Text Technologies

Probabilistic Model of IR
Victor Lavrenko

Probability Ranking Principle
- Robertson (1977)
  - "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 these data.
- Basis for most probabilistic approaches to IR

PRP = best possible ranking
- Let  \( D = \{ \text{Document}, \text{Query}, \text{User}, \text{Task}, \text{Context}, \ldots \} \)
- Rank documents by \( p_r = P(R=1|D) \)
  - \( p_r \) should be highest
  - expected precision at rank \( r \)
  - expected recall at rank \( r \)
  - \( P \) at rank, average precision, etc.
- Ranking version of Bayes error rate
  - best possible classification rate:
    - relevant if \( P(D,R=1) > P(D,R=0) \)

Optimality of PRP
- Retrieving a set of documents:
  - PRP equivalent to Bayes error criterion
  - optimal w.r.t. classification error
- Ranking a set of documents:
  - optimal w.r.t. precision / recall at a given rank average precision, etc.
- Need to estimate \( P(\text{relevant} | \text{document}, \text{query}) \)
  - many different attempts to do that
  - Classical Probabilistic Model (Robertson, Sparck-Jones)
    - also known as Binary Independence model, Okapi model
    - very influential, successful (BM25 ranking formula)

Let's dissect the PRP
- rank documents ... by probability of relevance
  - \( P(\text{relevant} | \text{document}) \) ...
  - \( P(\text{relevant} | \text{document}, \text{query}, \text{session}, \text{user}, \text{context}, \text{task}) \)
  - estimated as accurately as possible
  - \( P_{\text{rec}} \) (relevant | document) \( \Rightarrow P_{\text{rec}} \) (relevant | document, query)
  - based on whatever data is available to system
- best possible accuracy one can achieve with that data
  - recipe for a perfect IR system: just need \( P_{\text{rec}}(\text{relevant} | \ldots) \)
  - strong stuff, can this really be true?

Classical probabilistic model
- Assumption A0:
  - relevance of \( D \) doesn't depend on any other document
    - made by almost every retrieval model (exception: cluster-based)
- Rank documents by \( P(R=1|D) \)
  - \( R \sim \{0,1\} \)
  - Bernoulli RV indicating relevance
  - \( D \) represents content of the document
- Rank-equivalent:
  - \( P(R=1|D) = \frac{P(R=1|D) \cdot P(D | R=1) \cdot P(D)}{P(D)} \)
  - \( P(R=0|D) = \frac{P(R=0|D) \cdot P(D | R=0) \cdot P(D)}{P(D)} \)

Motivation
- Vector-space is very heuristic in nature
  - why does it work? no notion of relevance anywhere
  - any weighting scheme, similarity measure can be used
  - components not interpretable - no guide for what to try next
  - encourages ad-hoc engineering (tweak, rank, observe, tweak)
  - very popular, hard to beat, good baseline
- easy to assemble good ideas from other models
- Probabilistic Model of Retrieval
  - mathematical formalism for relevant / non-relevant sets
    - explicitly define random variables (R, D, Q)
  - be specific about what their values are
  - state the assumptions behind every step
  - watch out for contradictions

Probability of relevance
- What is \( P_{\text{rec}} \) (relevant | doc, qry, session, user, context, task)?
  - isn't relevance just the user's opinion?
  - user decides relevant or not, what is this "probability" thing?
- Search algorithm cannot look into your head (yet)
  - relevance depends on features that algorithm cannot observe
  - different users may disagree on relevance of same doc
    - even similar users, doing the same task, in the same context
- \( P_{\text{rec}} \) (relevant | Q, D)
  - proportion of all unseen users / contexts / tasks for which \( D \) would have been judged relevant to \( Q \)
- Analogy: \( P(D \in S | \text{even and not square}) \)
**Probabilistic model: assumptions**

- Want \( P(D|R=1) \) and \( P(D|R=0) \)

- **Assumptions:**
  - **A1:** \( D = \{D_1, ..., D_n\} \) - one RF for every word \( w \)
    - Bernoulli: values 0,1 (word either present or absent in a document)
  - **A2:** \( D_w \) are mutually independent given \( R \)
    - Blatantly false: presence of "Barack" tells you nothing about "Obama"
    - But must assume something: \( D \) represents subsets of vocabulary (without assumptions: \( 2^{|\text{all elements}|}\) possible events
    - Allows us to write:
    \[
    P(R=1|D) = \frac{P(D|R=1)}{P(D|R=0)} \prod_{w} P(D_w|R=R) \prod_{w} P(D_w|R=R) 
    \]
    - Note: identical to the Naive Bayes classifier
    - With equal priors

