CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computer Science

http://www.cit.ie

Natural Language Processing

Week12: Information Retrieval

Dr. Haithem Afli

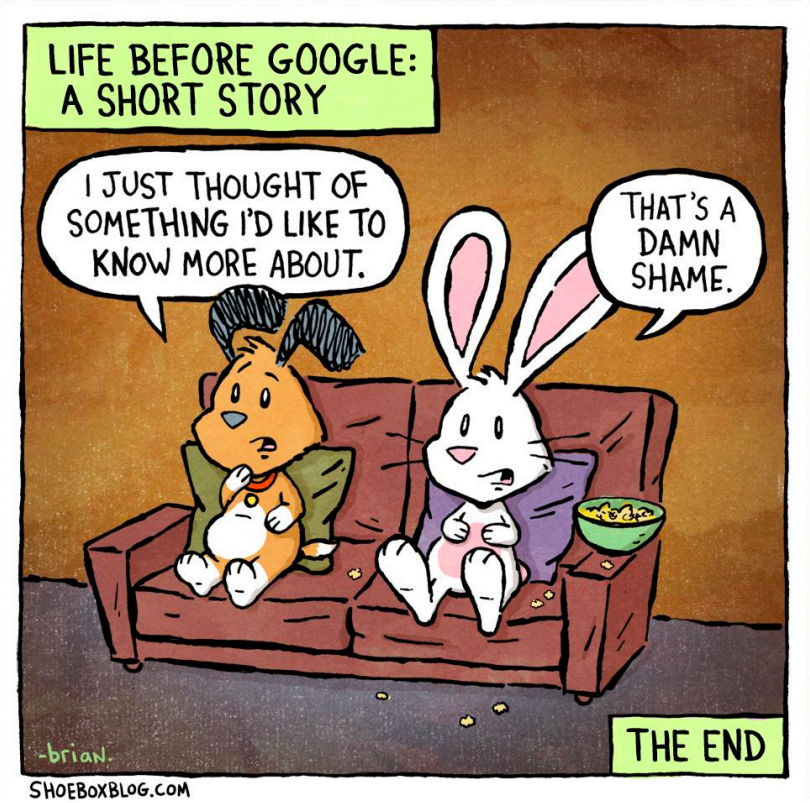
Haithem.afli@cit.ie

@AfliHaithem

2020/2021

What is information retrieval

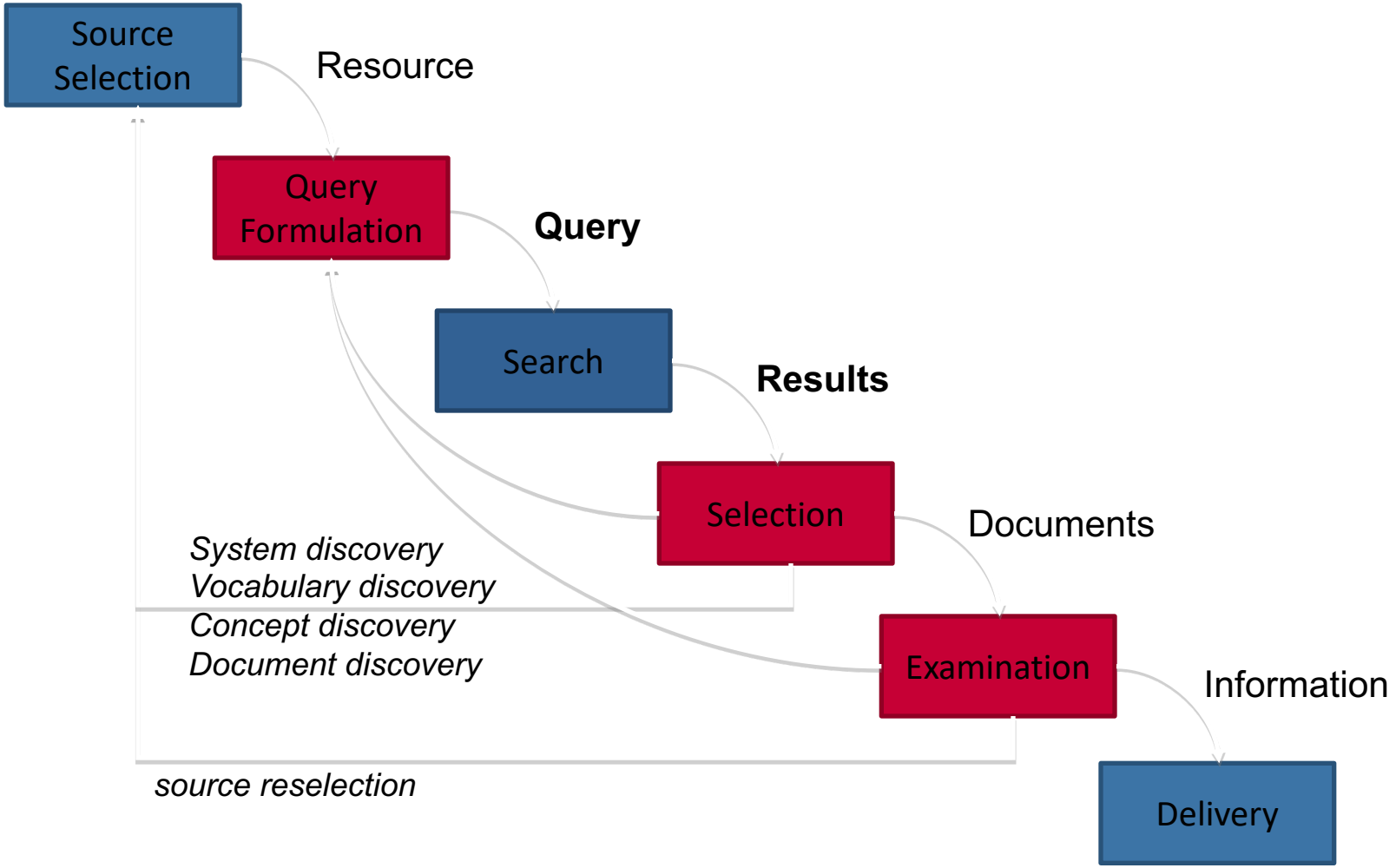
- Obtaining documents pertinent to a natural language query
  - What are documents?
  - What is pertinent?



