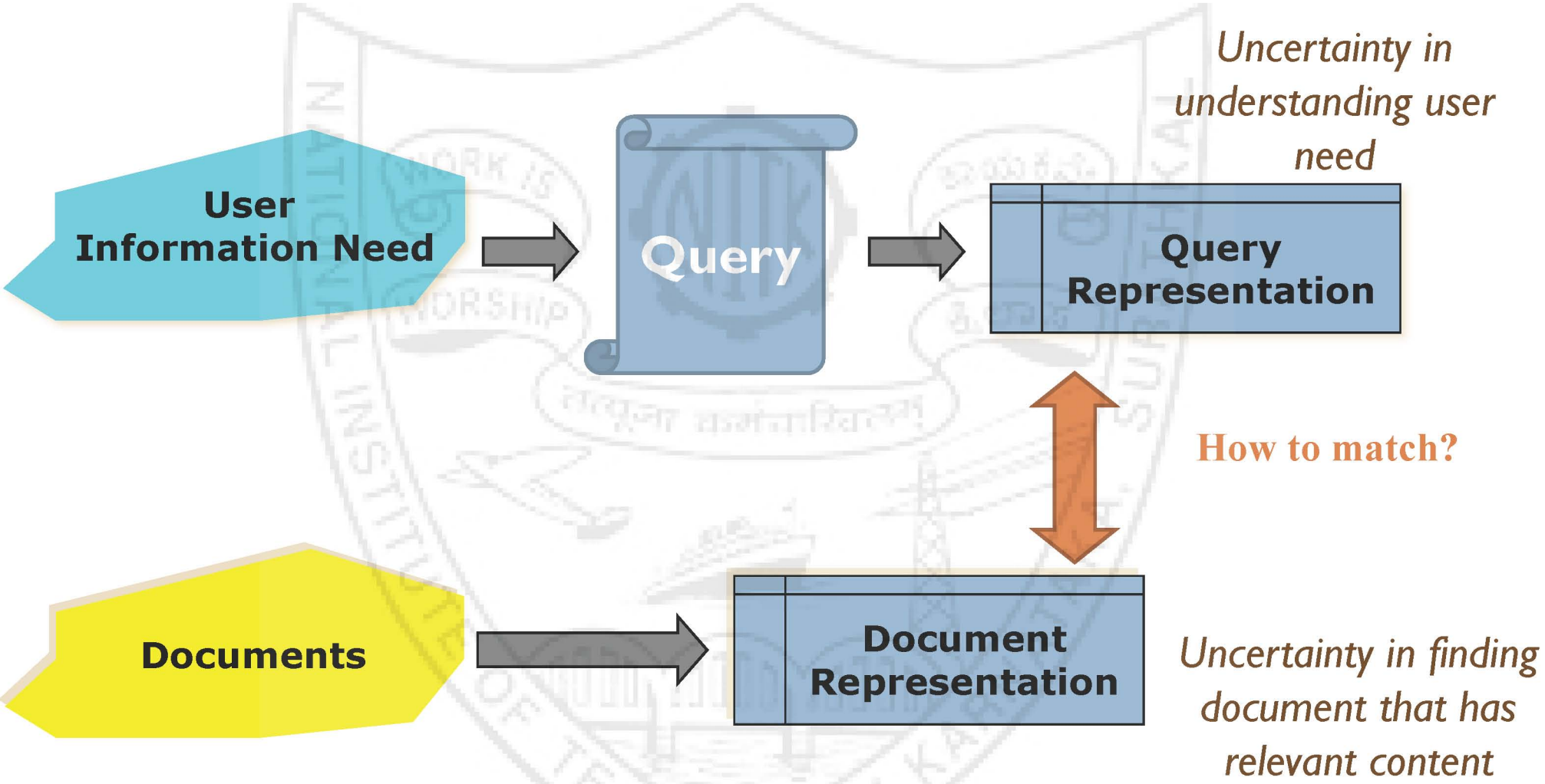
Jul – Nov 2022
IT458

IR Models for Unstructured Text

Probabilistic IR Models

Why probabilities in IR?



Solution: Use probability to quantify these uncertainties.

