

Probabilistic Retrieval

(CS429)
Nazli Goharian
nazli@ir.iit.edu

Slides are *mostly* based on Information Retrieval Algorithms and Heuristics, Grossman, Frieder

© Goharian, Grossman, Frieder, 2002, 2009

1

Probabilistic Model

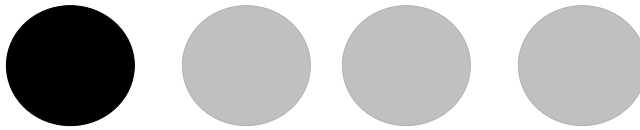
- Use **probability** to estimate the “odds” of relevance of a query to a document.
 - With having information about relevant and non-relevant sets
 - Without having such information
- Original model (binary independence model, BIM) does not consider the document and query term weight.
- An extended version that includes the document and query term weight has influenced search engines.

© Goharian, Grossman, Frieder, 2002, 2009

2

Some Background

- If we have four balls, three gray and one black, and *it is equally likely that we could pick any of the balls*, we can estimate the probability that of:



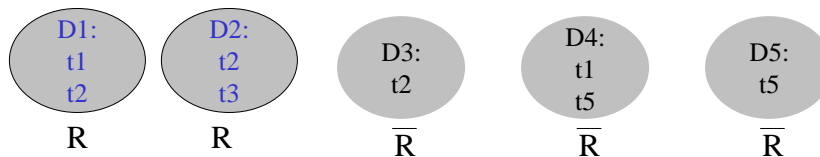
- Choosing a black ball = $1/4$
- Choosing two black balls in a row $(1/4)(1/4) = (1/16)$

© Goharian, Grossman, Frieder, 2002, 2009

3

Relevance Odds for One Term

- We want to estimate, for a given term, the odds of being in a relevant document.



- Assumption: D1 and D2 are relevant; D3, D4 and D5 are non-relevant.
- Need to compute the estimate that a document D_j is relevant given the query term $t1$
- Odds that R is relevant given $t1$:

$$O(R | t1) = \frac{\text{num relevant with } t1 / \text{num relevant}}{\text{num of docs with } t1 / \text{all documents}}$$

$$O(R | t1) = (1 / 2) / (2 / 5) = .5 / .4 = 1.25 : 1$$

© Goharian, Grossman, Frieder, 2002, 2009

4

Computing Odds of Relevance for Multiple Terms

- Given query terms t_1, t_2, \dots, t_n , must compute the odds of relevance given these terms:

$$O(R | t_1, t_2, \dots, t_n)$$

- Based on the Bayes theorem (independence assumption), we can take the product of these individual odds.

$$O(R | t_i) = \prod_{i=1}^{i=t} O(R | t_i)$$

- Note, since the log function is often used to scale the odds, the sum of the log odds (log of each odds) may be used:

$$\log\left(\prod_{i=1}^{i=t} O(R | t_i)\right) = \sum_{i=1}^{i=t} \log(O(R | t_i))$$

© Goharian, Grossman, Frieder, 2002, 2009

5

Principles surrounding 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.
- 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 absence from documents.

© Goharian, Grossman, Frieder, 2002, 2009

6

Parameters in Computing Term Weight

- N = total number of documents in collection
R = total number of relevant documents for a query
n = number of documents that contain the query term
r = number of relevant documents that contain the query term

© Goharian, Grossman, Frieder, 2002, 2009

7

Probabilistic Variations to Compute Term Weight

- **I1 and O1:**
 - **Ratio of** the ratio of relevant documents having the term **to** the ratio of all documents having the term
- **I2 and O1:**
 - **Ratio of** the ratio of relevant docs having the term **to** the ratio of the non-relevant documents having the term

$$\left(\frac{\frac{r}{R}}{\frac{n}{N}} \right)$$

$$\left(\frac{\frac{r}{R}}{\frac{n-r}{N-R}} \right)$$

© Goharian, Grossman, Frieder, 2002, 2009

8

Probabilistic Variations to Compute Term Weight

- I1 and O2:
 - **Ratio of the odds** of a relevant document having the term (i.e., ratio of relevant documents having the term to not having the term) **to the odds** of all documents having the term (i.e., ratio of all documents having the term to not having the term)

