

Text Technologies

Probabilistic Model of IR

Victor Lavrenko

Copyright © 2011 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 those data.”
- Basis for most probabilistic approaches to IR

Copyright © 2011 Victor Lavrenko

Motivation

- Vector-space is very heuristic in nature
 - no notion of relevance (= similarity?)
 - any weighting scheme, similarity measure can be used
 - components not interpretable → no guide for what to try next
 - encourages ad-hoc engineering: tweak, run, observe, tweak
 - very popular, hard to beat, good baseline
 - easy to assimilate 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

Copyright © 2011 Victor Lavrenko

Let's dissect the PRP

- rank documents ... by probability of relevance
 - $P(\text{relevant} | \text{document})$
- estimated as accurately as possible
 - $P_{\text{est}}(\text{relevant} | \text{document}) \rightarrow P_{\text{true}}(\text{rel} | \text{doc})$ in some way
- based on whatever data is available to system
 - $P_{\text{est}}(\text{relevant} | \text{document, query, context, user profile, ...})$
- best possible accuracy one can achieve with that data
 - recipe for a perfect IR system: just need $P_{\text{est}}(\text{relevant} | \dots)$
 - strong stuff, can this really be true?

Copyright © 2011 Victor Lavrenko

Probability of relevance

- What is: $P_{\text{true}}(\text{relevant} \mid \text{doc, qry, user, context})$?
 - isn't relevance just the user's opinion?
 - user decides relevant or not, what's the "probability" thing?
- "user" does not mean the human being
 - doc, qry, user, context ... *representations*
 - parts of the real thing that are available to the system
 - typical case: $P_{\text{true}}(\text{relevant} \mid \text{document, query})$
 - query: 2-3 keywords, user profile unknown, context not available
 - whether document is relevant is uncertain
 - depends on the factors which are not *available to our system*
 - think of $P_{\text{true}}(\text{rel} \mid \text{doc, qry})$ as proportion of all unseen users/contexts/... for which the document would have been judged relevant
- Analogy: $P(\text{die}=6 \mid \text{even and not square})$

Copyright © 2011 Victor Lavrenko

Optimality of PRP

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

Copyright © 2011 Victor Lavrenko

PRP = best possible ranking

- Let $D_i = \{\text{Document}_i, \text{Query}, \text{User}, \text{Task}, \text{Context}, \dots\}$
- Rank documents by $p_i = P(R_i=1 \mid D_i)$:

$$p_1 > \dots > p_i > \dots > p_j > \dots$$

$$\text{precision}_r = \frac{\text{rel}_r}{r} = \frac{1}{r} \sum_{i=1}^r \begin{cases} 1 \dots D_i \in \text{rel} \\ 0 \dots D_i \notin \text{rel} \end{cases}$$

- PRP gives highest:
 - expected precision at rank r
 - expected recall at r
 - F1 at r, average precision, ...

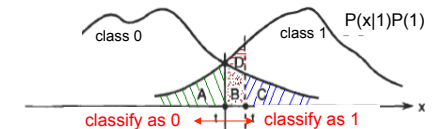
$$E[\text{precision}_r] = \frac{E[\text{rel}_r]}{r}$$

$$= \frac{1}{r} \sum_{i=1}^r \left\{ 1 \dots w.p.P(R=1 \mid D_i) \right\}$$

$$= \frac{1}{r} \sum_{i=1}^r p_i$$

- Ranking version of Bayes error rate

- best possible classification rate: relevant if $P(D, R=1) > P(D, R=0)$

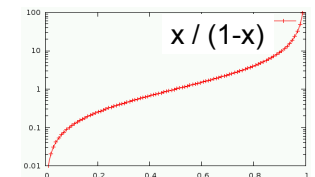


Copyright © 2011 Victor Lavrenko

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 $P(R=1 \mid D)$
 - $R = \{0, 1\}$... Bernoulli RV indicating relevance
 - D ... represents content of the document
- Rank-equivalent:

$$P(R=1 \mid D) = \frac{P(R=1 \mid D)}{P(R=0 \mid D)} = \frac{P(D \mid R=1)P(R=1)}{P(D \mid R=0)P(R=0)}$$
- Why Bayes? Want a generative model.
 - P (observation | class) sometimes easier with limited data
 - note: $P(R=1)$ and $P(R=0)$ don't affect the ranking



