Introduction to Information Retrieval

Probabilistic Information Retrieval

Christopher Manning and Pandu Nayak
FROM BOOLEAN TO RANKED RETRIEVAL ... IN TWO STEPS
Ranked retrieval

- Thus far, our queries have all been Boolean
  - Documents either match or don’t
- Can be good for expert users with precise understanding of their needs and the collection
  - Can also be good for applications: Applications can easily consume 1000s of results
- Not good for the majority of users
  - Most users incapable of writing Boolean queries
    - Or they are, but they think it’s too much work
  - Most users don’t want to wade through 1000s of results.
    - This is particularly true of web search
Problem with Boolean search: feast or famine

- Boolean queries often result in either too few (=0) or too many (1000s) results.
- Query 1: “standard user dlink 650” → 200,000 hits
- Query 2: “standard user dlink 650 no card found”: 0 hits
- It takes a lot of skill to come up with a query that produces a manageable number of hits.
  - AND gives too few; OR gives too many
Who are these people?

Karen Spärck Jones

Stephen Robertson

Keith van Rijsbergen
Why probabilities in IR?

In traditional IR systems, matching between each document and query is attempted in a semantically imprecise space of index terms.

Probabilities provide a principled foundation for uncertain reasoning.

Can we use probabilities to quantify our uncertainties?
Probabilistic IR topics

1. Classical probabilistic retrieval model
   - Probability ranking principle, etc.
   - Binary independence model (≈ Naïve Bayes text cat)
   - (Okapi) BM25

2. Bayesian networks for text retrieval

3. Language model approach to IR
   - An important emphasis in recent work

Probabilistic methods are one of the oldest but also one of the currently hot topics in IR
   - Traditionally: neat ideas, but didn’t win on performance
   - It seems to be different now
The document ranking problem

- We have a collection of documents
- User issues a query
- A list of documents needs to be returned
- Ranking method is the core of modern IR systems:
  - In what order do we present documents to the user?
  - We want the “best” document to be first, second best second, etc....
- Idea: Rank by probability of relevance of the document w.r.t. information need
  - \( P(R=1 \mid \text{document}_i, \text{query}) \)
Recall a few probability basics

- For events $A$ and $B$:

$$p(A, B) = p(A \cap B) = p(A \mid B)p(B) = p(B \mid A)p(A)$$

- Bayes’ Rule

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} = \frac{p(B \mid A)p(A)}{\sum_{X=A,A} p(B \mid X)p(X)}$$

- Odds:

$$O(A) = \frac{p(A)}{p(\overline{A})} = \frac{p(A)}{1 - p(A)}$$
The Probability Ranking Principle

“If a reference retrieval system’s response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.”

Probability Ranking Principle

Let \( x \) represent a document in the collection. Let \( R \) represent relevance of a document w.r.t. given (fixed) query and let \( R=1 \) represent relevant and \( R=0 \) not relevant.

Need to find \( p(R=1|x) \) – probability that a document \( x \) is relevant.

\[
p(R = 1 \mid x) = \frac{p(x \mid R = 1)p(R = 1)}{p(x)}
\]

\[
p(R = 0 \mid x) = \frac{p(x \mid R = 0)p(R = 0)}{p(x)}
\]

\[
p(R = 0 \mid x) + p(R = 1 \mid x) = 1
\]
Probability Ranking Principle (PRP)

- Simple case: no selection costs or other utility concerns that would differentially weight errors

- PRP in action: Rank all documents by \( p(R=1 \mid x) \)

- Theorem: Using the PRP is optimal, in that it minimizes the loss (Bayes risk) under 1/0 loss
  - Provable if all probabilities correct, etc.  [e.g., Ripley 1996]
Probability Ranking Principle

- More complex case: retrieval costs.
  - Let $d$ be a document
  - $C$ – cost of not retrieving a relevant document
  - $C'$ – cost of retrieving a non-relevant document