Probability of Relevance

- Three random variables
  - Query Q
  - Document D
  - Relevance R $\in$ {0,1}

- **Idea:** find document's probability of relevance w.r.t information need
  - $P(R=1 \mid Q, D_i)$ = ???

- **Goal**: to find rank of a D for given Q

Probabilistic IR - varieties

- Classical probabilistic retrieval models
  - Binary independence models (BIM)
    - When some relevance is known
    - When relevance is not known
  - Best Match Models
    - BM11, BM15
    - (Okapi) BM25

Probabilistic IR Models

Binary Independence Model (BIM)

Binary Independence Model

Basic concept:

- "For a given query, if we know some documents that are relevant, then terms that occur in those documents should be given greater weighting while searching for other relevant documents.

-Van Rijsbergen

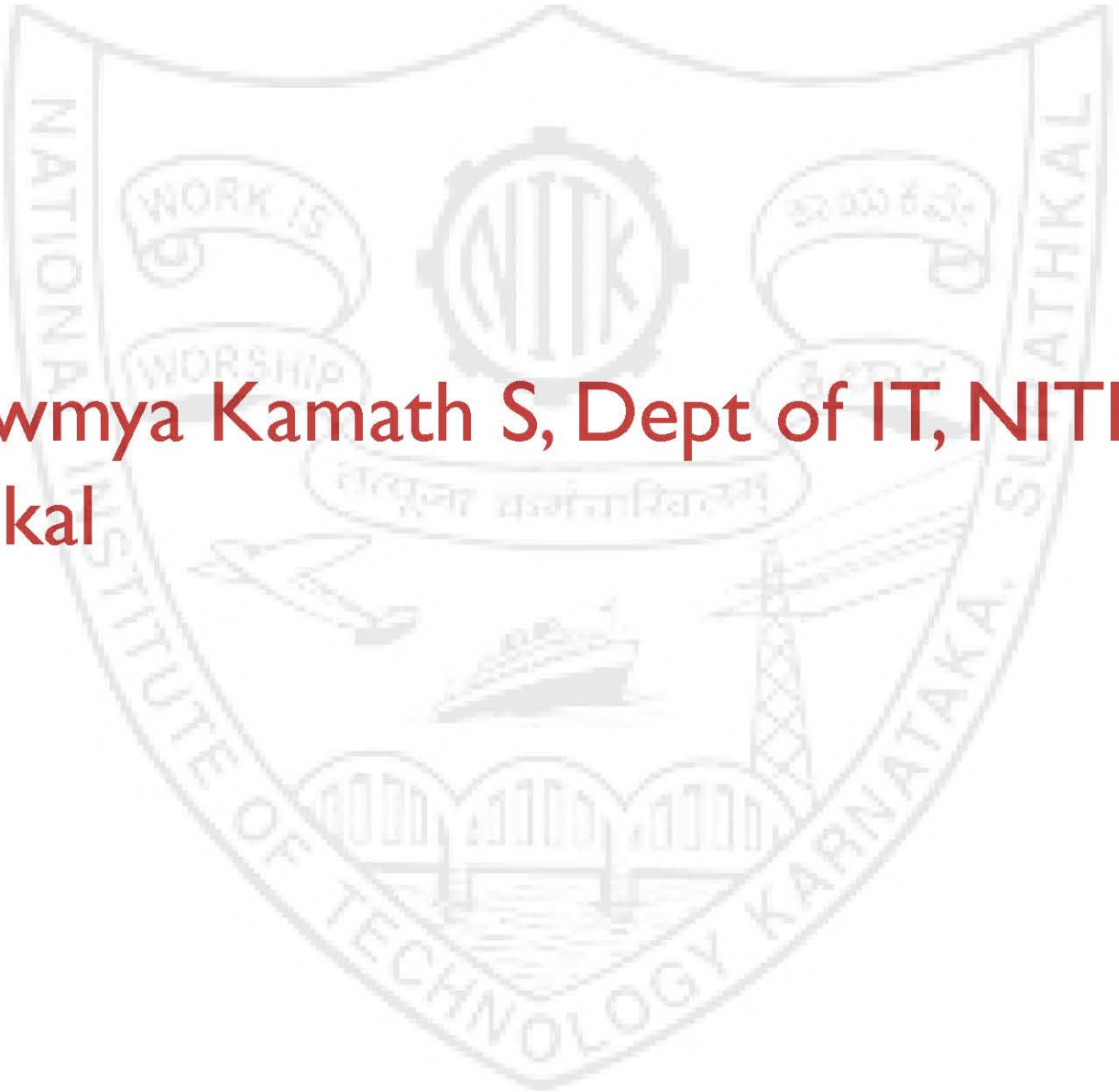
Van Rijsbergen, Cornelis. "A theoretical basis for the use of co-occurrence data in information retrieval." *Journal of documentation* 33.2 : 106-119.

