PROBABILISTIC

INFORMATION RETRIEVAL

Many authors divide the classical methods of information retrieval into three categories:

- **Boolean** (based on set theory, and Boolean logic)
- **Vector space** (based on linear algebra)
- **Probabilistic** (based on Bayesian statistics)

Probabilistic methods are one of the oldest but also one of the currently hottest topics in IR.

The *Probabilistic* model of *IR* (*PIR*)
is a classical model of *IR*.

Basic concept:

The probability that a document is relevant to a query is assumed to depend on

- the terms in the query and
- the terms used to index the document, only.

Given

- a user query *q* and
- a document *d* in the collection,

the probabilistic model **estimates the probability** that the user will find *d* relevant.

Initial probabilities:

Given a query *q* and a document *d* the model needs an estimate of the **probability** that the user finds *d* relevant. i.e., $P(R \mid d)$.

Similarity measure:

$s(q, d)$, the similarity of *d* to *q*, is the ratio:

$$\frac{\text{probability that document } d \text{ is relevant to } q}{\text{probability that document } d \text{ is not relevant to } q}$$

A mathematical formalisation of the Probabilistic Retrieval:

Given

- documents $\boldsymbol{D = \{d_1,\ldots, d_j,\ldots, d_m\}}$
- query $\boldsymbol{q}$,
- terms $\boldsymbol{T = \{t_1,\ldots,t_i,\ldots,t_n\}}$,
- binary weights matrix

  $matrix\ TD = (w_{ij})_{n\times m},\ (i = 1,\ldots, n, j = 1,\ldots, m)$

  where

$$w_{ij} = \begin{cases} 1 & if \quad t_i \quad occurs\ in \quad d_j \\ 0 & otherwise \end{cases},$$

Let

$P(R)$: probability that a randomly selected document is relevant to query q.

$P(\overline{R})$: probability that a randomly selected document is not relevant to query q.

DEFINITION:

Similarity between q and d_j :

$$s(q, d_j) = \frac{P(R|d_j)}{P(\overline{R}|d_j)}$$

where

$P(R \mid d_j)$ means the probability that the document d_j is relevant to query q.

Using Bayes′ Rule:

$$s(q, d_j) = \frac{P(R|d_j)}{P(\overline{R}|d_j)} = \frac{\frac{P(d_j|R)P(R)}{P(d_j)}}{\frac{P(d_j|\overline{R})P(\overline{R})}{P(d_j)}} = \frac{P(d_j|R)P(R)}{P(d_j|\overline{R})P(\overline{R})}$$

where

$P(R)$ is called the **prior** probability

$P(R \mid d_j)$ is called the **posterior** probability

$P(d_j \mid R)$ probability that if a relevant document is retrieved, it is d_j.

because $P(R)$ and $P(\overline{R})$ are constant for D,

and we are interested in ranking documents on their similarities with q, we may take:

$$s(q, d_j) = \frac{P(d_j|R)}{P(d_j|\overline{R})}\frac{P(R)}{P(\overline{R})} \approx \frac{P(d_j|R)}{P(d_j|\overline{R})}$$

Independence assumption:

Let

$\mathbf{t} = (t_1, t_2, \ldots t_n)$ be the term vector for document d_j.
$t_i = 1$ if term i is in the document and 0 otherwise.

We estimate $P(d_j \mid R)$ by $P(\mathbf{t}| R)$

If the index terms are independent

$$P(\mathbf{t}| R) = P(t_1 \cap R)\, P(t_2 \cap R) \ldots P(t_n \cap R)$$

$$= P(t_1 \mid R)\, P(t_2 \mid R) \ldots P(t_n \mid R)$$

$$= \prod_{i=1}^{n} P(t_i|R)$$

Because we assumed that terms are independent of each other, so

$$s\left(q, d_j\right) = \frac{P\left(d_j \middle| R\right)}{P\left(d_j \middle| \overline{R}\right)} = \frac{\prod_{i=1}^{n} P\left(t_i \middle| R\right)}{\prod_{i=1}^{n} P\left(t_i \middle| \overline{R}\right)}$$

(d_j is a compound event of simultaneous events t_i which occur in d_j)

Let

t_1, ..., t_i, ..., t_p be the terms occuring in query q

Then,

$s(q, d_j)$ is computed only for query terms:

$$s\left(q, d_j\right) = \frac{\prod_{t_i \in q} P\left(t_i \middle| R\right)}{\prod_{t_i \in q} P\left(t_i \middle| \overline{R}\right)}$$

Retrieval method:

Step 1. Assume $P(t_i \mid R) = 0.5$, and

$$P\left(t_i \mid \overline{R}\right) = \frac{m_i}{m}$$

where $i = 1, \ldots, n$

m_i is number of documents containing term t_i

Step 2. Compute $s(q, d_j)$.

Produce ranking.

$V \subseteq D$ denotes retrieved documents.

Step 3. Let $V_i \subseteq V$ denote documents containing $t_i \in q$

Revision:

$$P(t_i \mid R) = \frac{|V_i|}{|V|}$$

$$P\left(t_i \mid \overline{R}\right) = \frac{m_i - |V_i|}{m - |V|}$$

Step 4. Repeat from step 2.