Probabilistic Relevance Models

The probabilistic model [1] estimates the probability that the user will find a document d_j relevant for a user query q. The model assumes that this probability of relevance depends on the query and the document representations only. Further, it assumes that there is a subset of all documents that the user prefers as the answer set for the query q. Such an *ideal* answer set is labeled R and should maximize the overall probability of relevance to user. Documents in the set R are predicted to be relevant to the query, while documents not in this set are predicted to be non-relevant.

The BM25 measure [2] is one of the most successful ranking formula for the probabilistic model. The BM25 weighting scheme is a function of the number of occurrences of the term in a document, in the whole collection, and a function of the document length. Formally, the weight of a term i in a document d_j is defined as:

$$w_{i,j} = \frac{k_1 \times tf_{i,j}}{tf_{i,j} + k_1 \times \left(1 - b + b \times \frac{|\vec{d_j}|}{\overline{|\vec{d_j}|}}\right)} \times \log \frac{N - df_i + 0.5}{0.5}, \tag{1}$$

where $tf_{i,j}$ is the number of occurrences of the term i in the document d_j, k_1 and b are parameters, that depends on the collection and possibly on the nature of the user queries, $|\vec{d_j}|$ corresponds to a document length function, $\overline{|\vec{d_j}|}$ is the average document length, N is the number of documents in the collection, and df_i is the number of documents containing the term i.

The BM25 scheme also defines a weight of a term i in a query q. This weight is formally defined as:

$$w_{i,q} = \frac{(k_3 + 1) \times tf_{i,q}}{k_3 + tf_{i,q}}, \tag{2}$$

where $tf_{i,q}$ is the number of occurrences of the term i in the query q, and k_3 is a parameter, that depends on the collection and possibly on the nature of the user queries.

The probabilistic model computes the similarity between a document and the user query as the scalar product between the document vector $\vec{d_j}$, $1 \leq j \leq N$, and the query vector $\vec{q}$, as follows:

$$sim\,(q, d_j) = \vec{d_j} \bullet \vec{q} = \sum_{i \in q} w^{(1)} \times w_{i,j} \times w_{i,q} + k_2 \times \frac{\overline{|\vec{d_j}|} - |\vec{d_j}|}{\overline{|\vec{d_j}|} + |\vec{d_j}|}, \tag{3}$$

where $w_{i,j}$ is the weight associated with the term i in the document d_j, $w_{i,q}$ is the weight associated with the term i in the query q, $|\vec{d_j}|$ corresponds to a document length function, $\overline{|\vec{d_j}|}$ is the average document length, k_2 is another parameter, that also depends on the collection and possibly on the nature of the user queries, and $w^{(1)}$ is the Robertson-Sparck Jones weight [1], which is defined as:

$$w^{(1)} = \log \frac{(r+0.5)\,/\,(R-r+0.5)}{(df_i - r + 0.5)\,/\,(N - df_i - R + r + 0.5)}, \tag{4}$$

where N is the number of documents in the collection, df_i is the number of documents containing the term i, R is the number of relevant documents for the query q, and r is the number of relevant documents containing the term i.

To compute the ranking with regard to a user query q using the probabilistic model with the BM25 weighting scheme, we use the algorithm of Figure 1. The BM25 parameters was set with the following values to produce the baseline: $k_1 = 1.2$, $k_2 = 0$, $k_3 = 1000$, and $b = 0.75$.

BM25 (q, k_1, k_2, k_3, b, k)
    q : a set of query terms
    k_1, k_2, k_3, b : model parameters
    k : number of documents to be returned
    Let A be a set of accumulators
    Set R = 0 and r = 0
    for each term $t_i \in q$ do begin
        for each $[d_j, tf_{i,j}] \in t_i$ inverted list do begin
            if $A_j \notin A$ then $A = A \cup \{A_j\}$
            $A_j = A_j + w^{(1)} \times w_{i,j} \times w_{i,q} + k_2 \times \frac{\overline{|\vec{d_j}|} - |\vec{d_j}|}{\overline{|\vec{d_j}|} + |\vec{d_j}|}$*, from Eqs. (1, 2, and 4).*
        end
    end
    determine the k largest $A_j \in A$ and return the corresponding documents
end

Figure 1: The BM25 ranking algorithm.

References

[1] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, May-June 1976.

[2] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Proceedings of the Third Text Retrieval Conference (TReC-3)*, pages 109–126. NIST Special Publication 500-225, April 1995.