CS 6740, Spring 2010: Advanced Language Technologies
Lecture 7 - February 18, 2010
Lecturer: *Prof. Lillian Lee* Scribe: *Taiyang Chen*

1 Introduction

Recall the Robertson-Spärck Jones probabilistic retrieval scoring function [3]:

$$RSJ_q(d) = P(r \mid \vec{F} = \vec{f_d}, \vec{q}) \tag{1}$$

$RSJ_q(d)$ represents the relevance score of a document d against a query q, under a corpus C. The right-hand side of (1) can be interpreted as the probability of being relevant (r) given that an attribute vector ($\vec{F}$) is the same as the attribute vector of the document ($\vec{f_d}$), and the query vector ($\vec{q}$) itself.
Further derivation gives us the following form[1]:

$$RSJ_q(d) \stackrel{rank}{=} \sum_{j:q[j]>0, f_d[j]>0} \log\left(\frac{P(F[j] = f_d[j] \mid r, \vec{q})}{P(F[j] = f_d[j])} \times \frac{P(F[j] = 0)}{P(F[j] = 0 \mid r, \vec{q})}\right) \tag{2}$$

Note that (2) is derived by applying the log function on the Lecture 5 result (product of the fractional terms), as in the original RSJ model. This is the simplest form we have before applying the binary attribute assumption from Lecture 6. This form leads to the derivation of the Okapi BM25 scoring function which does not rely on any binary attribute function assumption [4].
First we only focus on the terms inside the log function. Label the two multiplying fractions for a j such that $f_d[j] > 0$ and $q[j] > 0$:

$$\frac{P(F[j] = f_d[j] \mid r, \vec{q})}{P(F[j] = f_d[j])} \times \frac{P(F[j] = 0)}{P(F[j] = 0 \mid r, \vec{q})} \equiv \frac{\mathsf{A}}{\mathsf{B}} \times \frac{\mathsf{C}}{\mathsf{D}} \tag{3}$$

Observe the above quantity. As a relevance scoring function, it should be high when the document d in question is relevant. $\frac{\mathsf{A}}{\mathsf{B}}$ is large when the attribute function $F[j]$ has value $f_d[j]$ in relevant documents more often that it does in general documents. On the other hand, $\frac{\mathsf{C}}{\mathsf{D}}$ is large when general documents are more likely to omit the attribute function $F[j]$ than relevant documents.

2 Estimating Term Frequencies

Now instead of applying the binary attribute assumption, we can assume $F[j]$'s in (3) to be term frequencies. Thus we need, as part of the RSJ weight A, to be able to compute

$$\hat{p}(F[j] = x \mid r, \vec{q}), x = 0, 1, 2, ... \tag{4}$$

To further evaluate the scoring function $RSJ_q(d)$, we need some way to estimate the probabilities of the term frequencies. Naturally, we need to estimate an easy-to-compute distribution for F. Continuous distribution is not considered because our x's in (4) are all integers. Possible discrete options are:

- Binomial Distribution

  Each document d has l word slots and each slot has probability p of having the term v_j, and $1 - p$ otherwise. This is a feasible scheme but the binomial coefficients can be messy.

- Poisson Distribution

  Assuming that all documents have the same length, for a set of documents with l word slots each, attribute v_j occurs at some steady rate on average (e.g. 14/100 words). We assume the documents have the same length for now (and will deal with different lengths later). This is not an entirely realistic approach. But it is convenient and also similar to the Binomial distribution in that, for big l and small p, we can model $Binomial(l, p)$ with $Poisson(lp)$. Poisson also has a cleaner form, and hence we pick it to simplify our RSJ scoring function.

