2/28/07

PREPARED BY: CKW
DATE:

B-Y & R-N 2.5.4

Notes on Probabilistic IR

Given a query q, and some documents $d_1, \ldots, d_n$, we want to estimate the probability

$P($ d_i is relevant to q $)$

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 q depends only on d_i and q; not on other documents or queries or whatever

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

$R = \{ d_i, \ldots, d_k \}$ relevant docs or we think so...

$\overline{R}$ is the rest of the corpus

$P(R \mid d_i)$ is the probability of d_i being in R (i.e. relevant)

$P(\overline{R} \mid d_i)$ prob. of irrelevance

$$\mathrm{sim}(d_i, q) = \frac{P(R \mid d_i)}{P(\overline{R} \mid d_i)}$$

unfortunately, we don't know exactly which documents are in R, nor even how many such documents there may be i.e. P(any di is relevant)

i.e. P(di is relevant)

$$P(R \mid d_i) = \frac{P(d_i \mid R)\, P(R)}{P(d_i)}$$

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

the denominators cancel out, and we have

$$sim(d_i, q) = \frac{P(d_i \mid R)\, P(R)}{P(d_i \mid \bar{R})\, P(\bar{R})}$$

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

$$sim(d_i, q) \sim \frac{P(d_i \mid R)}{P(d_i \mid \bar{R})}$$

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

$$\text{sim}(d_i, q) \sim \frac{\displaystyle\prod_{\substack{\text{all } k_j \\ \text{in } d_i}} P(k_j \mid R) \times \prod_{\substack{\text{all } k_j \\ \text{not in } d_i}} P(\overline{k_j} \mid R)}{\displaystyle\prod_{\substack{\text{all } k_j \\ \text{in } d_i}} P(k_j \mid \overline{R}) \times \prod_{\substack{\text{all } k_j \\ \text{not in} \\ d_i}} P(\overline{k_j} \mid \overline{R})}$$

· k_j in random doc. in R

k_j not in d_i

$$P(k_j \mid R) + P(\overline{k_j} \mid R) = 1$$

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

$$\text{sim}(d_i, q) \sim \frac{\prod P(k_j \mid R) \times \prod (1 - P(k_j \mid R))}{\prod P(k_j \mid \overline{R}) \times \prod (1 - P(k_j \mid \overline{R}))}$$

taking logs, noting that

$$\log \prod P(k_j \mid R) = \sum \log P(k_j \mid R)$$

we have

after some algebra
(and magic)

$$sim(d_i, q) \sim$$

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

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

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

n_i/N, where

n_i = # of docs containing k_i

N = # of docs in collection