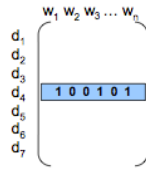
Copyright © 2011 Victor Lavrenko

Probabilistic model: assumptions

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

- Assumptions:

- A1: $D = \{D_w\}$... one RV 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: $10^6!$ possible events



- allows us to write:

$$P(R=1|D) = \frac{\text{rank } P(D|R=1)}{P(D|R=0)} = \frac{\prod_w P(D_w|R=1)}{\prod_w P(D_w|R=0)}$$

- Note: identical to the Naïve Bayes classifier

- with equal priors

Copyright © 2011 Victor Lavrenko

Probabilistic model: assumptions

- Define: $p_w = P(D_w=1|R=1)$ and $q_w = P(D_w=1|R=0)$
- Assumption A3 : $P(\vec{0}|R=1) = P(\vec{0}|R=0)$

- empty document (all words absent) is equally likely to be observed in relevant and non-relevant classes

$$P(R=1|D) \stackrel{\text{rank}}{=} \frac{\prod_w P(D_w|R=1)}{\prod_w P(D_w|R=0)} = \frac{\prod_w \left\{ \begin{matrix} p_w \dots w \in D \\ 1-p_w \dots w \notin D \end{matrix} \right\}}{\prod_w \left\{ \begin{matrix} q_w \dots w \in D \\ 1-q_w \dots w \notin D \end{matrix} \right\}}$$

$$= \frac{\prod_{w \in D} \left(\frac{p_w}{q_w} \right) \prod_{w \notin D} \left(\frac{1-p_w}{1-q_w} \right)}{\prod_w \left(\frac{1-p_w}{1-q_w} \right)} = \prod_{w \in D} \frac{p_w(1-q_w)}{q_w(1-p_w)}$$

- dividing by 1: no effect
- provides "natural zero"

$$\frac{P(\vec{0}|R=1)}{P(\vec{0}|R=0)} = 1$$
- practical reason: final product only over words present in D
 - fast: small % of total vocabulary + allows term-at-a-time execution

Copyright © 2011 Victor Lavrenko

Estimation (with R)

- Suppose we have (partial) relevance judgments:
 - N_1 ... relevant, N_0 ... non-relevant documents marked
 - word w observed in $N_1(w)$, $N_0(w)$ docs
 - $P(w)$ = % of docs that contain at least one mention of w
 - includes crude smoothing: avoids zeros, reduces variance

$$p_w = \frac{N_1(w) + 0.5}{N_1 + 1.0} \quad q_w = \frac{N_0(w) + 0.5}{N_0 + 1.0}$$

- What if we don't have relevance information?
 - no way to count words for relevant / non-relevant classes
 - things get messy...

Copyright © 2011 Victor Lavrenko

Example (with relevance)

- relevant docs: $D_1 = "a b c b d"$, $D_2 = "a b e f b"$
- non-relevant: $D_3 = "b g c d"$, $D_4 = "b d e"$, $D_5 = "a b e g"$

word:	a	b	c	d	e	f	g	h	
$N_1(w)$:	2	2	1	1	1	1	0	0	$N_1 = 2$
$N_0(w)$:	1	3	1	2	2	0	2	0	$N_0 = 3$
p_w :	$2.5/3$	$2.5/3$	$1.5/3$	$1.5/3$	$1.5/3$	$1.5/3$	$0.5/3$	$0.5/3$	
q_w :	$1.5/4$	$3.5/4$	$1.5/4$	$2.5/4$	$2.5/4$	$0.5/4$	$2.5/4$	$0.5/4$	