[1] All log functions used in the notes are base-10 logarithmic functions

3 Poisson Transformation

The general form of a Poisson distribution with the rate of occurrence μ (for a unit text length) is:

$$Poisson(X = x) = \frac{e^{-\mu}\mu^x}{x!} \tag{5}$$

Let ρ_j (rho for relevant) be the expected number of occurrences of v_j in relevant documents, and γ_j (gamma for general) be the expected number of occurrences of v_j in general documents. Then we take $P(F[j] = x \mid r, \vec{q})$ as a $Poisson(\rho_j)$ distribution and $P(F[j] = x)$ as a $Poisson(\gamma_j)$ distribution.
The four parts of (3) are turned into:

$$\mathsf{A} = \frac{e^{-\rho_j}\rho_j^{f_d[j]}}{(f_d[j])!} \tag{6}$$

$$\mathsf{B} = \frac{e^{-\gamma_j}\gamma_j^{f_d[j]}}{(f_d[j])!} \tag{7}$$

$$\mathsf{C} = e^{-\gamma_j} \tag{8}$$

$$\mathsf{D} = e^{-\rho_j} \tag{9}$$

Multiplying them together again:

$$\frac{\mathsf{A}}{\mathsf{B}} \times \frac{\mathsf{C}}{\mathsf{D}} = \left(\frac{\rho_j}{\gamma_j}\right)^{f_d[j]} \tag{10}$$

Adding back the log and sum functions from (2):

$$RSJ_q(d) \overset{rank}{=} \sum_{j:q[j]>0, f_d[j]>0} f_d[j] \log\left(\frac{\rho_j}{\gamma_j}\right) \tag{11}$$

This result is not as "IDF" as the other estimation approaches. The fraction is the number of occurrences of a term in relevant documents over those in general documents. It is not exactly an inverse of the number of occurrences over all documents.

4 2-Poisson Model

We need a better strategy to deal with $RSJ_q(d)$ and this is where the famous 2-Poisson model comes in [1] [2]. The idea is that relevance should really be based on subject matter (topic). This is not to be confused with today's "topic modeling". Here the topic refers to the content of the document expressed by a given term.
Our previous sample space was either relevant or non-relevant. But now it is finer-grained because we can define, for each term v_j whether the document is on-topic ($\mathsf{on}[j]$) or off-topic ($\mathsf{off}[j]$). The definition of the new sample space is shown below:

$$\text{documents} \times \text{queries} \times \{r, \bar{r}\} \times \{\mathsf{on}[1], \mathsf{off}[1]\} \times ... \times \{\mathsf{on}[m], \mathsf{off}[m]\} \tag{12}$$

We can estimate the RSJ weight $P(F[j] = x)$ based on whether a document d is on-topic or off-topic with respect to the term v_j (x can refer to the previous $f_d[j]$). This idea leads to the probability function:

$$P(F[j] = x) = P(F[j] = x \mid \mathsf{on}[j])P(\mathsf{on}[j]) + P(F[j] = x \mid \mathsf{off}[j])P(\mathsf{off}[j]) \tag{13}$$

Now we can start applying the 2-Poisson model to (13), one for the on-term and the second (thus 2) for the off-term. Similar to the previous approach, let $\tau[j]$ be the expected number of occurrences of v_j given that the document is on topic w.r.t. v_j ("tau" stands for "on topic") and $\mu[j]$ be the expected number of occurrences of v_j given that the document is not on topic w.r.t. v_j.
Hence we model $P(F[j] = x \mid \mathsf{on}[j])$ as a $Poisson(\tau[j])$ distribution and $P(F[j] = x \mid \mathsf{off}[j])$ as a $Poisson(\mu[j])$ distribution. Also assume that $\tau[j] > \mu[j]$ because it is more likely a term occurs when the document is on the term's topic than in general documents.

The 2-Poisson derivation of (4), i.e., A of the RSJ weight:

$$\begin{aligned} P(F[j] = x \mid r, \vec{q}) &= \sum_{o \in \{\mathsf{on}[j], \mathsf{off}[j]\}} P(F[j] = x \mid o, r, \vec{q}) P(o \mid r, \vec{q}) \\ &= \sum_{o \in \{\mathsf{on}[j], \mathsf{off}[j]\}} P(F[j] = x \mid o) P(o \mid r, \vec{q})^{\star} \\ &= \frac{e^{-\tau[j]} \tau[j]^x}{x!} P(\mathsf{on}[j] \mid r, \vec{q}) + \frac{e^{-\mu[j]} \mu[j]^x}{x!} P(\mathsf{off}[j] \mid r, \vec{q}) \\ &= \frac{e^{-\tau[j]} \tau[j]^x}{x!} P(\mathsf{on}[j] \mid r, \vec{q}) + \frac{e^{-\mu[j]} \mu[j]^x}{x!} (1 - P(\mathsf{on}[j] \mid r, \vec{q})) \end{aligned} \tag{14}$$

The step $\star$ is taken because the number of occurrences ($x = f_d[j]$) in d of a term v_j is independent of the binary relevance and the query vector given that d is on the topic of v_j. The only thing that affects the relationship between $F[j]$ and x is whether d is on the topic of v_j or not. Thus the expression only depends on $o \in \{\mathsf{on}[j], \mathsf{off}[j]\}$.

