

O modelo probabilístico é baseado no princípio da ordenação probabilística (*Probability Ranking Principle*) onde dada uma consulta q e um documento d_j relevante a q , o modelo tenta estimar a probabilidade do usuário encontrar o documento d_j . O modelo assume que para uma consulta q há um conjunto de documentos R que contém exatamente os documentos relevantes e nenhum outro, sendo este um conjunto resposta ideal que maximiza a probabilidade do usuário encontrar um documento d_j relevante a q .

Seja \overline{R}_q o complemento de R de forma que \overline{R}_q contém todos os documentos não relevantes à consulta q . Seja $P(R_q|d_j)$ a probabilidade do documento d_j ser relevante à consulta q e $P(\overline{R}_q|d_j)$ a probabilidade de d_j não ser relevante à q .

Seja $p_i = P(k_i|R_q)$ a probabilidade do termo k_i ocorrer em um documento relevante à consulta q , e $s_i = P(k_i|\overline{R}_q)$ a probabilidade do termo k_i estar presente em um documento não relevante.

A similaridade entre um documento d_j e uma consulta q é definida por:

$$sim(d_j, q) = \frac{P(R_q|d_j)}{P(\overline{R}_q|d_j)} = \prod_{i:d_i=1} \frac{p_i}{s_i} \cdot \prod_{i:d_i=0} \frac{1-p_i}{1-s_i} \quad (1)$$

A fim de obter-se uma estimativa numérica das probabilidades, o modelo probabilístico clássico atribui valores binários aos pesos os quais indicam a presença ou ausência de um termo, isto é, $w_{ij} \in \{0, 1\}$ e $w_{iq} \in \{0, 1\}$. O modelo assume o documento como uma combinação de palavras e seus pesos. O modelo também supõe que os termos ocorrem independentemente no documento, ou seja, a ocorrência de um termo não influencia a ocorrência de outro. Partindo dessas suposições, a Equação ?? passa por transformações que incluem aplicação da regra de Bayes e simplificações matemáticas, e chega-se a Equação ?? conhecida como equação de Robertson-Spaul Jones a qual é considerada a expressão clássica para ranqueamento no modelo probabilístico. Detalhes da dedução dessa equação pode ser encontrada em [1].

$$sim(d_j, q) = \sum_{i=1}^t w_{i,j} \cdot w_{i,q} \cdot \sigma_{i,R} \quad (2)$$