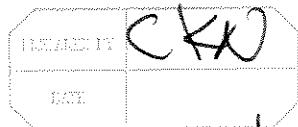


2/28/07



B-Y & R-N 2.5.4

Notes on Probabilistic IR

P. 11

Given a query g , and some documents d_1, \dots, d_n , we want to estimate the probability

$P(d_i \text{ is relevant to } g)$

and perhaps rank the documents with respect to this probability.

How to do this? Assume the query and documents are made up of terms t_i

Idea: terms that occur in known relevant docs more often than in the collection as a whole should indicate higher (probability of) relevance in documents whose relevance (yes, no) is still unknown.

User interaction may be used, but isn't necessary (!)

Assumption 1: probability of doc d_i being relevant to query g depends only on d_i and g ; not on other documents or queries or whatever

Assumption 2: a certain subset, R , of the collection exists which the user will consider relevant.

$$R = \{d_1, \dots, d_r\} \quad \begin{matrix} \text{relevant docs} \\ \text{or we think so...} \end{matrix}$$

\bar{R} is the rest of the corpus

$P(R|d_i)$ is the probability of d_i being in R (\therefore relevant)
 $P(\bar{R}|d_i)$ prob. of irrelevance

$$\text{sim}(d_i, g) = \frac{P(R|d_i)}{P(\bar{R}|d_i)}$$

Unfortunately, we don't know exactly which documents are in R , nor even how many such documents there may be

P.3
 RELEVANT BY
 DATE
 (any d.)
 i.e. is relevant

i.e. $P(d_i \text{ relevant})$

$$P(R | d_i) = \frac{P(d_i | R) P(R)}{P(d_i)}$$

$$P(\bar{R} | d_i) = \frac{P(d_i | \bar{R}) P(\bar{R})}{P(d_i)}$$

the denominators cancel out, and we have

$$\text{sim}(d_i, g) = \frac{P(d_i | R) P(R)}{P(d_i | \bar{R}) P(\bar{R})}$$

$P(R)$ (and therefore $P(\bar{R})$) are the same for all documents, so for ranking of documents

$$\text{sim}(d_i, g) \approx \frac{P(d_i | R)}{P(d_i | \bar{R})}$$

Assumption 3: index terms are independent, so ...

P. 4

$$\text{sim}(d_i, z) \sim$$

$\prod_{\text{all } k_j \text{ in } d_i}$ $P(k_j | R)$

$k_j \notin \text{indi}$

$$\frac{\prod_{\text{all } k_j \text{ in } d_i} P(k_j | R) \times \prod_{\text{all } k_j \text{ not in } d_i} P(\bar{k}_j | R)}{\prod_{\text{all } k_j \text{ in } d_i} P(k_j | \bar{R}) \times \prod_{\text{all } k_j \text{ not in } d_i} P(\bar{k}_j | \bar{R})}$$

$$P(k_j | R) + P(\bar{k}_j | R) = 1$$

since every keyword will be in
a random document in R , (or in \bar{R})
it won't - for sure. So we
have

$$\text{sim}(d_i, z) \sim$$

$$\frac{\prod_{\text{all } k_j \text{ in } d_i} P(k_j | R) \times \prod_{\text{all } k_j \text{ not in } d_i} (1 - P(k_j | R))}{\prod_{\text{all } k_j \text{ in } d_i} P(k_j | \bar{R}) \times \prod_{\text{all } k_j \text{ not in } d_i} (1 - P(k_j | \bar{R}))}$$

taking logs, noting that

$$\log \prod_{\text{all } k_j \text{ in } d_i} P(k_j | R) = \sum \log P(k_j | R)$$

we have

Q. 5

after some algebra
(and magic)

$$\text{sim}(d_i, g) \sim$$

$$\sum w_{i,j} * w_{i,j} \left(\log \frac{P(k_i | R)}{1 - P(k_i | R)} + \log \frac{1 - P(k_i | \bar{R})}{P(k_i | \bar{R})} \right)$$

$P(k_i | R)$ can start at 0.5

$P(k_i | \bar{R})$ can start at

n_i / N , where

$n_i = \# \text{ of docs containing } k_i$

$N = \# \text{ of docs in collection}$