- Probability Ranking Principle: if
  \[ C' \cdot p(R = 0 | d) - C \cdot p(R = 1 | d) \leq C' \cdot p(R = 0 | d') - C \cdot p(R = 1 | d') \]

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

- We won’t further consider cost/utility from now on
Probability Ranking Principle

- How do we compute all those probabilities?
  - Do not know exact probabilities, have to use estimates
  - Binary Independence Model (BIM) – which we discuss next – is the simplest model

- Questionable assumptions
  - “Relevance” of each document is independent of relevance of other documents.
    - Really, it’s bad to keep on returning duplicates
  - Boolean model of relevance
  - That one has a single step information need
    - Seeing a range of results might let user refine query
Probabilistic Retrieval Strategy

- Estimate how terms contribute to relevance
  - How do other things like term frequency and document length influence your judgments about document relevance?
    - Not at all in BIM
    - A more nuanced answer is the Okapi (BM25) formulae [next time]
      - Spärck Jones / Robertson

- Combine to find document relevance probability

- Order documents by decreasing probability
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 (cf. IIR Chapter 1):
  - \( \vec{x} = (x_1, \ldots, x_n) \)
  - \( x_i = 1 \) iff term \( i \) is present in document \( x \).
- "Independence": terms occur in documents independently
- Different documents can be modeled as the same vector
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 \( p(R|q,x) \) where \( x \) is binary term incidence vector representing \( d \).
- Interested only in ranking
- Will use odds and Bayes’ Rule:

\[
O(R|q,\bar{x}) = \frac{p(R = 1|q,\bar{x})}{p(R = 0|q,\bar{x})} = \frac{\frac{p(R = 1|q)p(\bar{x} | R = 1, q)}{p(\bar{x} | q)} \frac{p(\bar{x} | q)}{p(R = 0|q)p(\bar{x} | R = 0, q)}}{\frac{p(R = 0|q)p(\bar{x} | R = 0, q)}{p(\bar{x} | q)}}
\]
**Binary Independence Model**

\[ O(R \mid q, \bar{x}) = \frac{p(R = 1 \mid q, \bar{x})}{p(R = 0 \mid q, \bar{x})} = \frac{p(R = 1 \mid q)}{p(R = 0 \mid q)} \cdot \frac{p(\bar{x} \mid R = 1, q)}{p(\bar{x} \mid R = 0, q)} \]

Constant for a given query

• Using **Independence** Assumption:

\[ \frac{p(\bar{x} \mid R = 1, q)}{p(\bar{x} \mid R = 0, q)} = \prod_{i=1}^{n} \frac{p(x_i \mid R = 1, q)}{p(x_i \mid R = 0, q)} \]

\[ O(R \mid q, \bar{x}) = O(R \mid q) \cdot \prod_{i=1}^{n} \frac{p(x_i \mid R = 1, q)}{p(x_i \mid R = 0, q)} \]

Needs estimation
Binary Independence Model

\[ O(R \mid q, \vec{x}) = O(R \mid q) \cdot \prod_{i=1}^{n} \frac{p(x_i \mid R = 1, q)}{p(x_i \mid R = 0, q)} \]

- Since \( x_i \) is either 0 or 1:

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

- Let \( p_i = p(x_i = 1 \mid R = 1, q) \); \( r_i = p(x_i = 1 \mid R = 0, q) \);

- Assume, for all terms not occurring in the query \( q_i = 0 \) \( p_i = r_i \)

\[ O(R \mid q, \vec{x}) = O(R \mid q) \cdot \prod_{x_i=1} \frac{p_i}{r_i} \cdot \prod_{x_i=0} \frac{1 - p_i}{1 - r_i} \]
<table>
<thead>
<tr>
<th>document</th>
<th>relevant (R=1)</th>
<th>not relevant (R=0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>term present</td>
<td>$x_i = 1$</td>
<td>$p_i$</td>
</tr>
<tr>
<td>term absent</td>
<td>$x_i = 0$</td>
<td>$(1 - p_i)$</td>
</tr>
</tbody>
</table>
Binary Independence Model

\[ O(R \mid q, \tilde{x}) = O(R \mid q) \cdot \prod_{x_i=q_i=1} p_i \cdot \prod_{x_i=0 \quad q_i=1} \frac{1-p_i}{1-r_i} \]