Binary Independence Model

- **Idea 1**: Estimate how terms contribute to relevance
  - How do scores like tf, df, idf, and document length influence our judgments about document relevance?

Binary Independence Model

- Two possible scenarios
  - Some information about relevance of documents is known
    - Robertson-Sparck-Jones Model (1976)
  - Relevance information is not known
    - Croft & Harper Model (1979)



Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

Binary Independence Model
(Robertson & Sparck Jones, 1976)

RSJ model →

$$P(R=1|Q,D) \overset{rank}{\approx} \prod_{i=1,d_i=q_i=1}^{k} \left[\frac{\hat{p}_i(1-\hat{q}_i)}{\hat{q}_i(1-\hat{p}_i)}\right]$$

Where, for each term k_i,

$\hat{p}_i$ = prob. that term k_i occurs in a relevant doc

$\hat{q}_i$ = prob. that term k_i occurs in a non-relevant doc

Binary Independence Model
(Robertson & Sparck Jones, 1976)

Probabilities *(called Relevance judgments or Retrieval Status Value)* are given by,

$$\hat{p}_i = \frac{N_1(i) + 0.5}{N_1 + 1} \qquad \hat{q}_i = \frac{N_0(i) + 0.5}{N_0 + 1}$$

Where,

$N_1(i)$ = No.of relevant docs with term k_i

$N_0(i)$ = No.of non-relevant docs with term k_i

N_1 = No.of relevant docs;

N_0 = No.of non-relevant docs

BIM (Class Exercise)

- Given a corpus, with known relevance/non-relevance for each document w.r.t a query Q, compute the relevance judgment for a **newly added** document for a BIM based IR System.

RELEVANT DOCS

D_1 = {a, b, c, b, d}

D_2 = {b, e, f, b}

NON- RELEVANT DOCS

D_3 = {b, g, c, d}

D_4 = {b, d, e}

D_5 = {a, b, e, g}

NEWLY ADDED DOC

D_6 = {b, g, h}

BIM (Class Exercise)

- Given

RELEVANT DOCS

D_1 = {a, b, c, b, d}

D_2 = {b, e, f, b}

NON- RELEVANT DOCS

D_3 = {b, g, c, d}

D_4 = {b, d, e}

D_5 = {a, b, e, g}

Find relevance judgment for the newly added doc D_6 = {b, g, h}

→ Apply BIM - RSJ model

BIM (Class Exercise)

- BIM - RSJ model $P(R=1|Q,D) \overset{rank}{\approx} \prod_{i=1,d_i=q_i=1}^{k} \left[\frac{\hat{p}_i(1-\hat{q}_i)}{\hat{q}_i(1-\hat{p}_i)}\right]$

Where, for each term k_i,

$\hat{p}_i$ = prob. that term k_i occurs in a relevant doc

$\hat{q}_i$ = prob. that term k_i occurs in a non-relevant doc

$$\hat{p}_i = \frac{N_1(i)+0.5}{N_1+1}$$

$$\hat{q}_i = \frac{N_0(i)+0.5}{N_0+1}$$

Where, $N_1(i)$ = No.of relevant docs with term k_i

$N_0(i)$ = No.of non-relevant docs with term k_i

N_1 = No.of relevant docs;

N_0 = No.of non-relevant docs

BIM (Class Exercise)

- For given query Q
  - **RELEVANT DOCS**
    - $D_1 = \{a, b, c, b, d\}$
    - $D_2 = \{b, e, f, b\}$

