Okapi BM25 and Term Specificity

Paloma García de Herreros García
Eirini Tsakiri
Okapi BM25: a non-binary model

IntIR 11.4.3

- Okapi vs BIM (Binary Independence Model)
- A probabilistic model with a small number of parameters
- Used for document scoring
Retrieval Status Value

The simplest scoring systems are in the formulas below computing the sum of the idf for all terms in the query.

\[
RSV_d = \sum_{t \in q} \log \frac{N}{df_t}
\]

\[
RSV_d = \sum_{t \in q} \log \frac{N - df_t + \frac{1}{2}}{df_t + \frac{1}{2}}
\]
Build up of the standard form

By factoring in the frequency of each term and document length:

$$RSV_d = \sum_{t \in q} \log \left[ \frac{N}{df_t} \right] \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b \times (L_d/L_{ave})) + tf_{td}}$$

For longer queries it is possible to use similar weighting for the terms, taking into account the frequency of them:

$$RSV_d = \sum_{t \in q} \left[ \log \frac{N}{df_t} \right] \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b \times (L_d/L_{ave})) + tf_{td}} \cdot \frac{(k_3 + 1)tf_{tq}}{k_3 + tf_{tq}}$$
Build up of the standard form (cont’d)

If there is available information on the relevance of the documents, the full form can be used in place of the approximation of the idf:

\[
RSV_d = \sum_{t \in q} \log \left[ \frac{(|VR_t| + \frac{1}{2})/(|VNR_t| + \frac{1}{2})}{(df_t - |VR_t| + \frac{1}{2})/(N - df_t - |VR| + |VR_t| + \frac{1}{2})} \right] \\
\times \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b(L_d/L_{ave})) + tf_{td}} \times \frac{(k_3 + 1)tf_{tq}}{k_3 + tf_{tq}}
\]
A statistical interpretation of term specificity and its application in retrieval

Karen Spärck Jones
1972

Exhaustivity and Specificity:

- The **exhaustivity** of a document is the number of terms assigned to it
- “If the exhaustivity of a document description is *increased* by the assignment of more terms [...] the chance of the document matching a request is *increased*”
- In information retrieval a frequently used term is considered to be a non-specific term
Statistical Specificity

- The **specificity** of a term is the number of documents to which it pertains.
- Terms should be *weighted* according to collection frequency
- Matches on less frequent but more specific terms are of *greater value* than matches on frequent terms
- Three collections of documents were used
<table>
<thead>
<tr>
<th></th>
<th>Cranfield</th>
<th>INSPEC</th>
<th>Keen</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>number of documents</td>
<td>200</td>
<td>541</td>
</tr>
<tr>
<td></td>
<td>number of terms</td>
<td>712</td>
<td>1341</td>
</tr>
<tr>
<td></td>
<td>number of terms per document</td>
<td>32</td>
<td>12.2</td>
</tr>
<tr>
<td></td>
<td>number of documents per term</td>
<td>9</td>
<td>4.9</td>
</tr>
<tr>
<td>B</td>
<td>number of requests</td>
<td>42</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>number of terms represented</td>
<td>166</td>
<td>248</td>
</tr>
<tr>
<td></td>
<td>number of terms per request</td>
<td>6.9</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>number of documents per request term</td>
<td>31.6</td>
<td>11.5</td>
</tr>
<tr>
<td>C</td>
<td>number of retrieving terms per request</td>
<td>5</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>number of retrieving terms per document</td>
<td>1.8</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>number of retrieving terms per relevant document</td>
<td>3.6</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>number of frequent terms</td>
<td>96</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>number of frequent terms per request</td>
<td>4</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Specificity and Matching

- A possibility is to remove very frequent terms from requests.
- However, they are required for higher recall scores.

Figure 1. Recall/Precision graph for the Cranfield collection (retrieved from Spärck Jones, 1972)
Weighting by Specificity

- Proposed in the paper
- $f(n) = m$ such that
- $\frac{2^m}{2^{m-1}} < n \leq 2^m$
- The weight of a term which occurs $n$ times is $f(N) - f(n) + 1$, where $N$ equals the documents in the collection
Results of the term weighting by specificity for the 3 collections
Thank you for your attention!