for increasing *recall* in situations where recall is important.
- Empirically, one round of relevance feedback is often very useful. Two rounds is sometimes marginally useful.

Query Expansion?

Relevance Feedback: Problems

- Long queries are inefficient for typical IR engine.
  - High cost for retrieval system.
  - Long response times for user.
- Users are often reluctant to provide explicit feedback
- It's often harder to understand why a particular document was retrieved after applying relevance feedback.

Solution?

19

In relevance feedback, positive feedback is much better than negative feedback.

- Yes
- No

In Rocchio, there is no way to completely ignore the initial query.

- Yes
- No

20

Is there a way to apply relevance feedback <u>without</u> user's input?

Pseudo (or Blind) Relevance Feedback

- Automates the “manual” part of true relevance feedback.
- Algorithm:
  - Retrieve a ranked list of hits for the user’s query
  - Assume that the top k documents are relevant
  - Do relevance feedback (e.g., Rocchio)

Pseudo (or Blind) Relevance Feedback

- Automates the “manual” part of true relevance feedback.
- Algorithm:
  - Retrieve a ranked list of hits for the user’s query
  - Assume that the top k documents are relevant
  - Do relevance feedback (e.g., Rocchio)
    - Select top T terms based on term weights (e.g., tf-idf)
    - Add them to the query

PRF (BRF)

- Was proven to be useful for many IR applications
  - News search (learn names and entities)
  - Social media search (learn hashtags)
- But can go horribly wrong for some queries.

- Several iterations can cause ***query drift***.

- PRF is the most basic QE method for IR
  - Unsupervised
  - Language-independent
- Efficiency?

Co-occurrence with PRF

- **Pseudo-relevance feedback**
  - expansion terms based on term co-occurrence in top retrieved documents for initial query.

Implicit Feedback

- Less reliable than explicit
- But more useful than BRF

- Click on links
  - assumptions?
- Positive and negative?
- Other forms of implicit feedback?

19

20

PRF is more efficient than standard retrieval.

- ➢ Yes
- ➢ No

Eye tracking can be used for getting implicit feedback.

- ➢ Yes
- ➢ No

Can we represent terms by "meaning"?