    → i.e., $N_1 = 2$
  - **NON- RELEVANT DOCS**
    - $D_3 = \{b, g, c, d\}$
    - $D_4 = \{b, d, e\}$
    - $D_5 = \{a, b, e, g\}$

    → i.e., $N_0 = 3$

BIM (Class Exercise)

RELEVANT DOCS
D_1 = {a, b, c, b, d}
D_2 = {b, e, f, b}

NON- RELEVANT DOCS
D_3 = {b, g, c, d}
D_4 = {b, d, e}
D_5 = {a, b, e, g}

- Construct term matrix

$$P(R=1|Q,D) \overset{rank}{\approx} \prod_{i=1,d_i=q_i=1}^{k} \left[\frac{\hat{p}_i(1-\hat{q}_i)}{\hat{q}_i(1-\hat{p}_i)}\right] \qquad \hat{p}_i = \frac{N_1(i)+0.5}{N_1+1} \qquad \hat{q}_i = \frac{N_0(i)+0.5}{N_0+1}$$

term	a	b	c	d	e	f	g	h
$N_{1(i)}$	2	2	1	1	1	1	0	0
$N_{o(i)}$	1	3	1	2	2	0	2	0
$\hat{p}_i$	$\frac{2+0.5}{2+1}$	$\frac{2+0.5}{2+1}$	$\frac{1.5}{3}$	$\frac{1.5}{3}$	$\frac{1.5}{3}$	$\frac{1.5}{3}$	$\frac{0.5}{3}$	$\frac{0.5}{3}$
$\hat{q}_i$	$\frac{1.5}{4}$	$\frac{3.5}{4}$	$\frac{1.5}{4}$	$\frac{2.5}{4}$	$\frac{2.5}{4}$	$\frac{0.5}{4}$	$\frac{2.5}{4}$	$\frac{0.5}{4}$

BIM (Class Exercise)

- Compute relevance of newly added doc D_6 = {b, g, h}

$$P(R=1|Q,D) \overset{rank}{\approx} \prod_{i=1,d_i=q_i=1}^{k} \left[\frac{\hat{p}_i(1-\hat{q}_i)}{\hat{q}_i(1-\hat{p}_i)}\right]$$

$$P(R=1|Q,D_6) = \prod_{i\in D6} \frac{\hat{p}_i(1-\hat{q}_i)}{\hat{q}_i(1-\hat{p}_i)}$$

For *b* For *g* For *h*

$$= \prod_{i\in D6} \frac{\left(\frac{2.5}{3}(1-\frac{3.5}{4})\right).\left(\frac{0.5}{3}(1-\frac{2.5}{4})\right).\left(\frac{0.5}{3}(1-\frac{0.5}{4})\right)}{\left(\frac{3.5}{4}(1-\frac{2.5}{3})\right).\left(\frac{2.5}{4}(1-\frac{0.5}{4})\right).\left(\frac{0.5}{4}(1-\frac{0.5}{3})\right)}$$

= 0.1199

BIM (Class Exercise)

