O modelo probabilístico é baseado no princípio da ordenação probabilística (*Probability Ranking Principle*) onde dada um consulta q e um documento d_j ¡perfeitamente¿ relevante a q, o modelo tenta estimar a probabiliade 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 probablidade do usuário encontrar um documento d_j relevante a q.

Seja $\overline{R_q}$ o complemento de R de forma que $\bar{R}_q$ 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 definada por:

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