First, nomenclature…

- Information retrieval (IR)
  - Focus on textual information (= text/document retrieval)
  - Other possibilities include image, video, music, …
- What do we search?
  - Generically, “collections”
  - Less-frequently used, “corpora”
- What do we find?
  - Generically, “documents”
  - Even though we may be referring to web pages, PDFs, PowerPoint slides, paragraphs, etc.

Information Retrieval Cycle



The Central Problem in Search

Searcher

Author

Concepts

Concepts

Query Terms

"tragic love story"

Document Terms

"fateful star-crossed romance"

Do these represent the same concepts?

Nature of information is changing

- Big search engines are good generalists
  - *but our retrieval expectations are evolving*

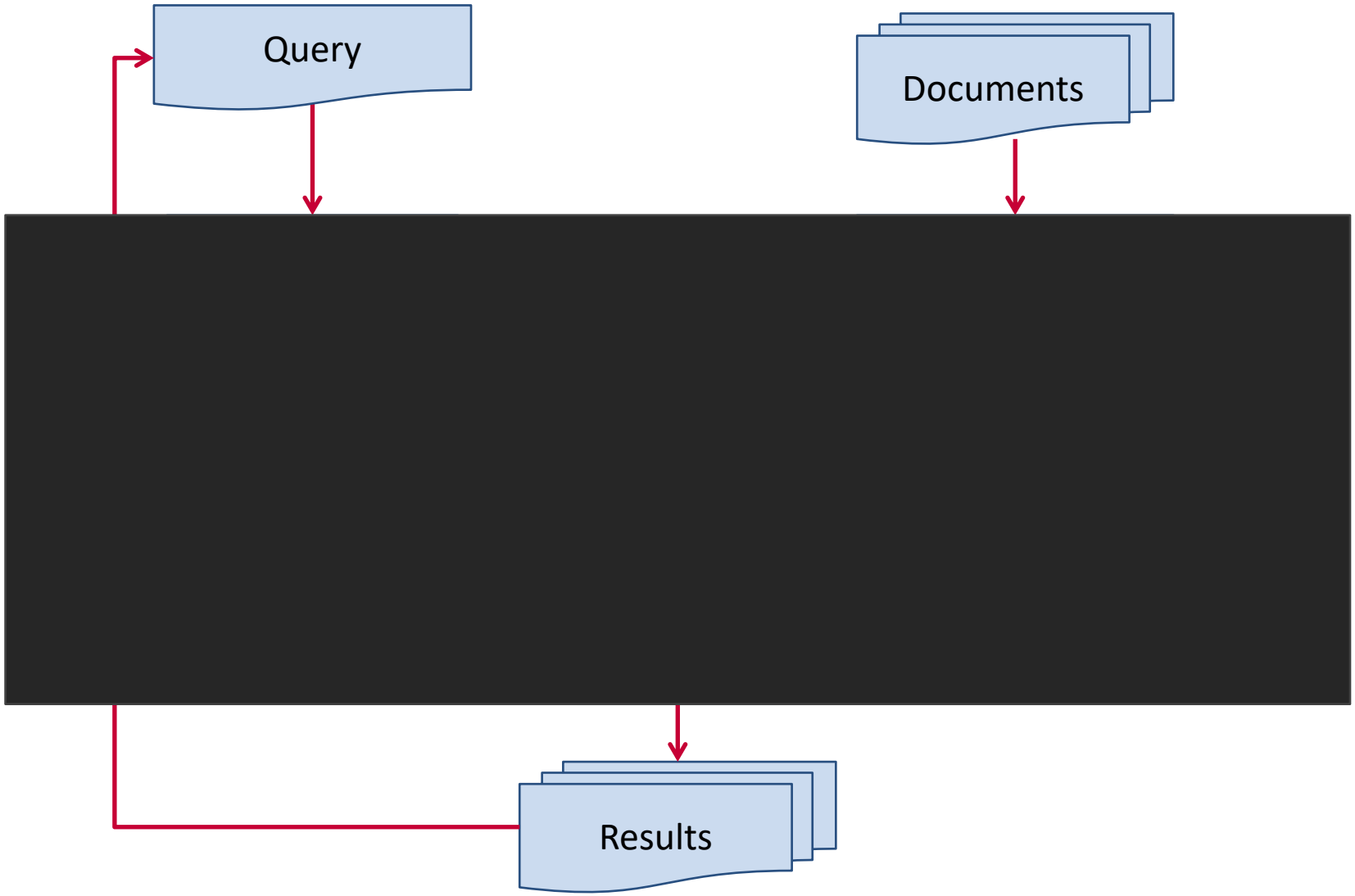
Partially as a result of retrieval!

	Documents are changing	Queries are changing	Consumption is changing
Then	• Print media • Plain text	• Single words • Simple expressions • Traditional keyboards	• Microfiche • Monitor
Now	• Structured text • Images • Audio • Video	• Thoughts • Images • Audio • Swipe • Digital languages	• Voice • Mobile • Wrist