$$\left(\frac{\frac{r}{R-r}}{\frac{n}{N-n}} \right)$$

- I2 and O2:
 - **Ratio of the odds** of a relevant document having the term (i.e., ratio of relevant documents having the term to not having the term) **to the odds** of all non-relevant documents having the term (i.e., ratio of all non-relevant documents having the term to not having the term)

$$\left(\frac{\frac{r}{R-r}}{\frac{n-r}{(N-n)-(R-r)}} \right)$$

© Goharian, Grossman, Frieder, 2002, 2009

9

Probabilistic Variations to Compute Term Weight

- To guarantee that the denominator is never zero, adding a minor 0.5 to all numerators and denominators:

$$\left(\frac{\frac{r+0.5}{R-r+0.5}}{\frac{n-r+0.5}{(N-n)-(R-r)+0.5}} \right) \leftarrow \text{Robertson/Spark Jones weight}$$

© Goharian, Grossman, Frieder, 2002, 2009

10

a priori Relevance Information

- *a priori* Relevance Information not always known
- In on-line systems not possible to have relevant information as training data (r, R)
- Alternative:
 - Relying on user's feedback
 - Without any relevance information

© Goharian, Grossman, Frieder, 2002, 2009

11

Probabilistic Retrieval Example

- D1: "Cost of paper is up." (*relevant*)
- D2: "Cost of jellybeans is up." (*not relevant*)
- D3: "Salaries of CEO's are up." (*not relevant*)
- D4: "Paper: CEO's labor cost up." (????)

| Q. Term | Relevant | Not relevant | Evidence |
|----------------|-----------------|---------------------|-----------------|
| paper | 1 | 0 | for (strong) |
| CEO | 0 | 1/2 | against |
| labor | 0 | 0 | none |
| cost | 1 | 1/2 | for (weak) |
| up | 1 | 1 | none |

© Goharian, Grossman, Frieder, 2002, 2009

12

Probabilistic Retrieval Example (Cont'd)

- *cost* appears in 1 of 1 relevant document
 - odds are $(1+.5)/(0+.5) = 3$ to 1 that *cost* will appear
- *cost* appears in 1 of 2 non-relevant documents
 - odds are $(1+.5)/(1+.5) = 1$ to 1 that *cost* will appear
- If *cost* appears in D, then the odds are $(3/1)/(1/1) = 3$ to 1 that D is relevant.

© Goharian, Grossman, Frieder, 2002, 2009

13

Probabilistic Retrieval Example (Cont'd)

- D1: “Cost of paper is up.” (*relevant*)
- D2: “Cost of jellybeans is up.” (*not relevant*)
- D3: “Salaries of CEO’s are up.” (*not relevant*)
- D4: “Paper: CEO’s labor cost up.” (????)

| Term | Odds of Relevance | |
|--|-----------------------|-------------------|
| paper | $(1.5/0.5)/(0.5/2.5)$ | = 15 |
| CEO | $(0.5/1.5)/(1.5/1.5)$ | = 1/3 |
| labor | $(0.5/1.5)/(0.5/2.5)$ | = 5/3 |
| cost | $(1.5/0.5)/(1.5/1.5)$ | = 3 |
| up | $(1.5/0.5)/(2.5/0.5)$ | = 3/5 |
| TOTAL ODDS (product of the individual odds) | | = 15 (RSV) |

© Goharian, Grossman, Frieder, 2002, 2009

14

Modifications to Basic Probabilistic Model

- Term frequency and document length are not considered in original probabilistic model (BIM – Binary Independence Model).
- Performed worse than vector space model (VSM).

Thus:

- Modification to Probabilistic model – a non-binary model:
 - Incorporating tf-idf (Croft and Harper, 1979)
 - Incorporating document length (Robertson and Walker 1995)

A Common Approach: BM25

$$SC(Q, D_i) = \sum_{j=1}^l w \left(\frac{(k_1 + 1)tf_{ij}}{tf_{ij} + k_1 \underbrace{\left(1 - b + b \frac{|D|}{avgdl}\right)}_K} \right) \left(\frac{(k_2 + 1)qtf_i}{k_2 + qtf_i} \right)$$

$$w = idf = \log\left(\frac{N - n + 0.5}{n + 0.5}\right) \quad \leftarrow \text{IDF is used and normally defined as this!}$$

k_1, k_2 and b are parameters to be empirically determined.
 $k_1: 1.2$; k_2 0 to 1000; $b=0.75$ (in many cases)