O modelo probabilístico é baseado no princípio da ordenação probabilística (*Probability Ranking Principle*) onde dada um 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 assumente 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. A similaridade entre um documento d_j e uma consulta q é definida por:

$$sim(d_j, q) = \frac{P(R_q|dj)}{P(\overline{R_q}|dj)} \tag{1}$$

A fim de obter-se uma estimativa numéricas das probabilidades, o modelo assume o documento como uma combinação de palavras e seus pesos aos quais atribui-se valores binários que indicam a presença ou ausência de um termo, isto é, $w_{ij} \in \{0, 1\}$ e $w_{iq} \in \{0, 1\}$. 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. Então, pode-se calcular:

$$sim(d_j, q) = \prod_{i:d_i=1} \frac{p_i}{s_i} \cdot \prod_{i:d_i=0} \frac{1 - p_i}{1 - s_i} \tag{2}$$

, onde $\prod_{i:d_i=1}$ significa o produto dos termos com valor 1.

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 2 passa por transformações que incluem aplicação da regra de Bayes e simplificações matemáticas, e chega-se a Equação 3 conhecida como equação de Robertson-Spark Jones a qual é considerada a expressão clássica para ranqueamento no modelo probabilístico. Detalhes da dedução dessa equação pode ser encontrada em [].

$$sim(d_j, q) = \sum_{i=1}^{t} w_{i,j} \cdot w_{i,q} \cdot \sigma_{i/R} \tag{3}$$

, onde

$$\sigma_{i/R} = \log \frac{p_i}{1 - p_i} + \log \frac{1 - s_i}{s_i} \tag{4}$$