Probabilistic Information Retrieval
Part II: In Depth

Alexander Dekhtyar

Department of Computer Science

University of Maryland

In this part

- ✔Probability Ranking Principle
  - simple case
  - case with retrieval costs
- ✔Binary Independence Retrieval (BIR)
  - Estimating the probabilities
- ✔Binary Independence Indexing (BII)
  - dual to BIR

The Basics

- ✔ Bayesian probability formulas

$$p(a \mid b)\, p(b) = p(a \cap b) = p(b \mid a)\, p(a)$$

$$p(a \mid b) = \frac{p(b \mid a)\, p(a)}{p(b)}$$

$$p(\overline{a} \mid b)\, p(b) = p(b \mid \overline{a})\, p(\overline{a})$$

- ✔ Odds: $O(y) = \dfrac{p(y)}{p(\overline{y})} = \dfrac{p(y)}{1 - p(y)}$

The Basics

- Document Relevance:

$$p(R \mid x) = \frac{p(x \mid R)\, p(R)}{p(x)}$$

$$p(NR \mid x) = \frac{p(x \mid NR)\, p(NR)}{p(x)}$$

- Note:

$$p(R \mid x) + p(NR \mid x) = 1$$

Probability Ranking Principle

- ✔ Simple case: no selection costs.
- ✔ ***x*** is **relevant** iff $p(R|x) > p(NR|x)$
- ✔ *(Bayes' Decision Rule)*
- ✔ PRP in action: Rank all documents by $p(R|x)$.

Probability Ranking Principle

- More complex case: retrieval costs.
  - *C* - cost of retrieval of relevant document
  - *C'* - cost of retrieval of non-relevant document
  - let *d,* be a document
- Probability Ranking Principle: if

$$C \cdot p(R|d) + C' \cdot (1 - p(R|d)) \leq C \cdot p(R|d') + C' \cdot (1 - p(R|d'))$$

for all *d' not yet retrieved*, then *d* **is the next document to be retrieved**

Next: Binary Independence Model

Binary Independence Model

- ✔ Traditionally used in conjunction with PRP
- ✔ **"Binary" = Boolean**: documents are represented as binary vectors of terms:
  - $\vec{x} = (x_1, \ldots, x_n)$
  - $x_i = 1$ iff term i is present in document x.
- ✔ **"Independence":** terms occur in documents independently
- ✔ Different documents can be modeled as same vector.

Binary Independence Model

- ✔ Queries: binary vectors of terms
- ✔ Given query ***q***,
  - for each document ***d*** need to compute ***p(R|q,d).***
  - replace with computing ***p(R|q,x)*** where ***x*** is vector representing ***d***
- ✔ Interested only in ranking
- ✔ Will use odds:

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

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

Binary Independence Model

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

- Since x_i is either 0 or 1:

$$O(R|q,d) = O(R|q) \cdot \prod_{x_i=1} \frac{p(x_i=1|R,q)}{p(x_i=1|NR,q)} \cdot \prod_{x_i=0} \frac{p(x_i=0|R,q)}{p(x_i=0|NR,q)}$$

- Let $p_i = p(x_i=1|R,q)$; $r_i = p(x_i=1|NR,q)$;

- Assume, for all terms not occuring 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

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

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.

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

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

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

Binary Independence Model

- Estimating RSV coefficients.
- For each term *i* look at the following table:

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

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

Add 0.5 to every expression

PRP and BIR: The lessons

- ✔ Getting reasonable approximations of probabilities is possible.
- ✔ Simple methods work only with restrictive assumptions:
  - – ***term independence***
  - – ***terms not in query do not affect the outcome***
  - – ***boolean representation of documents/queries***
  - – ***document relevance values are independent***
- ✔ Some of these assumptions can be removed

Next: Binary Independence Indexing

Binary Independence Indexing vs. Binary Independence Retrieval

•BIR

- ✔ Many Documents, One Query
- ✔ Bayesian Probability:

$$p(R|\vec{q},\vec{x}) = \frac{p(\vec{x}|\vec{q},R)\,p(R|\vec{q})}{p(\vec{x}|\vec{q})}$$

- ✔ Varies: document representation
- ✔ Constant: query (representation)

•BII

- ✔ One Document, Many Queries
- ✔ Bayesian Probability

$$p(R|\vec{q},\vec{x}) = \frac{p(\vec{q}|\vec{x},R)\,p(R|\vec{x})}{p(\vec{q}|\vec{x})}$$

- ✔ Varies: query
- ✔ Constant: document

Binary Independence Indexing

- ✔ "Learnng" from queries
  - – More queries: better results

$$p(R|\vec{q},\vec{x}) = \frac{p(\vec{q}|\vec{x},R)\,p(R|\vec{x})}{p(\vec{q}|\vec{x})}$$

- ✔ ***p(q|x,R)*** *- probability that if document* ***x*** *had been deemed relevant, query* ***q*** *had been asked*
- ✔ The rest of the framework is similar to BIR

Binary Independence Indexing: Key Assumptions

- ✔ Term occurrence in queries is ***conditionally independent:*** $p(\vec{q} \mid R, \vec{x}) = \prod_{i=1}^{n} p(q_i \mid R, \vec{x})$
- ✔ *Relevance* of document representation ***x*** w.r.t. query ***q*** *depends only* on the terms present in the query (***q_i=1***)
- ✔ For each term ***i*** not used in representation ***x*** of document ***d*** (***x_i=0***): $p(R \mid z_i, \vec{x}) = p(R \mid \vec{x})$
  - *only positive occurrences of terms count*

Binary Independence Indexing

$$p(R|\vec{q},\vec{x}) = \frac{p(\vec{q}|\vec{x},R)p(R|\vec{x})}{p(\vec{q}|\vec{x})}$$

$$= \frac{\prod_i p(q_i)}{p(\vec{q})} \cdot \prod_{i=1}^{n} \frac{p(R|q_i,\vec{x})}{p(R|\vec{x})}$$

$$= \frac{\prod_i p(q_i)}{p(\vec{q})} \cdot p(R|\vec{x}) \cdot \prod_{q_i=1} \frac{p(R|q_i=1,\vec{x})}{p(R|\vec{x})} \cdot \prod_{q_i=0} \frac{p(R|q_i=0,\vec{x})}{p(R|\vec{x})}$$

constant

Equal to 1