CIS 833 – Information Retrieval and Text Mining

Lecture 12

Probabilistic Models

October 1, 2015

Credits for slides: Hofmann, Mihalcea, Mobasher, Mooney, Schutze.

Assignments

- HW3 due October 2nd (extended)
- PA1 due October 16th (extended)
- Exam 1 – October 13th

Classes of Retrieval Models

- Boolean models (set theoretic)
  - Extended Boolean
- Vector space models (algebraic)
  - Generalized VS
  - Latent Semantic Indexing
- Probabilistic models
  - Inference Networks
  - Belief Networks

Exact match

Ranking - "Best" match

Required Reading

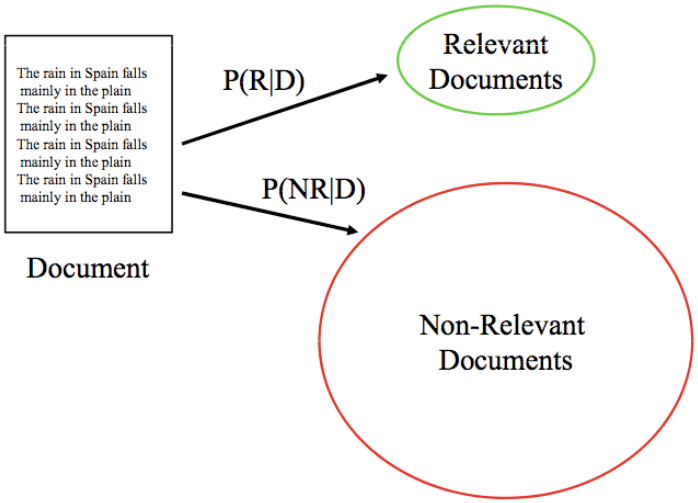
Probabilistic Retrieval Models

- Chapter 11: 11.2-11.4 - Probabilistic retrieval models

Basic Probabilistic Retrieval Model

- Retrieval is modeled as a classification process
- Two classes for each query: the *relevant* and *non-relevant* documents
- Given a particular document D, calculate the probability of belonging to the relevant class, retrieve if greater than probability of belonging to non-relevant class
  - i.e., retrieve if P(R|D) > P(NR|D)
- Equivalently, rank by *likelihood ratio* P(D|R)÷P(D|NR)
- Different ways of estimating these probabilities lead to different probabilistic models

Basic Probabilistic Model



Probabilistic Ranking

Basic concept:

"For a given query, if we know some documents that are relevant, terms that occur in those documents should be given greater weighting in searching for other relevant documents.

By making assumptions about the distribution of terms and applying Bayes Theorem, it is possible to derive weights theoretically."

Van Rijsbergen

Binary Independence Model

- Traditionally used in conjunction with PRP
- **"Binary" = Boolean**: documents are represented as binary incidence vectors of terms:
  - $\vec{x} = (x_1, \ldots, x_n)$
  - $x_i = 1$ iff term *i* is present in document *d* having representation *x*.
- **"Independence":** terms occur in documents independently
  - Different documents can be modeled as same vector
- Bernoulli Naive Bayes model (cf. text categorization!)

Binary Independence Model

- Queries: binary term incidence vectors
- Given query ***q***,
  - for each document ***d*** need to compute ***p(R|q,d).***
  - replace with computing $p(R \mid q,\vec{x})$ where $\vec{x}$ is binary term incidence vector representing ***d***
  - interested in ranking
- Will use odds and Bayes' Rule:

$$O(R \mid q,\vec{x}) = \frac{p(R \mid q,\vec{x})}{p(NR \mid q,\vec{x})} = \frac{\frac{p(R \mid q)p(\vec{x} \mid R,q)}{p(\vec{x} \mid q)}}{\frac{p(NR \mid q)p(\vec{x} \mid NR,q)}{p(\vec{x} \mid q)}}$$

Binary Independence Model

$$O(R \mid q,\vec{x}) = \frac{p(R \mid q,\vec{x})}{p(NR \mid q,\vec{x})} = \frac{p(R \mid q)}{p(NR \mid q)} \cdot \frac{p(\vec{x} \mid R,q)}{p(\vec{x} \mid NR,q)}$$

