Deriving the Robertson / Spärck Jones Model from Probability of Relevance

Chase Geigle
University of Illinois at Urbana-Champaign
Department of Computer Science
geigle1@illinois.edu

September 23, 2016

1 From Probability of Relevance to Document Generation

We wish to rank documents by the probability that they are relevant to a given query. In other words, we care about

$$O(R = 1 \mid Q, D) = \frac{P(R = 1 \mid Q, D)}{P(R = 0 \mid Q, D)}$$

where Q is our query, D is our document, and R is a binary random variable indicating whether the document is relevant to the query. Using Bayes' rule, we can rewrite this as

$$\frac{P(R = 1 \mid Q, D)}{P(R = 0 \mid Q, D)} = \frac{P(Q, D \mid R = 1)P(R = 1)}{P(Q, D \mid R = 0)P(R = 0)}$$

and we can further simplify by noting that the prior probability of relevance or non-relevance ($P(R = 1)$ and $P(R = 0)$, respectively) do *not vary between documents to be ranked* and thus simply a constant when computing all the scores for all of the documents.

Now we are faced with understanding the joint probability of a query and a document given a relevance status. We can further decompose this in two different ways by using the definition of conditional probability:

- $P(Q, D \mid R) = P(Q \mid D, R)P(D \mid R)$, which is known as "query generation", or
- $P(Q, D \mid R) = P(D \mid Q, R)P(Q \mid R)$, which is known as "document generation".

The derivation of the RSJ model follows the second formulation, where we wish to compute the probability of generating a document given a query and relevance status:

$$\frac{P(Q, D \mid R = 1)}{P(Q, D \mid R = 0)} = \frac{P(D \mid Q, R = 1)P(Q \mid R = 1)}{P(D \mid Q, R = 0)P(Q \mid R = 0)}$$

Note that we can even further simplify this ratio by noting that the probability $P(Q \mid R)$ does not depend on a specific document and thus will be a fixed constant as we vary the document when computing scores for ranking. Thus, we arrive at the final quantity for the document generation ranking formulation

$$\frac{P(D \mid Q, R = 1)}{P(D \mid Q, R = 0)}.$$

2 Deriving RSJ as Document Generation

Starting from the document generation setup, we have

$$\frac{P(R=1 \mid Q,D)}{P(R=0 \mid Q,D)} \propto \frac{P(D \mid Q,R=1)}{P(D \mid Q,R=0)}$$

From here, we need to decide upon a representation for the document D and the query Q. In the RSJ model, we assume that we represent the document D as a vector of dimension $|V|$, where V is our vocabulary set. Each element in this vector, d_i, takes a binary value indicating presence or absence of the specific term $w_i \in V$ in the document D. So we have $D = (d_1, d_2, \ldots, d_{|V|})$. We can now define the probability of generating such a document by letting A_i be a binary random variable that indicates the value taken by d_i. Then we can write

$$P(D \mid Q,R) = P(A_1 = d_1, A_2 = d_2, \ldots, A_{|V|} = d_{|V|} \mid Q,R).$$

If we make a further assumption that each A_i is independent of each of the other A_j, we have

$$P(D \mid Q,R) = \prod_{i=1}^{|V|} P(A_i = d_i \mid Q,R).$$

From here, we can split up the product into terms that are present in the document D and terms that are not present in the document D. Why might we want to do this? One reason is for computational efficiency: $|V|$ is very large for most search engines, whereas the number of present terms in D is likely very small in comparison. If we can separate these terms, and find some way to avoid actually computing the product over all terms that are not present in D, we can arrive at a retrieval scoring function that can be computed quickly in practice. Thus, returning to our odds ratio, we have

$$\frac{P(D \mid Q,R=1)}{P(D \mid Q,R=0)} = \prod_{i=1,d_i=1}^{|V|} \frac{P(A_i=1 \mid Q,R=1)}{P(A_i=1 \mid Q,R=0)} \prod_{i=1,d_i=0}^{|V|} \frac{P(A_i=0 \mid Q,R=1)}{P(A_i=0 \mid Q,R=0)}$$

Here, the first term is over only the terms *present* in D, and the second term is only over the terms *not present* in D. We could stop here, but as mentioned before it would be convenient at runtime to have this be only a product (or, in log-space, a sum) over the *present* terms in D only. Achieving this requires a few observations.

First, we note that a part of the formula can be dropped from this ratio (and leave the ranking the same) if it is independent with respect to the choice of the document D, as in such a case it is simply a constant factor in the computation for scoring each document.

Second, we note that the part of our formula we would like to drop (the second product) is ***not independent of the document***, as the set of non-present terms clearly depends on our choice of D. In order to drop this product, we would have to ensure that it does not depend on the document or, in other words, is the same quantity across all choices of D.