- new document $D_6 = "b g h"$:

$$P(R=1|D_6) \stackrel{\text{rank}}{=} \frac{\prod_{w \in d_6} \frac{p_w(1-q_w)}{q_w(1-p_w)}}{\prod_{w \in d_6} \frac{p_w(1-q_w)}{q_w(1-p_w)}} = \frac{\frac{2.5}{3} (1 - \frac{3.5}{4}) \frac{0.5}{3} (1 - \frac{2.5}{4}) \frac{0.5}{3} (1 - \frac{0.5}{4})}{\frac{3.5}{4} (1 - \frac{2.5}{3}) \frac{2.5}{4} (1 - \frac{0.5}{3}) \frac{0.5}{4} (1 - \frac{0.5}{3})} = \frac{1.64}{13.67}$$

only words present in D_6

Copyright © 2011 Victor Lavrenko

Estimation (without R)

- Assumption A4: $p_w = q_w \dots w \notin Q$
 - 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: $p_w = 0.5 \dots w \in Q$
 - a query word is equally likely to be present and absent in a randomly-picked relevant document (usually $p_w \ll 0.5$)
 - practical reason: p_w and $(1-p_w)$ cancel out

Assumption A6: $q_w \approx N_w / N$

- non-relevant set approximated by collection as a whole
- very reasonable: most documents are non-relevant

$$P(R=1|D) \stackrel{rank}{=} \prod_{w \in D} \frac{p_w(1-q_w)}{q_w(1-p_w)} = \prod_{w \in D \cap Q} \frac{p_w(1-q_w)}{q_w(1-p_w)} = \prod_{w \in D \cap Q} \frac{1-q_w}{q_w} = \prod_{w \in D \cap Q} \frac{N - N_w + 0.5}{N_w + 0.5}$$

IDF

Copyright © 2011 Victor Lavrenko

Example (no relevance)

- documents: $D_1 = "a b c b d", D_2 = "b e f b", D_3 = "b g c d", D_4 = "b d e", D_5 = "a b e g", D_6 = "b g h"$
- word: a b c d e f g h
 $N(w): 2 \quad 6 \quad 2 \quad 3 \quad 3 \quad 1 \quad 3 \quad 1 \quad N = 6$
- $N - N_w / N_w$: $4.5/2.5 \quad 0.5/6.5 \quad 4.5/2.5 \quad 3.5/3.5 \quad 3.5/3.5 \quad 5.5/1.5 \quad 3.5/3.5 \quad 5.5/1.5$
- query: $Q = "a c h"$

$$P(R=1|D_1) \stackrel{rank}{=} \prod_{w \in Q \cap D_1} \frac{N - N_w + 0.5}{N_w + 0.5} = \frac{4.5}{2.5} \cdot \frac{4.5}{2.5} \quad P(R=1|D_4) \stackrel{rank}{=} 1$$

$$P(R=1|D_2) \stackrel{rank}{=} 1 \quad P(R=1|D_5) \stackrel{rank}{=} \frac{4.5}{2.5}$$

$$P(R=1|D_3) \stackrel{rank}{=} \frac{4.5}{2.5} \quad P(R=1|D_6) \stackrel{rank}{=} \frac{5.5}{1.5}$$

only words present in both D & Q

Ranking: $D_6, D_1, D_3, D_5, D_2, D_4$

Copyright © 2011 Victor Lavrenko

Classical Probabilistic model: BIR

- Probability Ranking Principle: best possible ranking

Assumptions: $P(R=1|D) \stackrel{rank}{=} \prod_{w \in D} \frac{p_w(1-q_w)}{q_w(1-p_w)} = \prod_{w \in D \cap Q} \frac{N - N_w + 0.5}{N_w + 0.5}$

- A0: relevance for document in isolation
 - A1: words absent or present (can't model frequency)
 - A2: all words mutually independent (given relevance)
 - A3: empty document equally likely for R=0,1
 - A4: non-query words cancel out
 - A5: query words: relevant class doesn't matter
 - A6: non-relevant class ~ collection as a whole
- make it fast: D/Q overlap
- estimate p_w, q_w w/out relevance

- How can we improve the model?