Constant for a given query

Needs estimation

Using **Independence** Assumption:

$$\frac{p(\vec{x} \mid R,q)}{p(\vec{x} \mid NR,q)} = \prod_{i=1}^{n} \frac{p(x_i \mid R,q)}{p(x_i \mid NR,q)}$$

So : $$O(R \mid q,\vec{x}) = O(R \mid q) \cdot \prod_{i=1}^{n} \frac{p(x_i \mid R,q)}{p(x_i \mid NR,q)}$$

Binary Independence Model

$$O(R \mid q,\vec{x}) = O(R \mid q) \cdot \prod_{i=1}^{n} \frac{p(x_i \mid R,q)}{p(x_i \mid NR,q)}$$

- Since x_i is either 0 or 1:

$$O(R \mid q,\vec{x}) = O(R \mid q) \cdot \prod_{x_i=1} \frac{p(x_i = 1 \mid R,q)}{p(x_i = 1 \mid NR,q)} \cdot \prod_{x_i=0} \frac{p(x_i = 0 \mid R,q)}{p(x_i = 0 \mid NR,q)}$$

- Let $p_i = p(x_i = 1 \mid R,q); \quad r_i = p(x_i = 1 \mid NR,q);$
- Assume, for all terms not occurring in the query (q_i=0) $p_i = r_i$

Then...

Binary Independence Model

$$O(R \mid q,\vec{x}) = O(R \mid q) \cdot \prod_{x_i=q_i=1} \frac{p_i}{r_i} \cdot \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-r_i}$$

All matching terms

Non-matching query terms

$$= O(R \mid q) \cdot \prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)} \cdot \prod_{q_i=1} \frac{1-p_i}{1-r_i}$$

All matching terms

All query terms

Binary Independence Model

$$O(R \mid q, \vec{x}) = O(R \mid q) \cdot \prod_{x_i = q_i = 1} \frac{p_i(1-r_i)}{r_i(1-p_i)} \cdot \prod_{q_i = 1} \frac{1-p_i}{1-r_i}$$

Constant for each query

Only quantity to be estimated for rankings

- Retrieval Status Value:

$$RSV = \log \prod_{x_i = q_i = 1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i = q_i = 1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

Binary Independence Model

- All boils down to computing RSV.

$$RSV = \log \prod_{x_i = q_i = 1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i = q_i = 1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

$$RSV = \sum_{x_i = q_i = 1} c_i; \quad c_i = \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

So, how do we compute c_i's from our data ?

Remember: $p_i = p(x_i = 1 \mid R, q); \quad r_i = p(x_i = 1 \mid NR, q);$

Binary Independence Model

- Estimating RSV coefficients.
- For each term i look at this table of document counts:

Documens	Relevant	Non-Relevant	Total
$x_i=1$	s	$n-s$	n
$x_i=0$	$S-s$	$N-n-S+s$	$N-n$
Total	S	$N-S$	N

- Estimates: $p_i \approx \frac{s}{S} \quad r_i \approx \frac{(n-s)}{(N-S)}$

$$c_i \approx K(N,n,S,s) = \log \frac{s/(S-s)}{(n-s)/(N-n-S+s)}$$

For now, assume no zero terms.

	T1	T2	T3	T4	T5	T6	Relevance
D1	1	0	1	1	0	0	R
D2	0	1	0	1	0	1	R
D3	1	0	1	1	1	0	NR
D4	0	1	1	0	1	1	NR
D5	1	1	0	1	0	0	NR
D6	1	1	0	1	1	1	NR
D7	0	0	0	0	0	1	R
D8	0	0	1	1	1	0	NR
D9	1	1	1	0	1	1	R
D10	1	0	0	1	1	0	NR

Suppose the relevance judgments specified above represent some past user judgments on the relevance of these documents wrt a given query. Using the basic probabilistic retrieval model, compute the discriminant (i.e., P(R|D) vs P(NR|D)) for each of the two new documents

D11 = (0,1,1,0,0,1)
D12 = (1,0,1,1,0,1)

with respect to the query Q = (1,1,0,1,0,1). Based on this discriminant, should these documents be retrieved? Explain your answer.