**Example (with relevance)**

- Relevant docs: \( D_1 = \{a, b, c, d\} \), \( D_2 = \{a, b, e\} \)
- Non-relevant: \( D_3 = \{b, c, d\} \), \( D_4 = \{b, e\} \), \( D_5 = \{a, b, c, e\} \)
- Word: \( a \), \( b \), \( c \), \( d \), \( e \)
- \( N_w \) (w): \( \frac{2}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2} \)
- \( N_w(D_1) \): \( \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2} \)
- New document: \( D_6 = \{c, b, g\} \)

**Estimation (without R)**

- Assumption A4: \( p_w = q_w \) for \( w \in Q \)
  - If the word is not in the query, it is equally likely to occur in relevant and non-relevant populations
  - Practical reason: restrict product to query - document overlap

- Assumption A5: \( p_w = 0.5 \) for \( w \in Q \)
  - Assumption: \( p_w \) and \( q_w \) cancel out

- Assumption A6: \( q_w = N_{w} / N \) for \( w \in Q \)
  - Non-relevant set approximated by collection as a whole
  - Very reasonable: most documents are non-relevant

**Modeling word dependence**

- Classical model assumes all words independent
  - Blatantly false, made by almost all retrieval models
  - The most widely criticized assumption behind IR models
  - Should be able to do better, right?

- Tree dependence model (van Rijsbergen, 1977)
  - Structure dependencies as maximum spanning tree
  - Each word depends on its parent (and R)
  - Total parameters: twice that of BIR

**Do dependence models work?**

- Many similar attempts since the original
  - Dozens published results, probably hundreds of attempts
  - Many different ways to model dependence
  - Never consistent improvement: always "promising results"

- Why? It works in other fields.
  - Independence (unigram) would be a silly choice for ASR, MT
  - Need to handle subtle form of the string (is output grammatical?)
  - In IR we are already dealing with well-formed strings
  - Pointless to waste probability mass on grammatical
  - BIR doesn't really assume independence
  - Necessary condition significantly weaker than independence
BIR doesn't assume independence

\[ P_{\text{BIR}}(d) = \prod_{w} P(d, w) = \prod_{w} \frac{P(d, w)}{P(d)} \cdot \frac{k(w)}{k(w)} \]

- *Independence* will not affect ranking if all words w have the same dependence in relevant/relevant classes
- *1st order dependence* in the non-relevant class

Sufficient condition: proportional independence

the total amount of independence among all words in a document is approximately the same under R=1 and R=0

How to interpret independence

- **(conditional) Independence:**
  - seeing "Obama" doesn't affect chances of seeing "Romney"
  - holds for R=1 and R=0 (probabilities can be different)
- **Linked Dependence:**
  - seeing "Obama" increases chances of seeing "Romney"
  - by the same amount under R=1 and R=0
  - reasonable, unless topic is 2012 elections
- **Proportional Interdependence:**
  - seeing "Obama" increases chances of seeing "Romney"
  - can be more so-depended in the relevant class
  - as long as offset by other word sets under R=0
  - e.g., "world" and "top" more co-dependent in non-relevant class

Sufficient condition: proportional interdependence

Two-Poisson model [Harter]

- Idea: words generated by a mixture of two Poissons
  - "elite" words for a document: occur unusually frequently
  - "non-elite" words - occur as expected by chance
- document is a mixture: \[ P(d,w) = \alpha P_{\text{elite}}(w) + (1-\alpha) P_{\text{non-elite}}(w) \]
- estimate \( \alpha \) by fitting to data (max. likelihood)

Two-Poisson model [Harter]

- Problem: need probabilities conditioned on relevance
  - "eliteness" not the same as relevance
  - Robertson and Sparck Jones: condition eliteness on R=0, R=1
  - final form has too many parameters, and no data to fit them...

- **BM25:** an "approximation" to conditioned 2-Poisson

BM25: an intuitive view

Example: BM25

- documents: \( D_1 = \{a b c d, 2\}, D_2 = \{b e f d, 2\}, D_3 = \{b g c d, 3\}, D_4 = \{b g h h, 3\} \)
- query: Q = \"a c h\"; assume k = 1, b = 0.5

Summary: probabilistic model

- **Probability Ranking Principle**
  - ranking by \( P(R=1|D) \) is optimal
- **Classical probabilistic model**
  - words: binary events (relaxed in the 2-Poisson model)
  - words assumed independent (not accurate)
  - numerous attempts to model dependence, most without success
- **Formal, interpretable model**
  - explicit, elegant model of relevance (if observable)
  - very problematic if relevance not observable
  - authors resort to heuristics, devised BM25