Now our messy weight is expressed as a single function of x (besides the unknown but fixed on and off-topic probabilities). We try to understand how this function behaves in terms of x by considering the following 3 cases:

1. $x = 0$

   First we know that given some condition c,

   $$P(F[j] = x = 0 \mid c) = e^{-\tau[j]} P(\mathsf{on}[j] \mid c) + e^{-\mu[j]} (1 - P(\mathsf{on}[j] \mid c)) \tag{15}$$

   With (15), the two fractions in (3) become

   $$\frac{\mathsf{A}}{\mathsf{B}} = \frac{e^{-\tau[j]} P(\mathsf{on}[j] \mid r, \vec{q}) + e^{-\mu[j]} (1 - P(\mathsf{on}[j] \mid r, \vec{q}))}{e^{-\tau[j]} P(\mathsf{on}[j]) + e^{-\mu[j]} (1 - P(\mathsf{on}[j]))} \tag{16}$$

   $$\frac{\mathsf{C}}{\mathsf{D}} = \frac{e^{-\tau[j]} P(\mathsf{on}[j]) + e^{-\mu[j]} (1 - P(\mathsf{on}[j]))}{e^{-\tau[j]} P(\mathsf{on}[j] \mid r, \vec{q}) + e^{-\mu[j]} (1 - P(\mathsf{on}[j] \mid r, \vec{q}))} \tag{17}$$

   Notice that (16) and (17) are multiplicative inverses. Thus,

   $$\begin{aligned} RSJ_q(d) &= \sum_{j : q[j] > 0, f_d[j] > 0} \log \left(\frac{\mathsf{A}}{\mathsf{B}} \times \frac{\mathsf{C}}{\mathsf{D}} \right) \\ &= \sum_{j : q[j] > 0, f_d[j] > 0} \log(1) \\ &= 0 \end{aligned} \tag{18}$$

2. $x \to \infty$

   First for some constants a and b w.r.t. x and condition c, we claim

   $$\lim_{x \to \infty} P(F[j] = x \mid c) = \lim_{x \to \infty} \frac{a \tau[j]^x + b \mu[j]^x}{x!} \tag{19}$$

   Let $a, ..., f$ be some constants w.r.t. x. Combine (15) and (19) and given the assumption $\dfrac{\mu[j]}{\tau[j]} < 1$:

$$
\begin{aligned}
RSJ_q(d) &= \sum_{j:q[j]>0, f_d[j]>0} \lim_{x\to\infty} \log\left(\frac{\mathsf{A}}{\mathsf{B}} \times \frac{\mathsf{C}}{\mathsf{D}}\right) \\
&= \sum_{j:q[j]>0, f_d[j]>0} \lim_{x\to\infty} \log\left(\frac{a\tau[j]^x + b\mu[j]^x}{c\tau[j]^x + d\mu[j]^x} \times \frac{e}{f}\right) \\
&= \sum_{j:q[j]>0, f_d[j]>0} \lim_{x\to\infty} \log\left(\frac{ae + be\left(\frac{\mu[j]}{\tau[j]}\right)^x}{cf + df\left(\frac{\mu[j]}{\tau[j]}\right)^x}\right) \\
&= \sum_{j:q[j]>0, f_d[j]>0} \log\left(\frac{ae}{cf}\right)
\end{aligned}
\tag{20}
$$

The result (18) is better than (11) because based on some further (not necessarily highly plausible) assumptions, this formula approximates the original binary-model RSJ. It will produce the IDF via any of the three derivations we saw from Lecture 6.