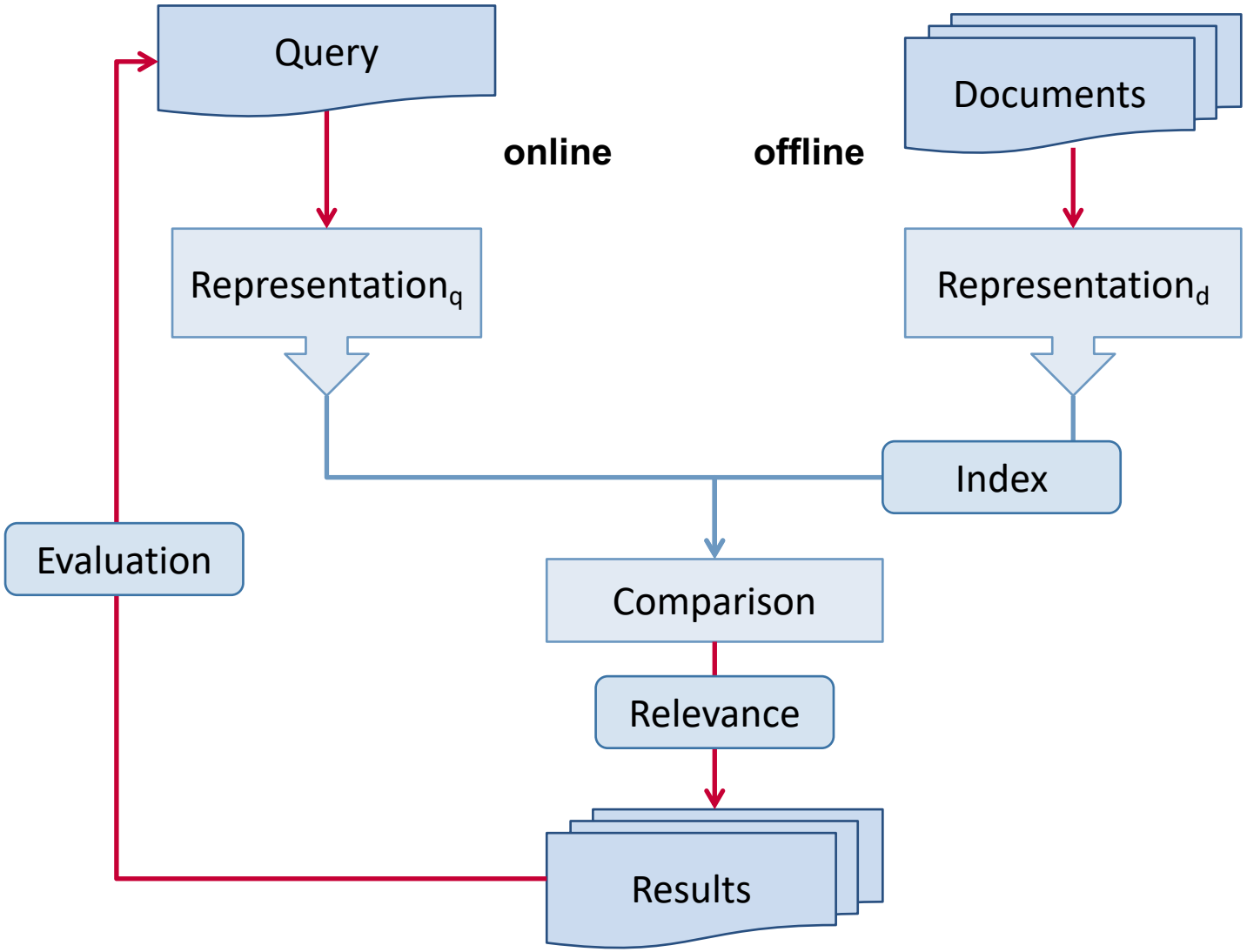
Search engines do this really well

Search engines are *trying* to do this really well

The IR black box



The IR black box



How do we represent text?

- "Bag of words"
  - Treat all the words in a document as index terms
  - Assign a "weight" to each term based on "importance" (or, in simplest case, presence/absence of word)
  - Disregard order, structure, meaning, etc. of the words
  - Simple, yet effective!
- Assumptions
  - Term occurrence is independent
  - Document relevance is independent
  - "Words" are well-defined

What's a word?

天主教教宗若望保祿二世因感冒再度住進醫院。
這是他今年第二度因同樣的病因住院。

وقال مارك ريجيف - الناطق باسم
الخارجية الإسرائيلية - إن شارون قبل
الدعوة وسيقوم للمرة الأولى بزيارة
تونس، التي كانت لفترة طويلة المقر
الرسمي لمنظمة التحرير الفلسطينية بعد خروجها من لبنان عام 1982.

Выступая в Мещанском суде Москвы экс-глава ЮКОСа заявил не совершал ничего противозаконного, в чем обвиняет его генпрокуратура России.

भारत सरकार ने आर्थिक सर्वेक्षण में वित्तीय वर्ष 2005-06 में सात फ़ीसदी विकास दर हासिल करने का आकलन किया है और कर सुधार पर ज़ोर दिया है

日米連合で台頭中国に対処…アーミテージ前副長官提言

조재영 기자= 서울시는 25일 이명박 시장이 `행정중심복합도시" 건설안에 대해 `군대라도 동원해 막고싶은 심정"이라고 말했다는 일부 언론의 보도를 부인했다.

Sample Document

McDonald's slims down spuds

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

…

"Bag of Words"

14 × McDonalds

12 × fat

11 × fries

8 × new

7 × french

6 × company, said, nutrition

5 × food, oil, percent, reduce, taste, Tuesday

…

Information retrieval models

- An IR model governs how a document and a query are represented and how the relevance of a document to a user query is defined.
- Main models:
  - **Boolean model**
  - Vector space model
  - Statistical language model
  - etc

Boolean model

- Each document or query is treated as a **"bag" of words** or **terms**. Word sequence is not considered.
- Given a collection of documents D, let $V = \{t_1, t_2, ..., t_{|V|}\}$ be the set of distinctive words/terms in the collection. V is called the vocabulary.
- A weight $w_{ij} > 0$ is associated with each term t_i of a document $\mathbf{d}_j \in D$. For a term that does not appear in document $\mathbf{d}_j$, $w_{ij} = 0$.

$$\mathbf{d}_j = (w_{1j}, w_{2j}, ..., w_{|V|j}),$$

Boolean model (contd)

- Query terms are combined logically using the Boolean operators **AND**, **OR**, and **NOT.**
  - E.g., ((*data* AND *mining*) AND (NOT *text*))
- Retrieval
  - Given a Boolean query, the system retrieves every document that makes the query logically true.
  - Called **exact match**.
- The retrieval results are usually quite poor because term frequency is not considered.