All matching terms

\[ O(R \mid q, \tilde{x}) = O(R \mid q) \cdot \prod_{x_i=q_i=1} p_i \cdot \prod_{x_i=1 \quad q_i=1} \left( \frac{1-r_i}{1-p_i} \cdot \frac{1-p_i}{1-r_i} \right) \prod_{x_i=0 \quad q_i=1} \frac{1-p_i}{1-r_i} \]

Non-matching query terms

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

All matching terms

All query terms
Binary Independence Model

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

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.

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

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

The \( c_i \) are log odds ratios
They function as the term weights in this model

So, how do we compute \( c_i \)'s from our data?
Binary Independence Model

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

<table>
<thead>
<tr>
<th>Documents</th>
<th>Relevant</th>
<th>Non-Related</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i = 1$</td>
<td>$s$</td>
<td>$n-s$</td>
<td>$n$</td>
</tr>
<tr>
<td>$x_i = 0$</td>
<td>$S-s$</td>
<td>$N-n-S+s$</td>
<td>$N-n$</td>
</tr>
<tr>
<td>Total</td>
<td>$S$</td>
<td>$N-S$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Estimates:

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

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

For now, assume no zero terms. Remember smoothing.
Estimation – key challenge

- If non-relevant documents are approximated by the whole collection, then $r_i$ (prob. of occurrence in non-relevant documents for query) is $n/N$ and

$$
\log \frac{1 - r_i}{r_i} = \log \frac{N - n - S + s}{n - s} \approx \log \frac{N - n}{n} \approx \log \frac{N}{n} = \text{IDF}!
$$
Collection vs. Document frequency

- Collection frequency of $t$ is the total number of occurrences of $t$ in the collection (incl. multiples)
- Document frequency is number of docs $t$ is in
- Example:

<table>
<thead>
<tr>
<th>Word</th>
<th>Collection frequency</th>
<th>Document frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>insurance</td>
<td>10440</td>
<td>3997</td>
</tr>
<tr>
<td>try</td>
<td>10422</td>
<td>8760</td>
</tr>
</tbody>
</table>

- Which word is a better search term (and should get a higher weight)?
Estimation – key challenge

- $p_i$ (probability of occurrence in relevant documents) cannot be approximated as easily
- $p_i$ can be estimated in various ways:
  - from relevant documents if you know some
    - Relevance weighting can be used in a feedback loop
  - constant (Croft and Harper combination match) – then just get idf weighting of terms (with $p_i=0.5$)
    \[
    RSV = \sum_{x_i=q_i=1} \log \frac{N}{n_i}
    \]
  - proportional to prob. of occurrence in collection
    - Greiff (SIGIR 1998) argues for $1/3 + 2/3 \frac{df_i}{N}$
Probabilistic Relevance Feedback

1. Guess a preliminary probabilistic description of $R=1$ documents; use it to retrieve a set of documents
2. Interact with the user to refine the description: learn some definite members with $R = 1$ and $R = 0$
3. Re-estimate $p_i$ and $r_i$ on the basis of these
   - If $i$ appears in $V_i$ within set of documents $V$: $p_i = |V_i|/|V|$
   - Or can combine new information with original guess (use Bayesian prior):
     \[ p_i^{(2)} = \frac{|V_i| + \kappa p_i^{(1)}}{|V| + \kappa} \]
     \( \kappa \) is prior weight
4. Repeat, thus generating a succession of approximations to relevant documents
Iteratively estimating $p_i$ and $r_i$ (= Pseudo-relevance feedback)

