

CSE 8337 Spring 2007
 Probabilistic Retrieval

Retrieve a document if:

$$P(R|\vec{d}_j) > P(\neg R|\vec{d}_j)$$

∴ Define similarity using this idea:

$$\text{sim}(q, d_j) = \frac{P(R|\vec{d}_j)}{P(\neg R|\vec{d}_j)}$$

Using Bayes Rule:

$$P(R|\vec{d}_j) = \frac{P(\vec{d}_j|R)P(R)}{P(\vec{d}_j)}$$

$$P(\bar{R}|\vec{d}_j) = \frac{P(\vec{d}_j|\bar{R})P(\bar{R})}{P(\vec{d}_j)}$$

$$\therefore \text{sim}(q, d_j) = \frac{P(\vec{d}_j|R)P(R)}{P(\vec{d}_j|\bar{R})P(\bar{R})}$$

Assuming that the individual terms are independent, we can rewrite the probabilities using products of probabilities for each term.

Recall that there are different probabilities for the presence and the absence of a term in a document.
(Note: that we assume $w_{ij} \in \{0, 1\}$)

Let $P(K_i | R)$ be the probability that K_i appears in a Relevant document.
Let $P(\bar{K}_i | R)$ be the probability that K_i does not appear in a Relevant document.

$$\therefore \text{sim}(q, d_j) \sim \frac{\prod_{w_{ij}=1} P(K_i | R) \prod_{w_{ij}=0} P(\bar{K}_i | R)}{\prod_{w_{ij}=1} P(K_i | \bar{R}) \prod_{w_{ij}=0} P(\bar{K}_i | \bar{R})}$$

Note that here the products are taken only over the terms which are present or absent in the document.

Instead, what we do now is take the products over all terms. To be sure that only the correct probabilities are used (term present or absent) we add superscripts that will be 1 or 0 where appropriate.

$$\therefore \text{sim}(g, d_j) \sim \frac{\prod P(k_i | R)^{w_{i,j}} \prod P(\bar{k}_i | R)^{(1-w_{i,j})}}{\prod P(k_i | \bar{R})^{w_{i,j}} \prod P(\bar{k}_i | \bar{R})^{(1-w_{i,j})}}$$

Note that now the products are all taken across all terms.

$$\text{Also notice } P(\bar{k}_i | R) = 1 - P(k_i | R) \\ \& P(\bar{k}_i | \bar{R}) = 1 - P(k_i | \bar{R})$$

$$\therefore \text{sim}(g, d_j) \sim \frac{\prod P(k_i | R)^{w_{i,j}} (1 - P(k_i | R))^{(1-w_{i,j})}}{\prod P(k_i | \bar{R})^{w_{i,j}} (1 - P(k_i | \bar{R}))^{(1-w_{i,j})}}$$

Now we transform these products using log.

$$\begin{aligned} \text{sim}(g, d_j) &\sim \log \left(\frac{\prod P(k_i | R)^{w_{i,j}} (1 - P(k_i | R))^{(1-w_{i,j})}}{\prod P(k_i | \bar{R})^{w_{i,j}} (1 - P(k_i | \bar{R}))^{(1-w_{i,j})}} \right) \\ &= \log \left(\prod P(k_i | R)^{w_{i,j}} \right) + \log \left(\prod (1 - P(k_i | R))^{(1-w_{i,j})} \right) \\ &\quad - \log \left(\prod P(k_i | \bar{R})^{w_{i,j}} \right) - \log \left(\prod (1 - P(k_i | \bar{R}))^{(1-w_{i,j})} \right) \\ &= \sum w_{i,j} \log(P(k_i | R)) + \sum (1-w_{i,j}) \log(1 - P(k_i | R)) \\ &\quad - \sum w_{i,j} \log(P(k_i | \bar{R})) - \sum (1-w_{i,j}) \log(1 - P(k_i | \bar{R})) \end{aligned}$$

$$= \sum w_{ij} \log \left(\frac{P(k_i | R)}{P(k_i | \bar{R})} \right) + \sum \log \left(\frac{(1 - P(k_i | R))}{(1 - P(k_i | \bar{R}))} \right) \\ - \sum w_{ij} \log \left(\frac{(1 - P(k_i | R))}{(1 - P(k_i | \bar{R}))} \right)$$

This 3rd term is a constant and does not impact ranking of documents.

$$\therefore \text{sim}(q, d_j) \sim \sum w_{ij} \left(\log \left(\frac{P(k_i | R)}{P(k_i | \bar{R})} \right) - \log \left(\frac{1 - P(k_i | R)}{1 - P(k_i | \bar{R})} \right) \right)$$

To include the weight of terms in query allows the similarity to be more closely related to query.

$$\therefore \text{sim}(q, d_j) \sim \sum w_{iq} w_{ij} \left(\log \left(\frac{P(k_i | R)}{P(k_i | \bar{R})} \right) - \log \left(\frac{1 - P(k_i | R)}{1 - P(k_i | \bar{R})} \right) \right)$$

This is equivalent to the formula found on bottom of p 32 in Baeza-Yates.