Boolean queries: Exact match

- The Boolean retrieval model is being able to ask a query that is a Boolean expression:
  - Boolean Queries are queries using *AND, OR* and *NOT* to join query terms
    - Views each document as a set of words
    - Is precise: document matches condition or not.
  - Perhaps the simplest model to build an IR system on
- Primary commercial retrieval tool for 3 decades.
- Many search systems you still use are Boolean:
  - Email, library catalog, Mac OS X Spotlight

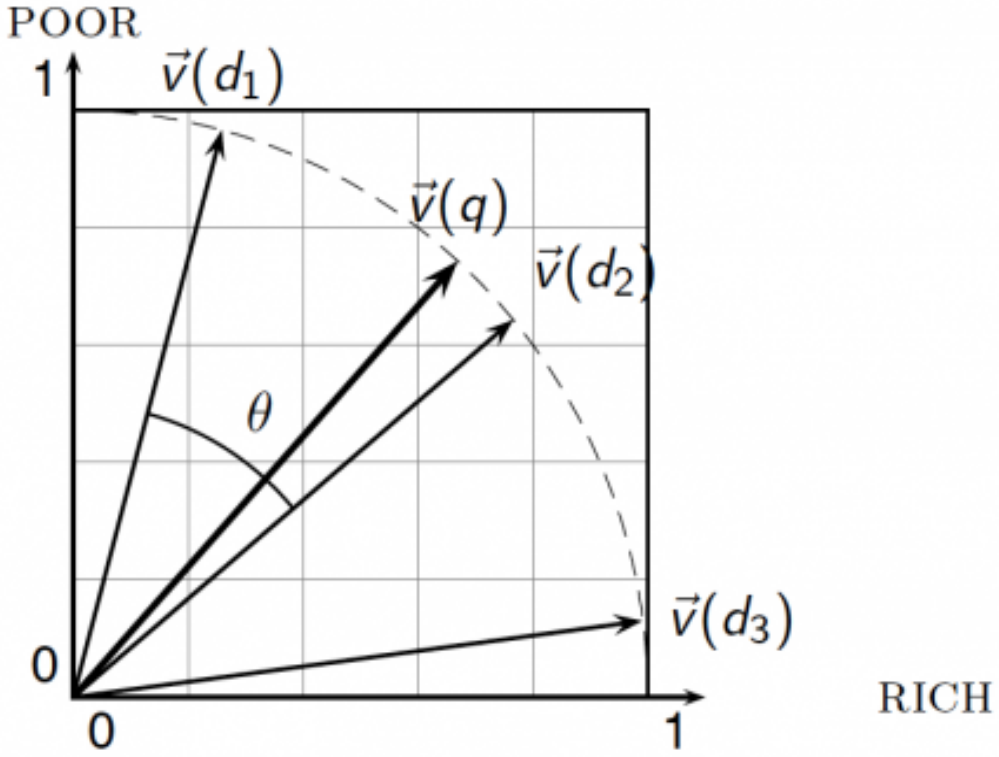
Strengths and Weaknesses

- Strengths
  - Precise, if you know the right strategies
  - Precise, if you have an idea of what you're looking for
  - Implementations are fast and efficient
- Weaknesses
  - Users must learn Boolean logic
  - Boolean logic insufficient to capture the richness of language
  - No control over size of result set: either too many hits or none
  - When do you stop reading? All documents in the result set are considered "equally good"
  - What about partial matches? Documents that "don't quite match" the query may be useful also

Information retrieval models

- An IR model governs how a document and a query are represented and how the relevance of a document to a user query is defined.
- Main models:
  - Boolean model
  - **Vector space model**
  - Statistical language model
  - etc

Vector Space Model



Assumption: Documents that are "close together" in vector space "talk about" the same things

Therefore, retrieve documents based on how close the document is to the query (i.e., similarity ~ "closeness")

Similarity Metric

- Use "angle" between the vectors:

$$\cos(\theta) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|}$$

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

- Or, more generally, inner products:

$$sim(d_j, d_k) = \vec{d}_j \cdot \vec{d}_k = \sum_{i=1}^{n} w_{i,j} w_{i,k}$$

Vector space model

- Documents are also treated as a “bag” of words or terms.
- Each document is represented as a vector.
- However, the term weights are no longer 0 or 1. Each term weight is computed based on some variations of TF or TF-IDF scheme.

Term Weighting

- Term weights consist of two components
  - Local: how important is the term in this document?
  - Global: how important is the term in the collection?
- Here's the intuition:
  - Terms that appear often in a document should get high weights
  - Terms that appear in many documents should get low weights
- How do we capture this mathematically?
  - Term frequency (local)
  - Inverse document frequency (global)

TF.IDF Term Weighting

$$w_{i,j} = \mathrm{tf}_{i,j} \cdot \log \frac{N}{n_i}$$

$w_{i,j}$ **weight assigned to term *i* in document *j***

$\mathrm{tf}_{i,j}$ **number of occurrence of term *i* in document *j***

N **number of documents in entire collection**

n_i **number of documents with term *i***

TF-IDF weighting

$$w_{t,d} = tf_{t,d} \times \log \frac{N}{df_t}$$

"Weight" of term *t* within document *d*

How often term *t* appears in document *d*

Number of documents in which term *t* appears, normalized, inverted, and scaled

Retrieval in vector space model

- Query **q** is represented in the same way or slightly differently.
- Relevance of $\mathbf{d}_j$ to **q**: Compare the similarity of query **q** and document $\mathbf{d}_j$.
- Cosine similarity (the cosine of the angle between the two vectors)

$$cosine(\mathbf{d}_j, \mathbf{q}) = \frac{\langle \mathbf{d}_j \bullet \mathbf{q} \rangle}{\| \mathbf{d}_j \| \times \| \mathbf{q} \|} = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}}$$

