Contents

1 Probabilistic Model

- To achieve optimal performance, documents should be ranked in order of their *probability of relevance* to the query.
- Whether or not document x should be retrieved depends on:
  - $Pr(rel|x)$, the probability that a given document x is relevant, and
  - $Pr(nonrel|x)$, the probability that a given document x is nonrelevant.
- Document x should be retrieved if:

$$a_2 Pr(rel|x) \geq a_1 Pr(nonrel|x)$$

  where a_1 and a_2 are, respectively, the costs associated with the retrieval of a nonrelevant document and the nonretrieval of a relevant document. For simplicity, we assume a_1 and a_2 are the same.

$$g(x) = \frac{Pr(rel|x)}{Pr(nonrel|x)} - \frac{a_1}{a_2} > 0.$$

  Since it is not possible to determine $Pr(rel|x)$, we apply Bayes theorem to rewrite it as:

$$Pr(rel|x) = \frac{Pr(x|rel)P(rel)}{Pr(x)}$$

  - $Pr(x)$ is the probability of observing x (whether x is relevant or not).
  - $P(rel)$ is the a priori probability of relevance (i.e., the probability of observing a set of relevant documents).
  - $Pr(x|rel)$ is the probability that x is in the given set of relevant documents.
  - Similar formulation can be obtained for $Pr(nonrel|x)$.
- The document ranking function (or discrimination function) can be rewritten as (after dropping a_1/a_2):

$$\log g(x) = \log \frac{Pr(x|rel)Pr(rel)}{Pr(x)} \frac{Pr(x)}{Pr(x|nonrel)Pr(nonrel)}$$

$$\log g(x) = \log \frac{Pr(x|rel)}{Pr(x|nonrel)} + \log \frac{Pr(rel)}{Pr(nonrel)}.$$

- Since $Pr(x|rel)$ and $Pr(x|nonrel)$ are unknown quantities, they need to be replaced in terms of the keywords in the document.
- Assuming terms occur independently in relevant and nonrelevant documents:

$$\log g(x) = \sum_{i=1}^{t} \log \frac{Pr(x_i|rel)}{Pr(x_i|nonrel)} + constant.$$

where $Pr(x_i|rel)$ is the probability that a relevant document contains term x_i; $Pr(x_i|nonrel)$ interpreted likewise.

- Considering document $D = \langle d_1, d_2, ...d_t \rangle$, where d_i is the weight of term i:

$$\log g(x) = \sum_{i=1}^{t} \log \frac{Pr(x_i = d_i|rel)}{Pr(x_i = d_i|nonrel)} + constant.$$

where $Pr(x_i = d_i|rel)$ is the probability that a relevant document x contains term x_i with weight d_i; $Pr(x_i|nonrel)$ interpreted likewise.

- When $d_i = 0$, we want the contribution of term i to $g(z)$ to be zero. This can be done by introducing a normalization factor (or a transformation):

$$\log g(x) = \sum_{i=1}^{t} \log \frac{Pr(x_i = d_i|rel)}{Pr(x_i = d_i|nonrel)} \frac{Pr(x_i = 0|nonrel)}{Pr(x_i = 0|rel)} + constant$$

$$\log g(x) = \sum_{i=1}^{t} \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + constant,$$

where

$$\begin{aligned} p_i &= Pr(x_i = d_i|rel) \\ &= \text{Probability of finding } x_i \text{ in a relevant document.} \\ q_i &= Pr(x_i = d_i|nonrel) \\ &= \text{Probability of finding } x_i \text{ in a nonrelevant document.} \\ Pr(x_i = 0|nonrel) &= \text{Probability of not finding } x_i \text{ in a nonrelevant document.} \\ &= 1 - q_i. \end{aligned}$$

- The term-relevance weight of term x_i is:

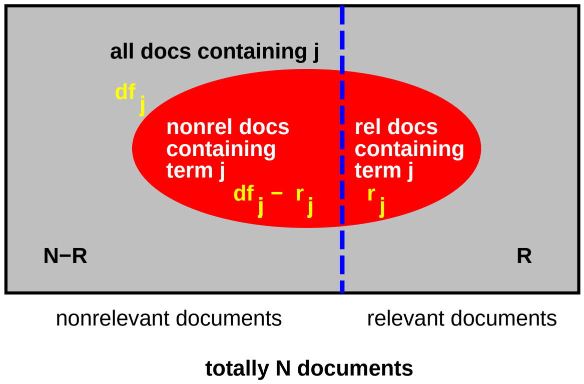
$$tr_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)} = \log \frac{Pr(x_i = d_i|rel)}{Pr(x_i = d_i|nonrel)} \frac{Pr(x_i = 0|nonrel)}{Pr(x_i = 0|rel)}.$$

- Weight of term j in document i is taken as:

$$w_{i,j} = tf_{i,j} \times tr_j.$$

1.1 Estimation of Term Occurrence Probability

- Given a query, a document collection can be partitioned into a set of relevant documents and a set of nonrelevant documents. The importance of index term j can be determined the role it plays in discriminating relevant and nonrelevant documents. The following diagram can be obtained:



- If we have *complete* information about the relevant and nonrelevant documents, p_j and q_j can be estimated by:

  $p_j = r_j/R \qquad q_j = \frac{df_j - r_j}{N-R}$

  $tr_j = \log \frac{r_j/R}{(df_j - rj)/(N-R)} \frac{1-(df_j-r_j)/(N-R)}{1-r_j/R} = \log \frac{r_j}{R-r_j} \frac{N-df_j-R+r_j}{df_j-r_j}$.

- Approximation 1: $tr_j = \log \frac{r_j}{R} \frac{N}{df_j}$.

- Approximation 2: $tr_j = \log \frac{r_j}{R} \frac{N-R}{df_j - r_j}$. (Eqns. 9.17 a and b in [S] and p. 368 in [FB])

- Approximation 3: $tr_j = \log \frac{r_j}{R-r_j} \frac{N-df_j}{df_j}$.

- tr_j can be interpreted as the power of term j in discriminating between the relevant and nonrelevant documents.

1.2 Term Occurrence Probability Without Relevance Information

- $q_j = df_j/N$ because most documents are nonrelevant; $p_j = 0.5$ for all j (quite arbitrary). Then, $tr_j = \log(N/df_j - 1)$, which is the same as the inverse document frequency.

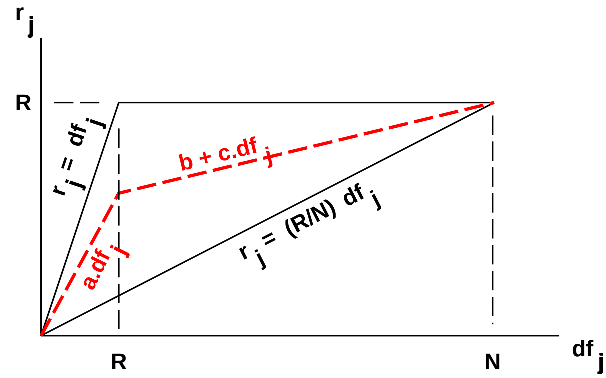
- A better approximation using Approximation 3:

  $tr_j = \log \frac{r_j}{R-r_j} + \log \frac{N-df_j}{df_j} \approx constant + \log(N/df_j - 1) \approx idf_j$.

- Approximation by interpolation:

  - Assume best case when term j appears *only* in relevant documents: $r_j = df_j$.
  - Worst case when term j spread *evenly* among the relevant and nonrelevant documents: $r_j = (R/N) \times df_j$.

– An estimation of r_j can be made somewhere in between:



The constants a, b, c can be obtained by calibration or simply assumed to be some medium values (e.g., a = 0.5).

1.3 Other Ranking Functions

- $tr_j = C + \log \frac{N - df_j}{df_j} = C + idf_j$.
- $sim_i = \sum_j (C + idf + j) f_{i,j}$

  where $f_{i,j} = k + (1-k)\frac{freq_{i,j}}{maxfreq_j}$, and $maxfreq_j$ is the frequency of the most frequent term in the document.
- When $k = 0$, $f_{i,j} = \frac{freq_{i,j}}{maxfreq_j}$.