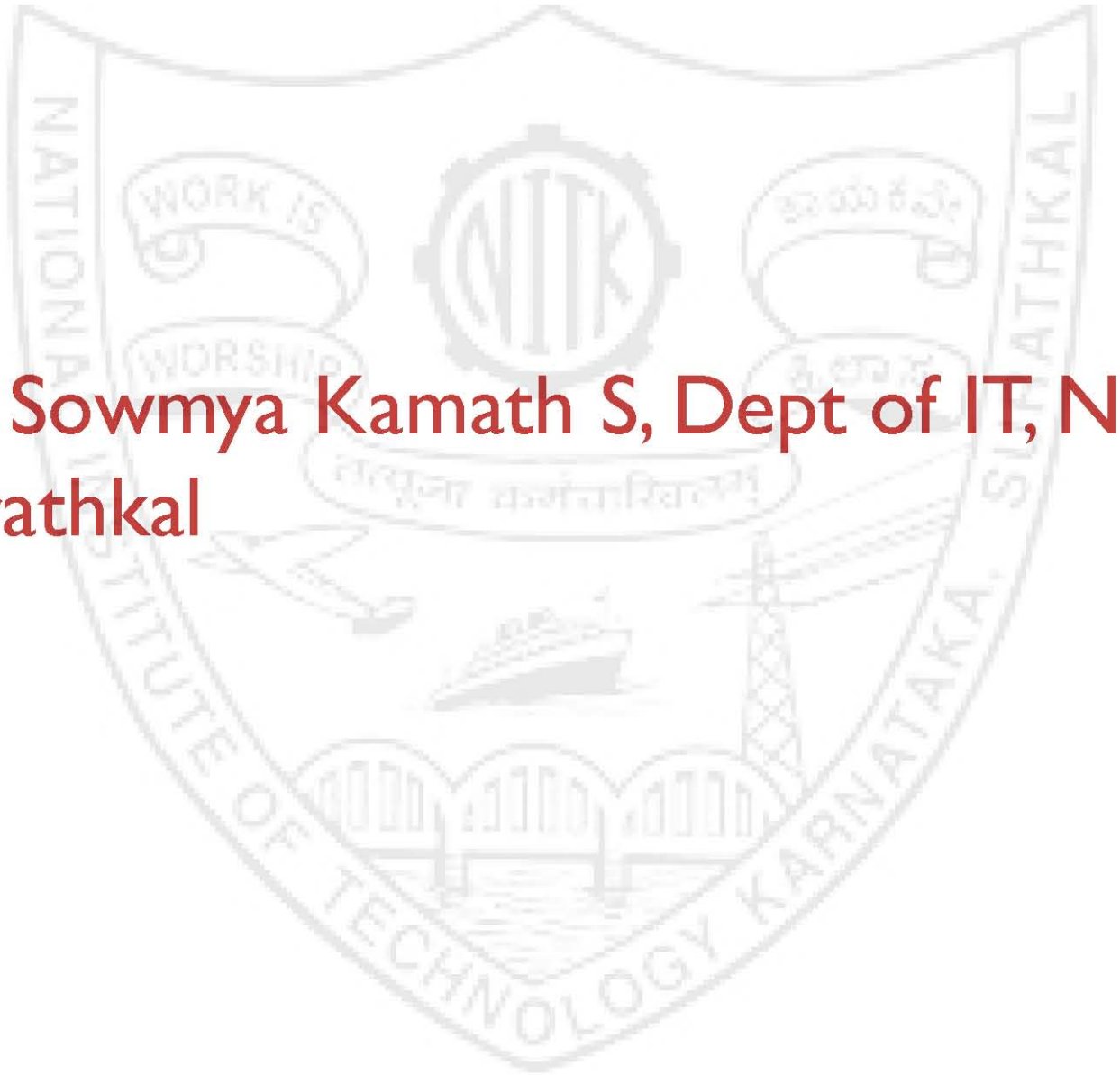
- Compute relevance of newly added doc D_6 = {b, g, h}

$$P(R=1|Q,D) \overset{rank}{\approx} \prod_{i=1,d_i=q_i=1}^{k} \left[\frac{\hat{p}_i(1-\hat{q}_i)}{\hat{q}_i(1-\hat{p}_i)}\right]$$

$$P(R=1|Q,D_6) = \prod_{i \in D6} \frac{\hat{p}_i(1-\hat{q}_i)}{\hat{q}_i(1-\hat{p}_i)}$$

= 0.1199

→ Newly added D_6 is not considered non-relevant, it starts with a probability of 0.1199 that it is at R=1 for the given Q



Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

BIM: No Relevance Info
(Croft & Harper 79)

- **Given a corpus, where relevance information is not known, the ranking of each document w.r.t to a given query Q for BIM based system is given by,**

$$Rel - value \overset{Rank}{\approx} \sum_{i = D \wedge Q} \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

Where,

N: no.of documents in collection

n_i: no.of documents in which term k_i occurs

BIM: No Relevance Info (Class Exercise)

- Given the new corpus of six documents, where relevance information is not known, compute the ranking of each document w.r.t to a given query Q for BIM based system

CORPUS

D_1 = {a, b, c, b, d}
D_2 = {b, e, f, b}
D_3 = {b, g, c, d}
D_4 = {b, d, e}
D_5 = {a, b, e, g}
D_6 = {b, g, h}