- Cosine is also commonly used in text clustering

Cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

- Cosine similarity is a measure of similarity between two vectors
- Ranges
  - Positive 1: A and B are exactly the same
  - Negative 1: A and B are exactly opposite
- In IR, the range is 0 to 1 since TF-IDF is always positive

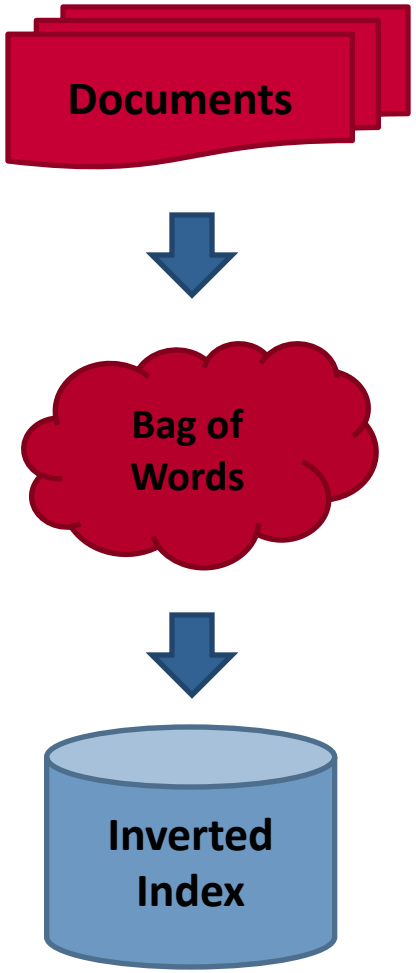
An Example

- A document space is defined by three terms:
  - hardware, software, users
  - the vocabulary
- A set of documents are defined as:
  - A1=(1, 0, 0), A2=(0, 1, 0), A3=(0, 0, 1)
  - A4=(1, 1, 0), A5=(1, 0, 1), A6=(0, 1, 1)
  - A7=(1, 1, 1) A8=(1, 0, 1). A9=(0, 1, 1)
- If the Query is “hardware and software”
- what documents should be retrieved?

An Example (cont.)

- In Boolean query matching:
  - document A4, A7 will be retrieved ("AND")
  - retrieved: A1, A2, A4, A5, A6, A7, A8, A9 ("OR")
- In similarity matching (cosine):
  - q=(1, 1, 0)
  - S(q, A1)=0.71, S(q, A2)=0.71, S(q, A3)=0
  - S(q, A4)=1, S(q, A5)=0.5, S(q, A6)=0.5
  - S(q, A7)=0.82, S(q, A8)=0.5, S(q, A9)=0.5
  - Document retrieved set (with ranking)=
    - {A4, A7, A1, A2, A5, A6, A8, A9}

Constructing Inverted Index (Word Counting)



case folding, tokenization, stopword removal, stemming

~~syntax~~, ~~semantics~~, word ~~knowledge~~, etc.

Stopwords removal

- Many of the most frequently used words in English are useless in IR and text mining – these words are called *stop words*.
  - the, of, and, to, ….
  - Typically about 400 to 500 such words
  - For an application, an additional domain specific stopwords list may be constructed
- Why do we need to remove stopwords?
  - Reduce indexing (or data) file size
    - stopwords accounts 20-30% of total word counts.
  - Improve efficiency and effectiveness
    - stopwords are not useful for searching or text mining
    - they may also confuse the retrieval system.

Stemming

- Techniques used to find out the root/stem of a word. E.g.,
  - user
  - users
  - used
  - using

  engineering
  engineered
  engineer
- stem: use engineer

Usefulness:

- improving effectiveness of IR and text mining
  - matching similar words
  - Mainly improve recall
- reducing indexing size
  - combing words with same roots may reduce indexing size as much as 40-50%.

Basic stemming methods

Using a set of rules. E.g.,

- remove ending
  - if a word ends with a consonant other than s,

    followed by an s, then delete s.
  - if a word ends in es, drop the s.
  - if a word ends in ing, delete the ing unless the remaining word consists only of one letter or of th.
  - If a word ends with ed, preceded by a consonant, delete the ed unless this leaves only a single letter.
  - ……
- transform words
  - if a word ends with “ies” but not “eies” or “aies” then “ies --> y.”

Discussion

And now… the results!

- We now have
  - A way of indexing documents
  - A structure for treating queries like documents
  - A scoring function for comparing the two
- How do we know if what's returned is any good?
- Can we alter the query to do better?

