Okapi at TREC-3
By Muhammad Nur Yanhaona (Yan)
Basic Overview

• Discusses a particular stage of evolution of Okapi IR System

Okapi IR System

• Developed by City University of London
• Initially a pure probabilistic approach to Information Retrieval
• Was already modified to include ad-hoc parameters at TREC-2

At TREC-3 Okapi team tried

• Refinement of document indexing (aka term weighting) function
  • This gives us BM25
• Query expansion without feedback
• Passage retrieval
• Interactive search
• Automatic and manual query routing
Focus of the Presentation

- Evolution of Probabilistic Document Ranking Through the Lenses of BM25
- That I understood from
  - Probabilistic Models for Automatic Indexing
    Abraham Bookstein and Don R. Swanson (1976)
  - A Probabilistic Approach to Automatic Keyword Indexing (Part II)
    Stephen P. Harter (1975)
  - Relevance Weighting of Search Terms
    S E Robertson and K. Sparck Jones (1976)
  - Okapi at TREC-2
    S E Robertson, S Walker S Jones, MM Hancock-Beaulieu, and M Gatford (1994)
  - And, Okapi at TREC-3
    S E Robertson, S Walker S Jones, MM Hancock-Beaulieu, and M Gatford (1995)
Probabilistic Document Ranking

- “How occurrence of words relate to the content of the documents in which they occur”
- Model of word occurrence that can serve as the basis of indexing algorithm
- Indexing suitability of a word based on its clustering tendency in entire collection of documents
  - Non-indexing words should be distributed randomly
  - Indexing words should form clusters
- Empirically verified
Index Term Distribution Model

- Collections of documents are divided into subclasses for individual words
- Within each class, occurrence of that word follows Poisson distribution

\[ P_{\text{relevance}}(k) = \sum_i r_i P_r\{i|K\} = \frac{\sum_i r_i \pi_i \lambda_i^k e^{-\lambda_i}}{\sum_i \pi_i \lambda_i^k e^{-\lambda_i}} \]

- \( \pi_i = \) Probability of a document to be in \( i^{th} \) group
- \( r_i = \) Probability of a document in \( i^{th} \) group to be relevant
- \( \lambda_i = \) expected number of word occurrences in the document

- Simplifying assumption is the 2-Poisson model (two document subclasses)
\[ f(k) = \pi(\lambda_1^k/k!) e^{-\lambda_1} + (1 - \pi)(\lambda_2^k/k!) e^{-\lambda_2} \]

A Probabilistic Approach to Automatic Keyword Indexing
And, Probabilistic Models for Automatic Indexing
Criteria for Retrieving a Document

- A IR system may retrieve some unrelated documents
- It may not some related documents
- The criteria for retrieval is $P(\text{Relevance}) > \text{Constant Cost}$
- Assuming I representing the relevant document subclass

$$P(d \in I|K) = \frac{\pi(\lambda^k/k!)}{f(k)} e^{-\lambda_1}$$

return $d$ if $P(d \in I|K)$ is above certain threshold

- However, rarely we have a single term query

Probabilistic Models for Automatic Indexing
Foundation of Relevance Weighting

- Explicit document ranking function embody the assumption that rank of a document is the sum of the rank of occurred terms.
- In addition two empirically validated assumptions regarding significance of a term’s occurrence in a document
  1. Independence: Distribution of terms in relevant document is independent and distribution of terms in irrelevant documents is independent
  2. Ordering: Probable relevance depends on both presence and absence of a term

\[
W^{(1)} = \log \frac{r + 0.5}{R - r + 0.5} \cdot \frac{N - r + 0.5}{N - n - R + r + 0.5}
\]

Okapi at TREC-1
Okapi at TREC-2

- Performance at TREC-1 was not particularly good
- How to incorporate term frequency in the equation?
- Given 2-Poisson Distribution Model is too costly
- Approximate that distribution with a simpler equation

1. Introducing document term frequency
   \[ w = w^{(i)} \times \frac{tf}{(k_1 + tf)} \]
2. Introducing query term frequency
   \[ w = w^{(i)} \times \frac{qtf}{(k_3 + qtf)} \]
3. Introducing document length
   \[ w = w^{(i)} \times \frac{tf}{((k \times dl)/avdl) + tf} \text{ and a global correction} \]
Okapi at TREC-2

**BM 11**

\[
w = s_1 s_3 \times \frac{tf}{k_1 \times dl + tf} \times w^{(1)} \times \frac{qtf}{k_3 + qtf} \text{ and } k_2 \times n_q \times \frac{avdl - dl}{avdl + dl}
\]

**BM 15**

\[
w = s_1 s_3 \times \frac{tf}{k_1 + tf} \times w^{(1)} \times \frac{qtf}{k_3 + qtf} \text{ and } k_2 \times n_q \times \frac{avdl - dl}{avdl + dl}
\]

$s_1$ and $s_3$ are scaling factors related to $k_1$ and $k_3$ respectively.
Okapi at TREC-3

- Makes the term frequency element flexible
- By adding yet another parameter!

\[ w = s_1 s_2 \times \frac{tf^c}{K^c + tf^c} \times w^{(1)} \times \frac{qt_f}{k_3 + qt_f} \text{ and } k_2 \times nq \frac{avdl - dl}{avdl + dl} \]

Here, \( K = k_1((1 - b) + b \frac{dl}{avdl}) \)

Yet after all that,

- Query expansion
- Passage retrieval
- Feedback based query routing
Final Assessment

- Impreciseness of relevance judgment necessitates ad-hoc modifications
  - 2-Poisson model itself was a simplification
  - The final model deviates far from even that
  - Diminishing gain with further complexity in the ranking model

Thanks