Query Q = {a, c, h}

BIM (Class Exercise 2)

- Given

CORPUS

D_1 = {a, b, c, b, d}
D_2 = {b, e, f, b}
D_3 = {b, g, c, d}
D_4 = {b, d, e}
D_5 = {a, b, e, g}
D_6 = {b, g, h}

Query Q = {a, c, h}

Find relevance judgment for each document w.r.t Q

→ Apply BIM – Croft Harper model

BIM (Class Exercise 2)

CORPUS N = 6

- Construct term matrix

D_1 = {a, b, c, b, d}; D_2 = {b, e, f, b}; D_3 = {b, g, c, d}
D_4 = {b, d, e}, D_5 = {a, b, e, g}; D_6 = {b, g, h}

Term →	a	b	c	d	e	f	g	h
n_i	2	6	2	3	3	1	3	1
$\frac{N-n_i+0.5}{n_i+0.5}$	$\frac{4.5}{2.5}$	$\frac{0.5}{6.5}$	$\frac{4.5}{2.5}$	$\frac{3.5}{3.5}$	$\frac{3.5}{3.5}$	$\frac{5.5}{1.5}$	$\frac{3.5}{3.5}$	$\frac{5.5}{1.5}$

BIM (Class Exercise 2)

CORPUS

D_1 = {a, b, c, b, d}; D_2 = {b, e, f, b};
D_3 = {b, g, c, d}; D_4 = {b, d, e},
D_5 = {a, b, e, g}; D_6 = {b, g, h}

- Compute relevance judgment of each document for the Query = {a, c, h}

$$Rel-value \overset{Rank}{\approx} \prod_{i \in D \cap Q}^{k} \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

First, compute relevance judgment of document D_1 for Q

$$P(R=1 \mid D_1) \overset{Rank}{\approx} \prod_{i \in Q \cap D_1}^{k} \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

Only *a* and c are common for D_1 and Q

$$= \log \frac{4.5}{2.5} \cdot \frac{4.5}{2.5}$$

= 0.511

BIM (Class Exercise 2)

CORPUS

D_1 = {a, b, c, b, d}; D_2 = {b, e, f, b};
D_3 = {b, g, c, d}; D_4 = {b, d, e},
D_5 = {a, b, e, g}; D_6 = {b, g, h}

Similarly, compute relevance judgment of documents D_2 to D_6 for Q = {a, c, h}

$$P(R=1|D_2)=0$$

$$P(R=1|D_3)=\log\frac{4.5}{2.5}=0.255$$

$$P(R=1|D_4)=0$$

$$P(R=1|D_5)=\log\frac{4.5}{2.5}=0.255$$

$$P(R=1|D_6)=\log\frac{5.5}{1.5}=0.564$$

BIM (Class Exercise 2)

CORPUS

D_1 = {a, b, c, b, d}; D_2 = {b, e, f, b};
D_3 = {b, g, c, d}; D_4 = {b, d, e},
D_5 = {a, b, e, g}; D_6 = {b, g, h}

Similarly, compute relevance judgment of documents D_2 to D_6 for Q

$$P(R=1|D_2)=0$$

$$P(R=1|D_3)=\log\frac{4.5}{2.5}=0.255$$

$$P(R=1|D_4)=0$$

$$P(R=1|D_5)=\log\frac{4.5}{2.5}=0.255$$

$$P(R=1|D_6)=\log\frac{5.5}{1.5}=0.564$$

Ranking R|Q = D_6 D_1 D_3 D_5 D_2 D_4

BIM: Summary

- Uses only term presence/absence, thus also referred to as **Binary Independence Model**
  - Essentially Naïve Bayes adapted for document ranking
- Can be adapted for both no relevance or pseudo-relevance feedback
- *Issues*: performance isn't as good as tuned VS model

Further reading...

- Robertson, S E, and Spark Jones, Karen. "The probability ranking principle in IR." *Journal of documentation* 33.4 (1977): 294-304.

- Croft, W. B, and Harper, D. J. "Using probabilistic models of document retrieval without relevance information." *Journal of documentation* 35.4 (1979): 285-295.