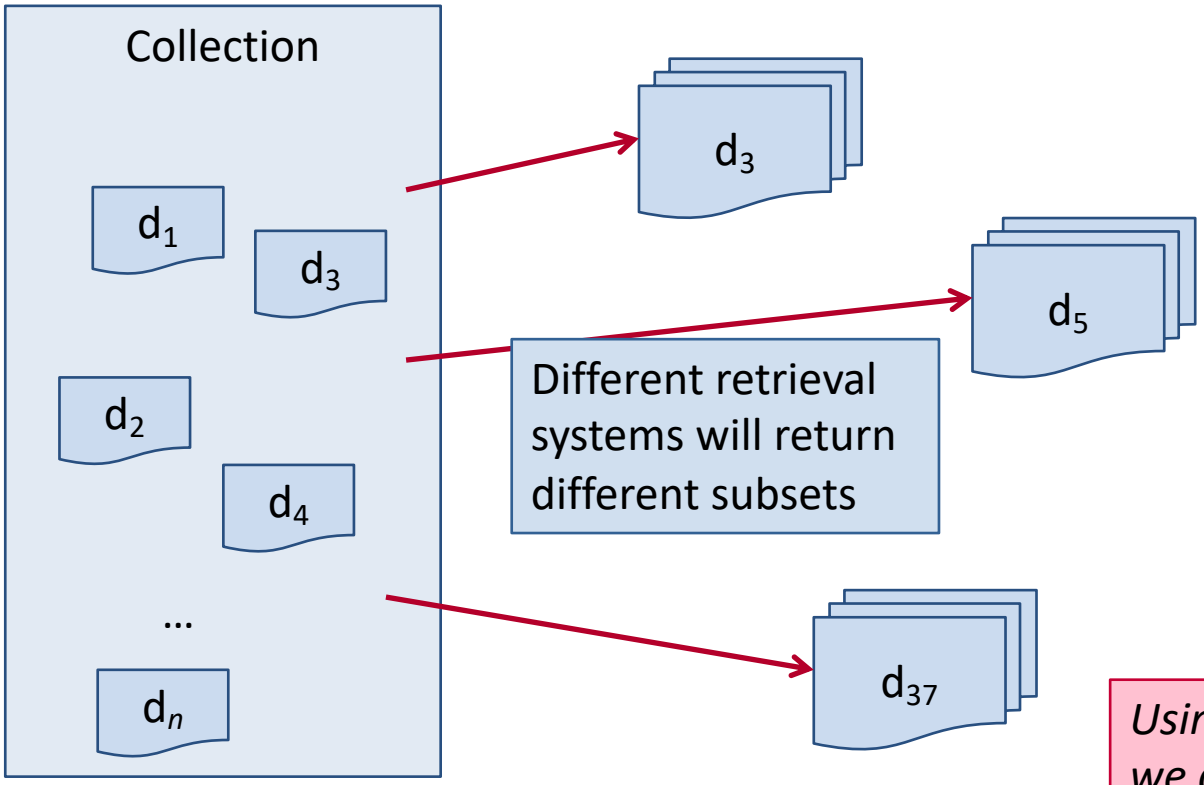
Relevance

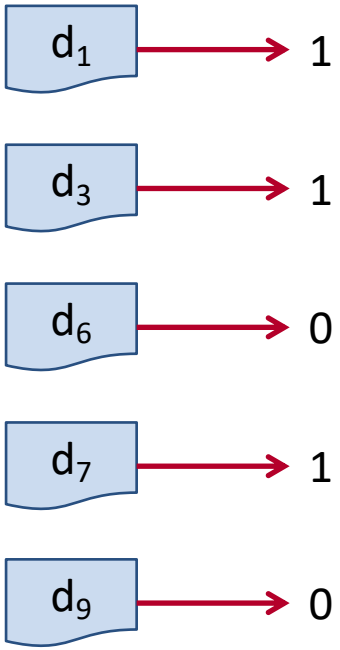
- TF-IDF is nice, but true relevance is subjective
- We can get closer by asking humans to judge
  - Mark documents within a collection as relevant or not relevant to a given *topic* (not a query!)
- New systems then have a way of estimating whether a query was effective at satisfying an information need

Judging relevance

Topic: Information on how to apply natural language processing on real-life applications

Human judges determine which documents are relevant to the topic





Using these relevance judgments, we can now evaluate new queries that are inline with the topic

Precision and recall

Precision

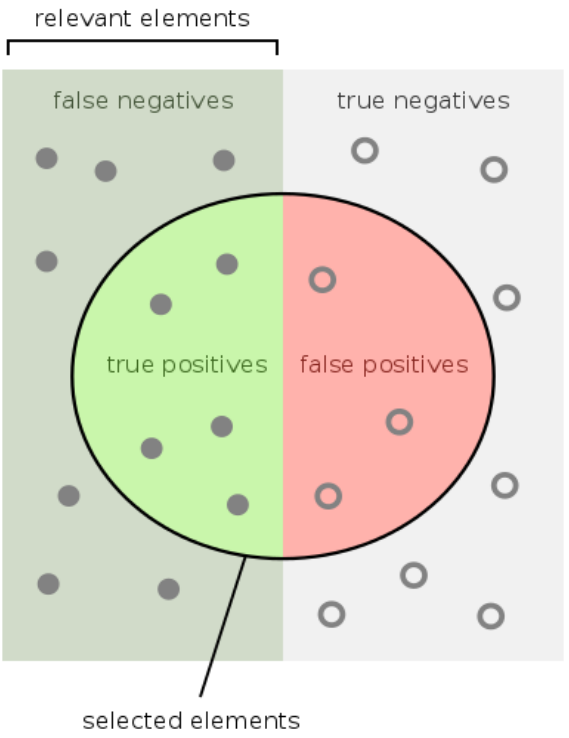
- The fraction of retrieved documents that are relevant
- How many selected documents are relevant?

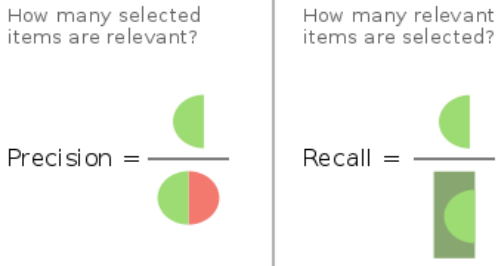
Recall

- Fraction of relevant instances that are retrieved
- How many relevant documents are selected?