1. Assume that $p_i$ is constant over all $x_i$ in query and $r_i$ as before
   - $p_i = 0.5$ (even odds) for any given doc

2. Determine guess of relevant document set:
   - $V$ is fixed size set of highest ranked documents on this model

3. We need to improve our guesses for $p_i$ and $r_i$, so
   - Use distribution of $x_i$ in docs in $V$. Let $V_i$ be set of documents containing $x_i$
     - $p_i = |V_i| / |V|$ 
     - Assume if not retrieved then not relevant
     - $r_i = (n_i - |V_i|) / (N - |V|)$

4. Go to 2. until converges then return ranking
PRP and BIM

- Getting reasonable approximations of probabilities is possible.
- Requires restrictive assumptions:
  - Term independence
  - Terms not in query don’t affect the outcome
  - Boolean representation of documents/queries/relevance
  - Document relevance values are independent
- Some of these assumptions can be removed
- Problem: either require partial relevance information or seemingly only can derive somewhat inferior term weights
Removing term independence

- In general, index terms aren’t independent
- Dependencies can be complex
- van Rijsbergen (1979) proposed simple model of dependencies as a tree
  - Exactly Friedman and Goldszmidt’s Tree Augmented Naive Bayes (AAAI 13, 1996)
- Each term dependent on one other
- In 1970s, estimation problems held back success of this model
Second step: Term frequency

- Right in the first lecture, we said that a page should rank higher if it mentions a word more
  - Perhaps modulated by things like page length

- We might want a model with term frequency in it.

- We’ll see a probabilistic one next time – BM25

- Quick summary of vector space model
Summary – vector space ranking

- Represent the query as a weighted term frequency/inverse document frequency (tf-idf) vector
- Represent each document as a weighted tf-idf vector
- Compute the cosine similarity score for the query vector and each document vector
- Rank documents with respect to the query by score
- Return the top $K$ (e.g., $K = 10$) to the user
Introduction to Information Retrieval
Cosine similarity
tf-idf weighting has many variants

<table>
<thead>
<tr>
<th>Term frequency</th>
<th>Document frequency</th>
<th>Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (natural)</td>
<td>tf&lt;sub&gt;t,d&lt;/sub&gt;</td>
<td>n (no) 1</td>
</tr>
<tr>
<td>l (logarithm)</td>
<td>1 + log(tf&lt;sub&gt;t,d&lt;/sub&gt;)</td>
<td>t (idf) log N/df&lt;sub&gt;t&lt;/sub&gt;</td>
</tr>
<tr>
<td>a (augmented)</td>
<td>0.5 + \frac{0.5 \times tf&lt;sub&gt;t,d&lt;/sub&gt;}{\max_t(tf&lt;sub&gt;t,d&lt;/sub)})</td>
<td>p (prob idf) max{0, log N−df&lt;sub&gt;t&lt;/sub&gt;/df&lt;sub&gt;t&lt;/sub&gt;}</td>
</tr>
<tr>
<td>b (boolean)</td>
<td>\begin{cases} 1 &amp; \text{if } tf&lt;sub&gt;t,d&lt;/sub&gt; &gt; 0 \ 0 &amp; \text{otherwise} \end{cases}</td>
<td>b (byte size) 1/CharLength&lt;sup&gt;\alpha&lt;/sup&gt;, ( \alpha &lt; 1 )</td>
</tr>
<tr>
<td>L (log ave)</td>
<td>\frac{1 + \log(tf&lt;sub&gt;t,d&lt;/sub&gt;)}{1 + \log(\text{ave}_{t \in d}(tf&lt;sub&gt;t,d&lt;/sub)))}</td>
<td></td>
</tr>
</tbody>
</table>
Resources


http://www.dcs.gla.ac.uk/Keith/Preface.html
