Jul – Nov 2022
IT458

IR Models for Unstructured Text

Probabilistic IR – Best Match Models

From BIM to BM25...

- BIM was originally designed for
  - short catalog records
  - abstracts of fairly consistent length

- For modern full-text search collections, a model should also pay attention to other document aspects -
  - term frequency
  - document frequency
  - document length

BM25 (Best Match 25/Okapi Weighting)

- Created using variations of the probabilistic model.
  - First applied to and tested on the Okapi system.

- Incorporates all three rank influencing factors -
  - term frequency
  - document frequency
  - document length normalization

* The classic probabilistic model considers only the first one of these principles.

BM1 Formula

- BM1 Ranking formula *(Best Match 1)*

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

Where,

N: no.of documents in collection

n_i: no.of documents in which term k_i occurs

BM1 Formula

- BM1 Ranking formula *(Best Match 1)*

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

Where,

N: no.of documents in collection

n_i: no.of documents in which term k_i occurs

* used as a ranking model when relevance information is not given.

Towards BM11 and BM15 Formulae

- STEP 1: improving ranking by introducing the **term-frequency factor $F_{i,j}$** in the BM1 formula.

Towards BM11 and BM15 Formulae

- STEP 1: improving ranking by introducing the **term-frequency factor $F_{i,j}$** in the BM1 formula.

$$\mathcal{F}_{i,j} = S_1 \times \frac{f_{i,j}}{K_1 + f_{i,j}}$$

- Where,
  - $f_{i,j}$ is the frequency of term k_i within document d_j
  - K_1 is a constant chosen heuristically for each collection
  - S_1 is a scaling constant, normally set to $S_1 = (K_1 + 1)$

Towards BM11 and BM15 Formulae

- STEP 1: improving the ranking by introducing the **term-frequency factor $F_{i,j}$** in the BM1 formula

$$\mathcal{F}_{i,j} = S_1 \times \frac{f_{i,j}}{K_1 + f_{i,j}}$$

- Where,
  - $f_{i,j}$ is the frequency of term k_i within document d_j
  - K_1 is a constant chosen heuristically for each collection
  - S_1 is a scaling constant, normally set to $S_1 = (K_1 + 1)$

* If $K_1 = 0$, $F_{i,j}$ becomes equal to 1, thus no effect on the ranking.

Towards BM11 and BM15 Formulae

- STEP 2: modify the $F_{i,j}$ factor by adding **document length normalization** to it.

Towards BM11 and BM15 Formulae

- STEP 2: modify the $F_{i,j}$ factor by adding **document length normalization** to it.

$$\mathcal{F}'_{i,j} = S_1 \times \frac{f_{i,j}}{\frac{K_1 \times len(d_j)}{avg\_doclen} + f_{i,j}}$$

- Where,
  - $len(d_j)$ = length of document d_j (for e.g. the number of terms in the document)
  - avg_doclen = average document length for the collection.

Towards BM11 and BM15 Formulae

- STEP 3: Next, add **a correction factor $G_{j,q}$** dependent on the document and query lengths.

Towards BM11 and BM15 Formulae

- STEP 3: Next, add a **correction factor $G_{j,q}$** dependent on the document and query lengths.

$$\mathcal{G}_{j,q} = K_2 \times len(q) \times \frac{avg\_doclen - len(d_j)}{avg\_doclen + len(d_j)}$$

- Where,
  - *len(q)* = query length (number of terms in the query)
  - K_2 = a constant

Towards BM11 and BM15 Formulae

- STEP 4: add an additional factor for taking into account **term frequencies within queries.**

Towards BM11 and BM15 Formulae

- STEP 4: add an additional factor for taking into account **term frequencies within queries.**

$$\mathcal{F}_{i,q} = S_3 \times \frac{f_{i,q}}{K_3 + f_{i,q}}$$

- Where,
  - $f_{i,q}$ = frequency of term k_i within query q
  - K_3 = a constant
  - S_3 = an scaling constant related to K_3, normally set to $S_3 = (K_3 + 1)$

Various BM* Formulae

- Various BM formulae were defined based on these factors -

$$BM1(d_j,q) \approx \prod_{i \in D \cap Q}^{k} \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

$$BM15(d_j,q) \approx G_{j,q} + \prod_{i \in D \cap Q}^{k} F_{i,j} \times F_{i,q} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

$$BM11(d_j,q) \approx G_{j,q} + \prod_{i \in D \cap Q}^{k} F'_{i,j} \times F_{i,q} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Various BM* Formulae

- Various BM formulae were defined based on these factors -

$$BM1(d_j,q) \approx \prod_{i \in D \cap Q}^{k} \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

$$BM15(d_j,q) \approx G_{j,q} + \prod_{i \in D \cap Q}^{k} F_{i,j} \times F_{i,q} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

$$BM11(d_j,q) \approx G_{j,q} + \prod_{i \in D \cap Q}^{k} F'_{i,j} \times F_{i,q} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

* Experiments using TREC data have shown that BM11 outperforms BM15

Simplifying the BM* Formulae

- Simplified BM* equations based on empirical results on TREC data –
  - best value of K2 is very near to 0, which almost eliminates the $G_{j,q}$ factor from these equations.