Foundation

- Precision/recall are actively used to describe system performance
- Form the basis of many IR evaluation metrics
  - "This is a recall-based metric"
- Ranked retrieval isn't necessarily required

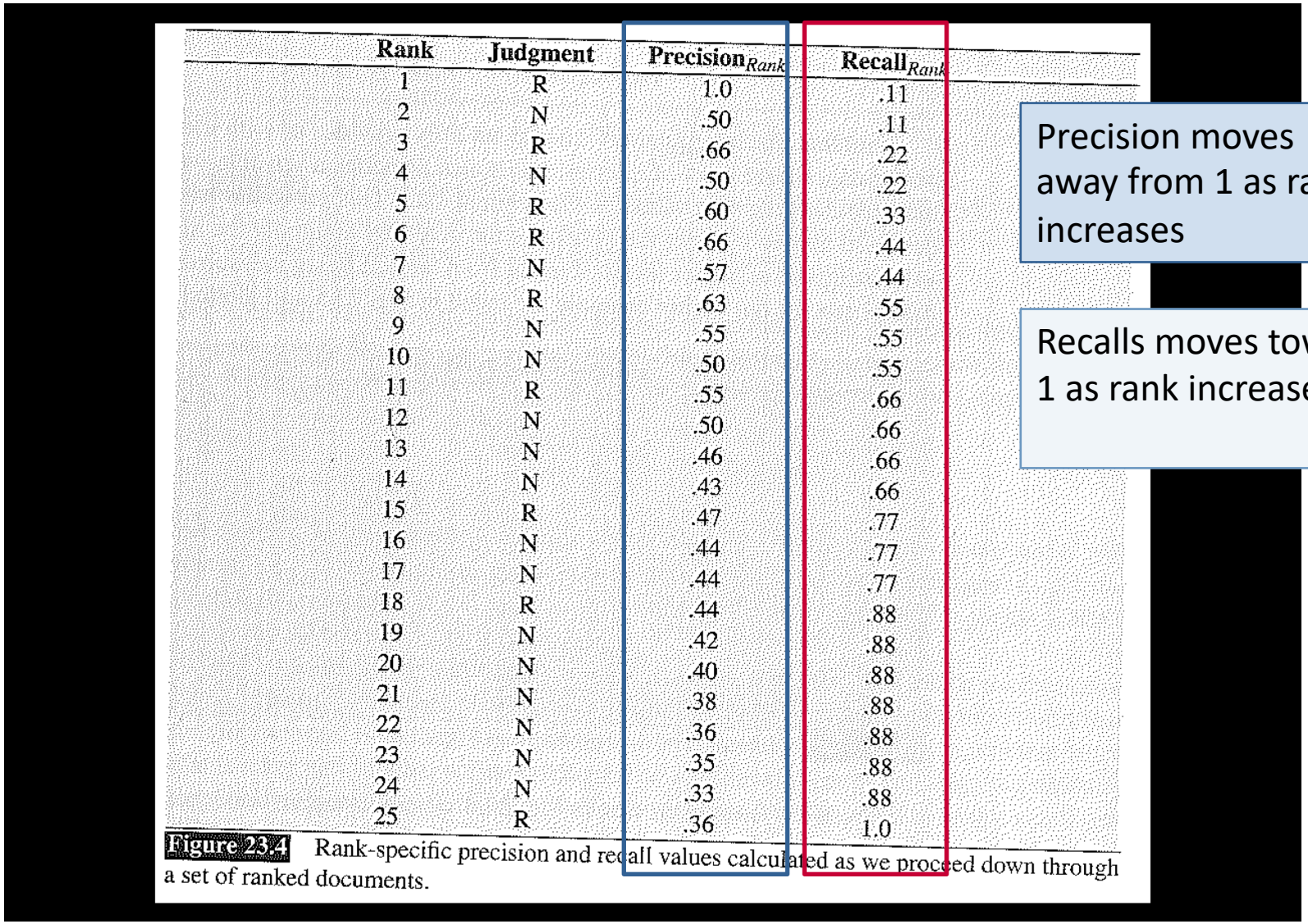




Cutoffs

- It is common to quote precision/recall metrics at certain *cutoffs*
- Rather than the entire set of retrieved documents, just consider a top-subset
  - Precision at *n*: score assuming *n* was the total number of documents retrieved
- Cutoffs make the metric more meaningful
  - Most users don't consider all retrieved documents
  - At some point, recall will always be 1

Nature of the beast



Rank	Judgment	Precision_{Rank}	Recall_{Rank}
1	R	1.0	.11
2	N	.50	.11
3	R	.66	.22
4	N	.50	.22
5	R	.60	.33
6	R	.66	.44
7	N	.57	.44
8	R	.63	.55
9	N	.55	.55
10	N	.50	.55
11	R	.55	.66
12	N	.50	.66
13	N	.46	.66
14	N	.43	.66
15	R	.47	.77
16	N	.44	.77
17	N	.44	.77
18	R	.44	.88
19	N	.42	.88
20	N	.40	.88
21	N	.38	.88
22	N	.36	.88
23	N	.35	.88
24	N	.33	.88
25	R	.36	1.0

Figure 23.4 Rank-specific precision and recall values calculated as we proceed down through a set of ranked documents.

Precision moves away from 1 as rank increases

Recalls moves toward 1 as rank increases

Improving the query

- Can the system be “smart” about how it issues a query?
  - Does a user actually know what they want?
  - Does a user know how to express what they want?
  - Perhaps they need some help ☺
- Known as *query reformulation*
  1. Language-based
  2. Relevance-based

Language-based reformulation

- Common reformulation techniques used during indexing
  - Normalization, stemming, and stop word removal
- Identify and *deal with* spelling errors
- Expand the users query to include related terms
  - Thesaurus
  - Co-occurrence evaluation
- *How does reformulation effect recall and precision?*

Relevance-based reformulation

- A user might feel differently about the relevance of their results
- Use their feedback to determine which documents to return
- “Rocchio Algorithm”

$$\overrightarrow{Q_m} = (a \cdot \overrightarrow{Q_o}) + \underbrace{\left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right)}_{\text{Add a bit of the relevant documents to the new query}} - \underbrace{\left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right)}_{\text{Subtract a bit of the non-relevant documents from the new query}}$$