3. $x \in (0, \infty)$

   Robertson and Walker [4] assert that this function is monotonically increasing in x.

5 Okapi BM25

So far we have shown that the 2-Poisson model yields a function that is monotonically increasing and approaches the desired IDF. But we still have unknowns such as $P(\mathsf{on}[j])$, and we need to find such a function $f(x)$. The idea is that the function is small when $x = 0$, and increases monotonically close to IDF_j. The proposal in [4] is, for some parameter k, the following:

$$
f(x) = \frac{x}{k + x} IDF_j \tag{21}
$$

But wait - what about our same length assumption? We can now add in (a simple) normalization resembling and predating the pivoted document length approach [6], the idea being that k is meant for "unit length" documents:

$$
f(x) = \frac{x}{k\dfrac{length(d)}{\text{average document length}} + x} IDF_j \tag{22}
$$

Now we have analyzed the intrinsics of the BM-style scoring function family [5]. As a comparison, the middle fraction is similar to what we have just derived:

$$
BM25 = \sum_{t\in Q,D} ln\frac{N - df + 0.5}{df + 0.5} \times \frac{(k_1 + 1)tf}{k_1((1 - b) + b\dfrac{dl}{avdl}) + tf} \times \frac{(k_3 + 1)qtf}{k_3 + qtf} \tag{23}
$$

6 Finger Exercises

1. The 2-Poisson model explores the idea that a query term is either on or off-topic with respect to the document. What about a n-Poisson model? For instance, in a corpus of sports magazines, readers may query for their desired type of sport (say skiing). However, expert skiers are definitely not interested in Skiing 101. They could find a perfectly matching article on how to brake and prevent being hurt, but these techniques may not fit for high mountain trails. If we use a (hypothesized) 4-Poisson model with Olympian, Expert, Normal and Beginner "levels", the definition of on-topic is more fine-grained. Now given this proposal, analyze the new model and give reasons why or why not this is a feasible approach.

2. We mentioned a crude RSJ weight approximation function in (22) based on the tunable parameter k. But it is unknown how feasible it is to compute k. Work through the following straightforward exercise and suggest (potential) methods for choosing this parameter.

   The documents:

id	**content**
d_1	cat dog dog dog
d_2	dog cat dog
d_3	you me and nothing
d_4	cat dog nothing
d_5	you and

   Suppose we have a query $q =$"you me dog" and we hope the most relevant document is d_3 because we're only missing the dog reference. Compute all the relevance scores using the $RSJ_q(d)$ with (22). Find out a k that makes the most sense.

7 Solutions

1. Start by defining the concept of a "level". Let $r_l[j]$ be the random variable that a document d is on-topic for a term v_j at level l, for $0 \leq l \leq n$. Trivially, $r_0[j] = \mathsf{on}[j]$ and $r_n[j] = \mathsf{off}[j]$, and there are $n+1$ levels in total.

   Trivially, we have:

$$P(F[j] = x) = \sum_{0 \leq l \leq n} P(F[j] = x \mid r_l[j]) P(r_l[j]) \tag{24}$$

   Each $P(F[j] = x \mid r_l[j])$ is modeled as a $Poisson(\mu_l[j])$ distribution. And assume $\mu_0[j] > ... > \mu_n[j]$.

   Similar to (14), we have the following for A of the RSJ scoring function:

$$\begin{aligned} P(F[j] = x \mid r, \vec{q}) &= \sum_{0 \leq l \leq n} P(F[j] = x \mid r_l[j], r, \vec{q}) P(r_l[j] \mid r, \vec{q}) \\ &= \sum_{0 \leq l \leq n} P(F[j] = x \mid r_l[j]) P(r_l[j] \mid r, \vec{q}) \\ &= \sum_{0 \leq l \leq n} \frac{e^{-\mu_l[j]} \mu_l[j]^x}{x!} P(r_l[j] \mid r, \vec{q}) \end{aligned} \tag{25}$$

   Finally, we consider the 3 cases of x again.