- relax particularly bad assumptions

Copyright © 2011 Victor Lavrenko

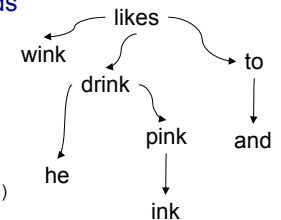
Modeling word dependence

- 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
 - edge weights: mutual information between words
- each word depends on its parent (and R)

- total # parameters: twice that of BIR

$$P("he likes to wink and drink pink ink") = P(\text{likes}) * P(\text{to}|\text{likes}) * P(\text{wink}|\text{likes}) * P(\text{and}|\text{to}) * P(\text{drink}|\text{likes}) * P(\text{he}|\text{drink}) * P(\text{pink}|\text{drink}) * P(\text{ink}|\text{pink})$$



Copyright © 2011 Victor Lavrenko

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 surface form of the string (is output grammatical?)
 - in IR we are already dealing with well-formed strings
 - pointless to waste probability mass on grammaticality
 - BIR doesn't really assume independence
 - necessary condition significantly weaker than independence

Copyright © 2011 Victor Lavrenko

Meaning of independence

- Independence:
 - seeing “subprime” doesn't affect chances of seeing “loan”
- Linked Dependence:
 - seeing “subprime” increases chance of seeing “loan”
 - by the same amount under R=1 and R=0
 - reasonable... unless topic is financial crisis
- Proportional Interdependence:
 - “subprime” increases chance of “loan”
 - can be more co-dependent in relevant class
 - as long as offset by other word sets under R=0
 - “world cup” more co-dependent in non-relevant class

Copyright © 2011 Victor Lavrenko

BIR doesn't assume independence

$$\frac{P_{R=1}(\vec{d})}{P_{R=0}(\vec{d})} = \prod_w \underbrace{\frac{P_1(d_w)}{P_0(d_w)}}_{\text{independence}} \times \underbrace{\prod_w \frac{k_1(w)}{k_0(w)}}_{\text{will not affect ranking if}} = \prod_w \underbrace{\frac{P_1(d_w | d_{\pi(w)})}{P_0(d_w | d_{\pi(w)})}}_{\text{1st order dependence}}$$

$$k_r(w) = \frac{P_r(d_w, d_{\pi(w)})}{P_r(d_w)P_r(d_{\pi(w)})}$$

$$\sum_w \log \frac{P_1(d_w, d_{\pi(w)})}{P_1(d_w)P_1(d_{\pi(w)})} \sim \sum_w \log \frac{P_0(d_w, d_{\pi(w)})}{P_0(d_w)P_0(d_{\pi(w)})}$$

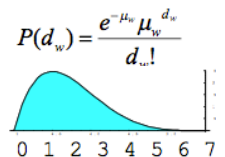
aggregate dependence between word and parent in the relevant class aggregate dependence in the non-relevant class

Sufficient condition: **proportional interdependence**
the total amount of interdependence among all words in a document is approximately the same under R=1 and R=0

Copyright © 2011 Victor Lavrenko

Modeling word frequencies

- Want to model TF (empirically useful) $P(R=1|D) = \prod_{w \in D} \frac{P(D_w | R=1)}{P(D_w | R=0)}$
 - A1': assume $D_w = d_w \dots$ # times word w occurs in document D
 - estimate $P(d_w|R)$: e.g. “obama” occurs 5 times in a rel. doc
 - naive: estimate probability for every outcome:
 - $P(D_w=0), P(D_w=1), P(D_w=2), P(D_w=3) \dots$ for R=1 and R=0
 - many outcomes → many parameters (BIR had only one p_w)
 - want “smoothness”: $d_w=5$ similar to $d_w=6$, but not $d_w=1$
 - parametric model: assume $d_w \sim \text{Poisson}$
 - single parameter $\mu_w \dots$ expected frequency
 - problem: Poisson a poor fit to observations
 - does not capture bursty nature of words



Copyright © 2011 Victor Lavrenko

Two-Poisson model [Harter]