Example

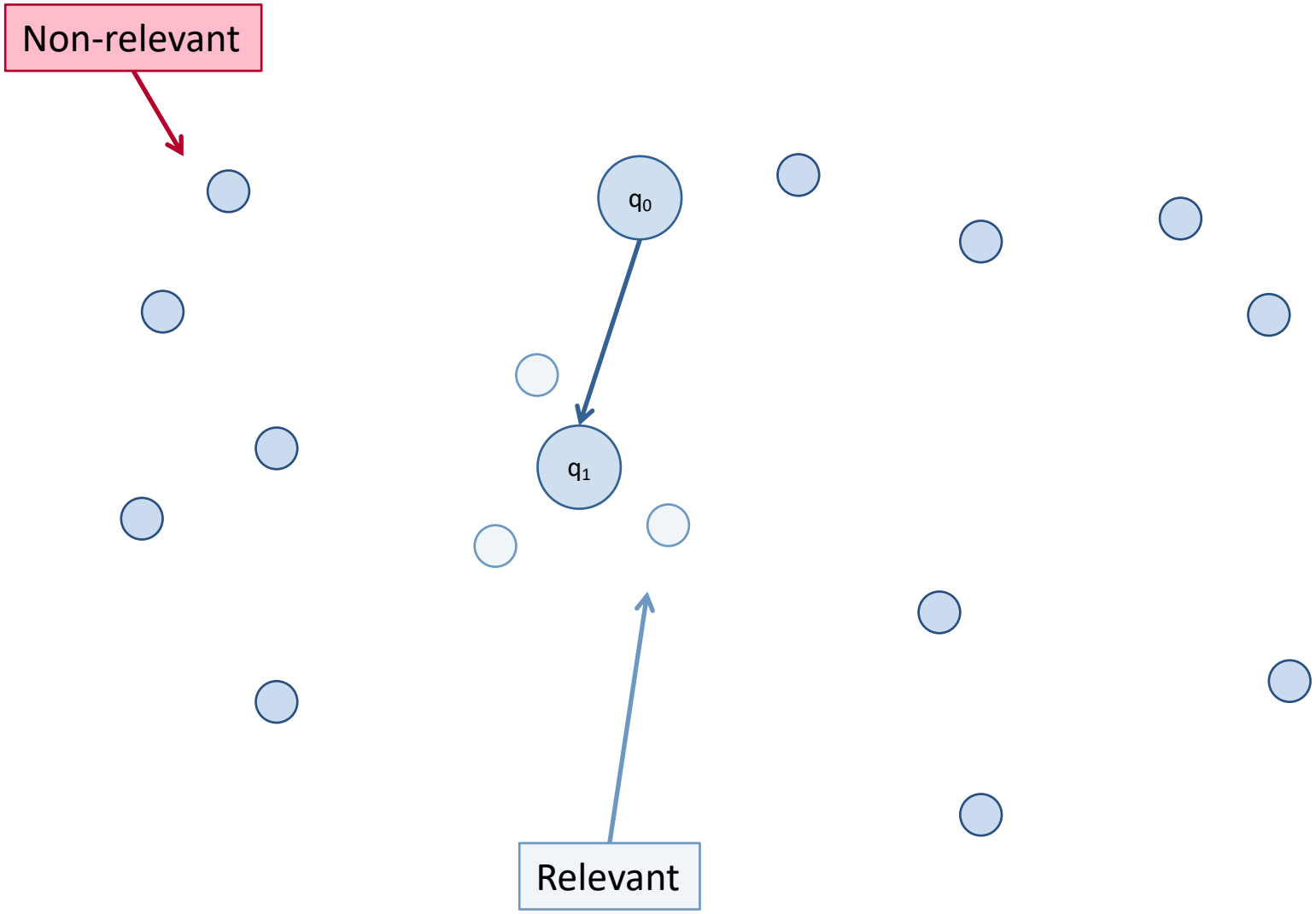
- Words in the corpus:

  { run, lion, cat, dog, program }

Original	q_0 = [1, 0, 1, 0, 0]	[run, cat]
Relevant	d_r = [2, 2, 1, 0, 0]	[run, run, lion, lion, cat]
Non-relevant	d_n = [2, 0, 1, 0, 3]	[run, run, cat, program, program, program]

$$
\begin{aligned}
q_1 &= q_0 + 1.0 d_r - 0.5 d_n \\
&= [1,0,1,0,0] + 1.0 \times [2,2,1,0,0] - 0.5 \times [2,0,1,0,3] \\
&= [2,2,1.5,0,-1.5] \\
&= [2 \times run, 2 \times lion, 1.5 \times cat]
\end{aligned}
$$

Graphical view

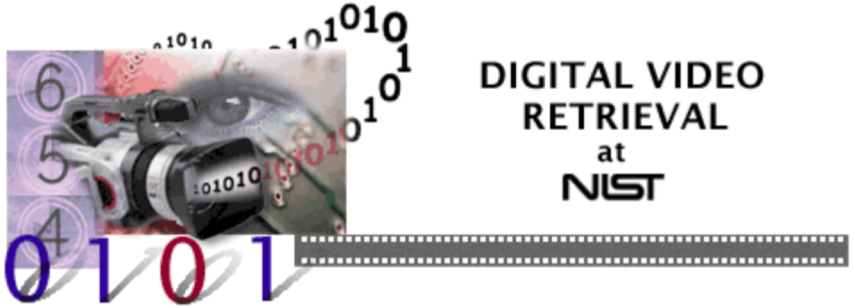


Get your hands dirty!

- The field of information retrieval is very accessible
- Lots of opportunities to try your indexing/query/evaluation ideas on various data sets
  - Compare results to how others have done

(Two) Forums for evaluation

- Text Retrieval Conference (TREC)
  - "support and encourage research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval"



TREC Video Retrieval Evaluation: TRECVID

- MediaEval
  - "benchmarking initiative dedicated to evaluating new algorithms for multimedia access and retrieval"

There are several others!

References

Book references
Introduction to information retrieval, Manning, Raghavan and Schütze
Speech and language processing, Jurafsky and Martin

Web references
Information retrieval (INST 734), UMD
Introduction to information retrieval (CS160), Pomona College

Discussion

CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computer Science

http://www.cit.ie

Thank you

Haithem.afli@cit.ie

@AfliHaithem