Information Retrieval

Dra. Mireya Paredes

Probabilistic Retrieval Model

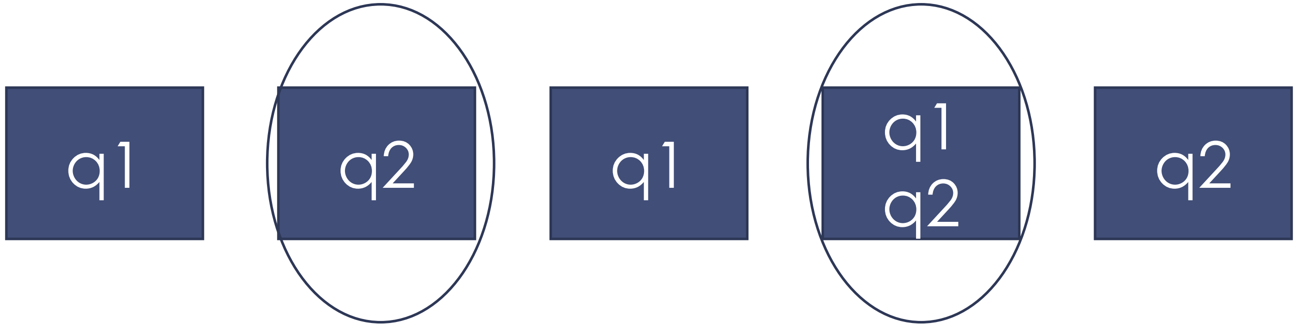
Probability Ranking Principle:

Given a user *query* ***q*** and a *document* ***dj*** in the collection, the probabilistic model tries to estimate the probability that the user will find the document ***dj*** interesting (i. e. relevant).

Probabilistic Retrieval Model

1. Usage patterns to predict relevance [Maron and Kunhs 1960].
2. Usage of each term in the query as **clues** as to whether or not a document is **relevant**. [Robertson and Sparck Jones, 1976].

- Query q(q1, q2)
- Run ***q*** and retrieve top ***n*** documents (***let n=5***) ***di*** represents an arbitrary document



Assume **d2** and **d4** are relevant:

- P(q1 | di is relevant) = ½
- P(q1 | di is not relevant) = 2/3
- P(q2 | di is relevant) = 1
- P(q2 | di is not relevant) = 1/3

Estimating the individual term weights Robertson and Sparck Jones, 1976 INDEPENDENCE ASSUMPTIONS

I1 - The distribution of ***terms*** in ***relevant*** documents is ***independent*** and their distribution in all documents is independent.

I2 - The distribution of ***terms*** in ***relevant*** documents is ***independent*** and their distribution in non-relevant documents is independent.

Estimating the individual term weights

ORDERING PRINCIPLES

O1.- Probable **relevance** is based only on the presence of search terms in the documents.

O2.- Probable relevance is based on both the presence of search terms in documents and their absense from documents.

Four **Weights** are derived **I1, I2, O1, O2**

N = number of documents in the collection
R = number of **relevant** documents for a given query ***q***.
n = number of documents that contain term **t**.
r = number of **relevant** documents that contain term ***t***.

Choosing I1 and O1 yields the following weight

$$\mathbf{w1} = \boldsymbol{log\,\frac{\frac{r}{R}}{\frac{n}{N}}}$$

Choosing I2 and O1 yields the following weight

$$\mathbf{w2} = \boldsymbol{log\frac{\frac{r}{R}}{\frac{n-r}{N-R}}}$$

Choosing **I1** and **O2** yields the following weight

$$\mathbf{w3} = \boldsymbol{log}\,\frac{\frac{\boldsymbol{r}}{\boldsymbol{R-r}}}{\frac{\boldsymbol{n}}{\boldsymbol{N-n}}}$$

Choosing **I2** and **O2** yields the following weight

$$\mathbf{w4} = \boldsymbol{log}\,\frac{\frac{\boldsymbol{r}}{\boldsymbol{R-r}}}{\frac{\boldsymbol{n-r}}{\boldsymbol{(N-n)-(R-r)}}}$$

Weight for incomplete relevant inf.

$$\mathbf{w} = \boldsymbol{log}\,\frac{\frac{\boldsymbol{r}+\mathbf{0.5}}{(\boldsymbol{R}-\boldsymbol{r})+\mathbf{0.5}}}{\frac{(\boldsymbol{n}-\boldsymbol{r})+\mathbf{0.5}}{(\boldsymbol{N}-\boldsymbol{n})-(\boldsymbol{R}-\boldsymbol{r})+\mathbf{0.5}}}$$

Q: “**gold silver truck**”
D1 = “Shipment of **gold** damaged in a fire”
D2 = “Delivery of **silver** arrived in a **silver truck**”
D3 = “Shipment of **gold** arrived in a **truck**”

variable	***gold***	***silver***	***truck***
N	3	3	3
n	2	1	2
R	2	2	2
r	1	1	2

N = number of documents in the collection
R = number of relevant documents for a given query ***q***.
n = number of documents that contain term t.
R = number of relevant documents that contain term ***t***.

Example: Term Weights

term	*w1*	*w2*	*w3*	*w4*
gold	-0.079	-0.176	-0.176	-0.477
silver	0.097	0.301	0.176	.477
truck	0.143	0.523	0.523	1.176

Example: Document Weights

term	*w1*	*w2*	*w3*	*w4*
D1	-0.079	-0.176	-0.176	-0.477
D2	0.240	0.824	0.699	1.653
D3	0.063	0.347	0.347	0.699

Disadvantages

- The need to guess the initial separation of documents into relevant and non-relevant sets.
- The fact that the method does not take into account the frequency
- Lack of length normalization.

Homework

- To study the topics we have seen so far because exam is after 4 lessons.
- To bring questions