   - $x = 0$
     Substituting $f_d[j] = x = 0$ in (3),

$$\frac{\mathsf{A}}{\mathsf{B}} = \frac{P(F[j] = 0 \mid r, \vec{q})}{P(F[j] = 0)} \tag{26}$$

$$\frac{\mathsf{C}}{\mathsf{D}} = \frac{P(F[j] = 0)}{P(F[j] = 0 \mid r, \vec{q})} \tag{27}$$

     We see that (26) and (27) are always multiplicative inverses no matter what $P(F[j] = 0 \mid c)$ is. Thus given (18), the scoring function is still 0.

   - $x \to \infty$
     This case now involves more Poisson distributions than before. Let a_l, b_l, c, and d denote constants for $0 \leq l \leq n$.

$$
\begin{aligned}
RSJ_q(d) &= \sum_{j:q[j]>0, f_d[j]>0} \lim_{x\to\infty} \log\left(\frac{\mathsf{A}}{\mathsf{B}} \times \frac{\mathsf{C}}{\mathsf{D}}\right) \\
&= \sum_{j:q[j]>0, f_d[j]>0} \lim_{x\to\infty} \log\left(\frac{\sum\limits_{0\le l\le n} a_l \mu_l[j]^x}{\sum\limits_{0\le l\le n} b_l \mu_l[j]^x} \times \frac{c}{d}\right) \\
&= \sum_{j:q[j]>0, f_d[j]>0} \lim_{x\to\infty} \log\left(\frac{a_0 c + \sum\limits_{1\le l\le n} a_l c \left(\frac{\mu_l[j]}{\mu_0[j]}\right)^x}{b_0 d + \sum\limits_{1\le l\le n} b_l d \left(\frac{\mu_l[j]}{\mu_0[j]}\right)^x}\right) \\
&= \sum_{j:q[j]>0, f_d[j]>0} \log\left(\frac{a_0 c}{b_0 d}\right)
\end{aligned}
\tag{28}
$$

The n-Poisson approach also exhibits the IDF approximation when using the same assumptions, as x goes to infinity.

- $x \in (0, \infty)$
  So far, n-Poisson is similar in behavior to 2-Poisson. They only differ in how the scoring function grows. In 2-Poisson, as (20) shows, $\frac{\mu[j]}{\tau[j]}$ dominates $\frac{\mathsf{A}}{\mathsf{B}}$. This quantity controls the rate of growth towards the ideal IDF. Similarly the n-Poisson scoring function is controlled by $\frac{\mu_l[j]}{\mu_0[j]}$'s for all $1 \le l \le n$. If we assume that $\tau[j] \approx \mu_0[j]$ and $\mu[j] \approx \sum_{l=1}^{n} \mu_l[j]$, then, $\left(\frac{\mu[j]}{\tau[j]}\right)^x \ge \sum_{l=1}^{n} \left(\frac{\mu_l[j]}{\mu_0[j]}\right)^x$ because the exponentiation terms are all positive and $x \ge 1$. The n-Poisson growth is somewhat "finer-grained" (steadier) than the 2-Poisson approach.

Feasibility. The only advantage of having n-Poisson is a steadier scoring function. However, the assumption $\mu_0[j] > ... > \mu_n[j]$ is harder to maintain. It is not always the case a "higher" level topic is more likely to occur in a corpus. This is analogous to cutting a loaf of bread. The first cut (into 2 parts) is always easy such that one is bigger than the other. The more times you cut the parts, the less likely you can separate them into distinct sizes. Hence, unless there is a big penalty for having very distinct scores, we should choose the 2-Poisson model. An example of this penalty would be the case where the scores in product marketing should be relatively close - to avoid monopoly claims.