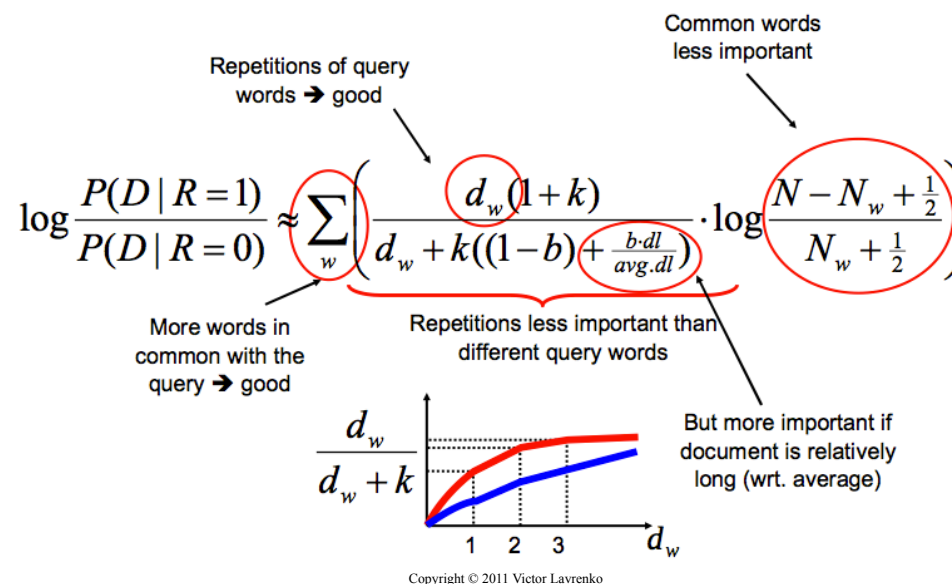
- Idea: words generated by a mixture of two Poissons
 - “elite” words for a document: occur unusually frequently
 - “non-elite” words – occur as expected by chance
 - document is a mixture: $P(d_w) = P(E=1) \frac{e^{-\mu_{1,w}} \mu_{1,w}^{d_w}}{d_w!} + P(E=0) \frac{e^{-\mu_{0,w}} \mu_{0,w}^{d_w}}{d_w!}$
 - estimate $m_{0,w}, m_{1,w}, P(E=1)$ by fitting to data (max. likelihood)
- 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

$$\frac{p_w(d_w)q_w(0)}{q_w(d_w)p_w(0)} \approx \exp\left(\underbrace{\frac{d_w(1+k)}{d_w + k((1-b) + b \cdot dl / avg.dl)}_{\text{squashed TF, k,b: parameters}} \cdot \log \frac{N - N_w + 0.5}{N_w + 0.5}}_{\text{IDF}}\right)$$

Copyright © 2011 Victor Lavrenko

BM25: an intuitive view



Example: BM25

- documents: $D_1 = \text{“a b c b d”}$, $D_2 = \text{“b e f b”}$, $D_3 = \text{“b g c d”}$, $D_4 = \text{“b d e”}$, $D_5 = \text{“a b e g”}$, $D_6 = \text{“b g h h”}$
- query: $Q = \text{“a c h”}$, assume $k = 1, b = 0.5$

word:	a	b	c	d	e	f	g	h	
$N(w)$:	2	6	2	3	3	1	3	1	$N = 6$
$N - N_w / N_w$:	$4.5/2.5$	$0.5/6.5$	$4.5/2.5$	$3.5/3.5$	$3.5/3.5$	$5.5/1.5$	$3.5/3.5$	$5.5/1.5$	

$$\log \frac{P(D_1 | R = 1)}{P(D_1 | R = 0)} \approx 2 \times \left(\frac{1 \cdot (1 + 1)}{1 + 1 \cdot (0.5 + \frac{0.5 \cdot 5}{4})} \cdot \log \frac{4.5}{2.5} \right)$$

$$\log \frac{P(D_6 | R = 1)}{P(D_6 | R = 0)} \approx \left(\frac{2 \cdot (1 + 1)}{2 + 1 \cdot (0.5 + \frac{0.5 \cdot 4}{4})} \cdot \log \frac{5.5}{1.5} \right)$$

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, develop BM25