Simplifying the BM* Formulae

- Simplified BM* equations based on empirical results on TREC data –
  - best value of K2 is very near to 0, which almost eliminates the $G_{j,q}$ factor from these equations.
  - Good values for scaling constants S1 and S3 are K1 + 1 and K3 + 1, respectively

Simplifying the BM* Formulae

- Simplified BM* equations based on empirical results on TREC data –
  - best value of K2 is very near to 0, which almost eliminates the $G_{j,q}$ factor from these equations.
  - Good values for scaling constants S1 and S3 are K1 + 1 and K3 + 1, respectively
  - making K3 very large gives better results.
    - As a result, the $F_{i,q}$ factor is reduced simply to $f_{i,q}$

Simplifying the BM* Formulae

- Simplified BM* equations based on empirical results on TREC data –
  - best value of K2 is very near to 0, which almost eliminates the $G_{j,q}$ factor from these equations.
  - Good values for scaling constants S1 and S3 are K1 + 1 and K3 + 1, respectively
  - making K3 very large gives better results
    - As a result, the $F_{i,q}$ factor is reduced simply to $f_{i,q}$
  - For short queries, it can be assumed that $f_{i,q}$ is 1 for all terms.

Simpler BM1, BM11 and BM15 Formulae

$$BM1(d_j, q) \approx \prod_{i \in D \cap Q}^{k} \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

$$BM15(d_j, q) \approx \prod_{i \in D \cap Q}^{k} \frac{(K_1 + 1) f_{i,q}}{(K_1 + f_{i,q})} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

$$BM11(d_j, q) \approx \prod_{i \in D \cap Q}^{k} \frac{(K_1 + 1) f_{i,q}}{\left(\frac{K_1 . len(d_j)}{avg\_doclen} + f_{i,q}\right)} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

BM25 (Best Match 25/Okapi Weighting)

- not a single function, but actually a whole family of scoring functions, with slightly different components and parameters.
- ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document.
- Named for the first system to use it (Okapi information retrieval system), implemented at City University, London in the 1980s.

BM25 Ranking Formula

- BM25: a combination of the BM11 and BM15 models
  - defined by an additional factor $B_{i,j}$

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1) f_{i,j}}{K_1 \left[(1 - b) + b \frac{len(d_j)}{avg\_doclen} \right] + f_{i,j}}$$

- where b is a constant with values in the interval [0, 1]
  - If $b = 0$, reduces to the BM15 term frequency factor
  - If $b = 1$, reduces to the BM11 term frequency factor

BM25 Ranking Formula

- BM25: a combination of the BM11 and BM15 models
  - defined by an additional factor $B_{i,j}$

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1) f_{i,j}}{K_1 \left[(1 - b) + b \frac{len(d_j)}{avg\_doclen} \right] + f_{i,j}}$$

- where b is a constant with values in the interval [0, 1]
  - If $b = 0$, reduces to the BM15 term frequency factor
  - If $b = 1$, reduces to the BM11 term frequency factor

* For values of b between 0 and 1, the equation provides a combination of BM11 and BM15.