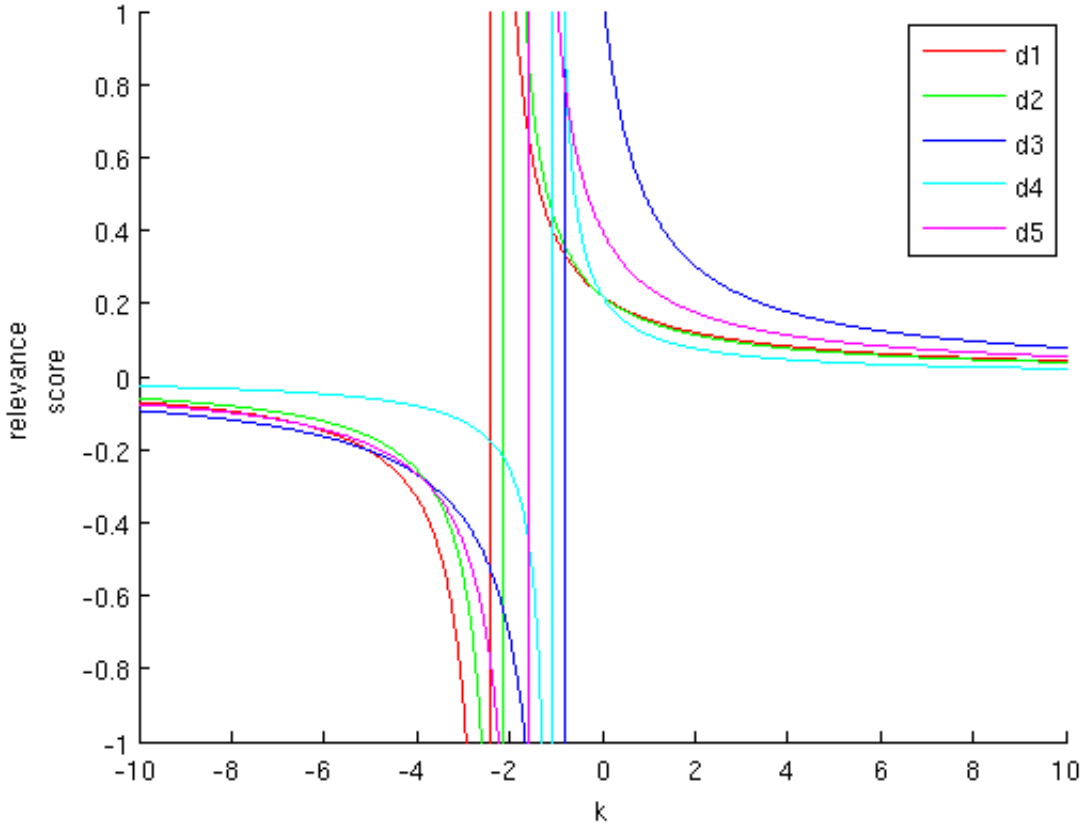
2. The IDF table:

v_j	$IDF_j = log(N/n_j)$
and	log(5/2)=0.3979
cat	log(5/3)=0.2218
dog	log(5/3)=0.2218
me	log(5/1)=0.6690
nothing	log(5/2)=0.3979
you	log(5/2)=0.3979

Average document length = 3.2. Relevance score table with query "you me dog":

id	$RSJ_q(d)$
d_1	$0.6654/(1.25k+3)$
d_2	$0.4436/(0.9375k+2)$
d_3	$0.3979/(1.25k+1) + 0.6690/(1.25k+1) = 1.0669/(1.25k+1)$
d_4	$0.2218/(0.9375k+1)$
d_5	$0.3979/(0.625k+1)$

We plot a k vs. $RSJ_q(d_1) \dots RSJ_q(d_5)$ graph to determine sensible values of k:



When $k \in [0, 4]$, the relevance score for d_3 is dominating all the others. Note that $[-2, 0]$ is a particularly bad range for this problem because of undefined points ($RSJ_q(d)$'s denominator cannot be 0). These values of k should obviously be avoided. Negative values do not make sense either: d_4 becomes the high-scorer. In fact, since the denominator of (22) contains a normalization term, only $k \geq 0$ (perhaps small negatives) is usually considered.

References

1. Abraham Bookstein and D. R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):312318 (1974).

2. Stephen P. Harter. A probabilistic approach to automatic keyword indexing, part I: On the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science* 26(4):197206, 1975.

3. Stephen E. Robertson and Karen Spärck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science* 27(3), 12946 (1976).

4. Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. *SIGIR*, pp. 232241 (1994).

5. Amit Singhal. Modern Information Retrieval: A Brief Overview. *IEEE Data Engineering Bulletin*, 2001.

6. Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. *SIGIR* 1996. See also Singhal, Salton, Mitra and Buckley, Document length normalization, *IPM* 32(5):619633, 1996.