BM25 Ranking Formula

- The ranking equation for the Simplified BM25 model -

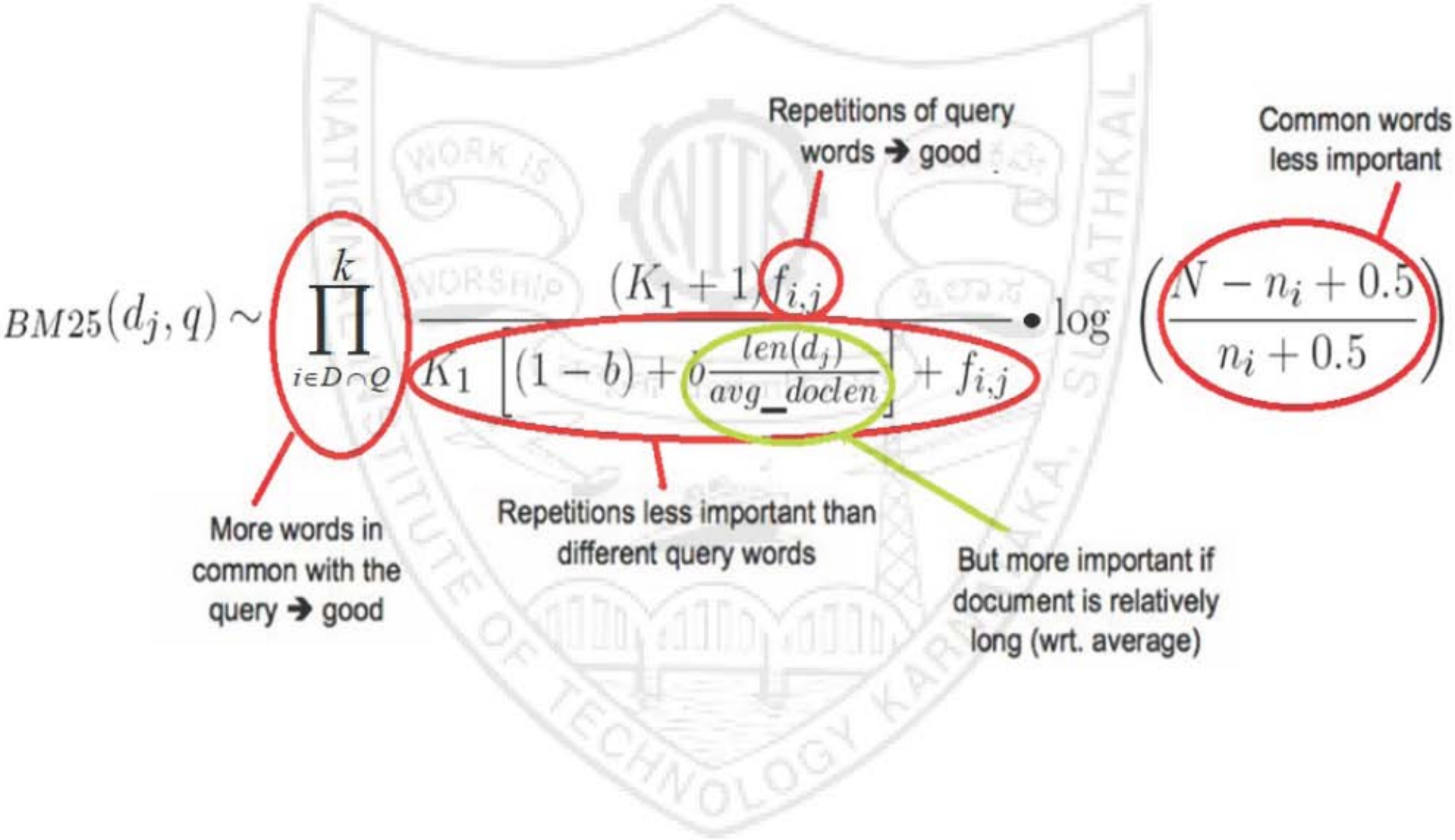
$$BM25(d_j, q) \approx \prod_{i \in D \cap Q}^{k} B_{i,j} \times \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

  - Where,

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1) f_{i,j}}{K_1 \left[(1 - b) + b \frac{len(d_j)}{avg\_doclen} \right] + f_{i,j}}$$

  - K1 and b are empirical constants.

BM25: an intuitive view



BM25 Ranking Formula

- Some Observations :
  - $K_1 = 1$ works quite well with real collections
  - *b* should be kept closer to 1 to emphasize the document length
    - This is due to the normalization effect present in the BM11 formula (b = 0.75 is a reasonable assumption)

* Constant values can be fine-tuned for particular collections through proper experimentation.

BM25 (Homework)

- Given a corpus of six documents, where relevance information is not known, compute the ranking of each document w.r.t to a given query Q, with due consideration to ranking factors such as term freq, inv. doc. freq and doc length normalization. Assume $K_1 = 1$ and $b = 0.5$.

CORPUS

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

Other models built on BM25...

- **BM25+**
  - Ranks long documents which match the query term fairly in comparison to shorter documents.
  - Achieves better term frequency normalization by document length.

- **BM25F** -
  - For documents composed of several fields (such as headlines, main text, anchor text)
  - Allows for possibly different degrees of importance, term relevance saturation and length normalization.

Further reading...

- Spärck Jones, K.; Walker, S.; Robertson, S. E. "A probabilistic model of information retrieval: Development and comparative experiments". Information Processing & Management. **36** (6): 779–808.

- Stephen Robertson & Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". **3** (4). Trends in Inf. Retr.: 333–389.