Third, we make the following observation:

$$\prod_{i=1}^{|V|} \frac{P(A_i=0 \mid Q,R=1)}{P(A_i=0 \mid Q,R=0)} = \prod_{i=1,d_i=0}^{|V|} \frac{P(A_i=0 \mid Q,R=1)}{P(A_i=0 \mid Q,R=0)} \prod_{i=1,d_i=1}^{|V|} \frac{P(A_i=0 \mid Q,R=1)}{P(A_i=0 \mid Q,R=0)}$$

This is clear because the sets $\{w_i \mid d_i = 1\}$ and $\{w_i \mid d_i = 0\}$ are disjoint and their union is V. Notice that this expression *is* constant with respect to our choice of D because, no matter what the distribution

of terms is for the first and second product, we still end up with a product over the entire vocabulary. Thus, we can transform our equation above as follows:

$$\begin{aligned}
\frac{P(D \mid Q, R=1)}{P(D \mid Q, R=0)} &= \prod_{i=1, d_i=1}^{|V|} \frac{P(A_i=1 \mid Q, R=1)}{P(A_i=1 \mid Q, R=0)} \prod_{i=1, d_i=0}^{|V|} \frac{P(A_i=0 \mid Q, R=1)}{P(A_i=0 \mid Q, R=0)} \\
&= \prod_{i=1, d_i=1}^{|V|} \frac{P(A_i=1 \mid Q, R=1)}{P(A_i=1 \mid Q, R=0)} \prod_{i=1, d_i=0}^{|V|} \frac{P(A_i=0 \mid Q, R=1)}{P(A_i=0 \mid Q, R=0)} \cdot \frac{\prod_{i=1, d_i=1}^{|V|} \frac{P(A_i=0|Q,R=1)}{P(A_i=0|Q,R=0)}}{\prod_{i=1, d_i=1}^{|V|} \frac{P(A_i=0|Q,R=1)}{P(A_i=0|Q,R=0)}} \\
&= \prod_{i=1, d_i=1}^{|V|} \frac{P(A_i=1 \mid Q, R=1)}{P(A_i=1 \mid Q, R=0)} \prod_{i=1}^{|V|} \frac{P(A_i=0 \mid Q, R=1)}{P(A_i=0 \mid Q, R=0)} \cdot \frac{1}{\prod_{i=1, d_i=1}^{|V|} \frac{P(A_i=0|Q,R=1)}{P(A_i=0|Q,R=0)}} \\
&= \prod_{i=1, d_i=1}^{|V|} \frac{\frac{P(A_i=1|Q,R=1)}{P(A_i=1|Q,R=0)}}{\frac{P(A_i=0|Q,R=1)}{P(A_i=0|Q,R=0)}} \prod_{i=1}^{|V|} \frac{P(A_i=0 \mid Q, R=1)}{P(A_i=0 \mid Q, R=0)} \\
&= \prod_{i=1, d_i=1}^{|V|} \frac{P(A_i=1 \mid Q, R=1) P(A_i=0 \mid Q, R=0)}{P(A_i=1 \mid Q, R=0) P(A_i=0 \mid Q, R=1)} \prod_{i=1}^{|V|} \frac{P(A_i=0 \mid Q, R=1)}{P(A_i=0 \mid Q, R=0)} \\
&\propto \prod_{i=1, d_i=1}^{|V|} \frac{P(A_i=1 \mid Q, R=1) P(A_i=0 \mid Q, R=0)}{P(A_i=1 \mid Q, R=0) P(A_i=0 \mid Q, R=1)}
\end{aligned}$$

If we further make the assumption that, for a term w_i such that $w_i \notin Q$, $P(A_i=1 \mid Q, R=1) = P(A_i = 1 \mid Q, R=0)$ (thus we assume words that do not appear in the query are equally likely to be present in relevant and non-relevant documents), we arrive at:

$$\frac{P(D \mid Q, R=1)}{P(D \mid Q, R=0)} \propto \prod_{i=1, d_i=q_i=1}^{|V|} \frac{P(A_i=1 \mid Q, R=1) P(A_i=0 \mid Q, R=0)}{P(A_i=1 \mid Q, R=0) P(A_i=0 \mid Q, R=1)}$$

From here, we can arrive at the RSJ model by taking the logarithm of our odds ratio:

$$\begin{aligned}
\log \frac{P(D \mid Q, R=1)}{P(D \mid Q, R=0)} &\propto \log \prod_{i=1, d_i=q_i=1}^{|V|} \frac{P(A_i=1 \mid Q, R=1) P(A_i=0 \mid Q, R=0)}{P(A_i=1 \mid Q, R=0) P(A_i=0 \mid Q, R=1)} \\
&= \sum_{i=1, d_i=q_i=1}^{|V|} \log \frac{P(A_i=1 \mid Q, R=1) P(A_i=0 \mid Q, R=0)}{P(A_i=1 \mid Q, R=0) P(A_i=0 \mid Q, R=1)} \\
&= \sum_{i=1, d_i=q_i=1}^{|V|} \log \frac{P(A_i=1 \mid Q, R=1)(1 - P(A_i=1 \mid Q, R=0))}{P(A_i=1 \mid Q, R=0)(1 - P(A_i=1 \mid Q, R=1))}
\end{aligned}$$

Rewriting using the definitions $p_i = P(A_i=1 \mid Q, R=1)$ and $q_i = P(A_i=1 \mid Q, R=0)$, we have the final formula that

$$\log O(R=1 \mid Q, D) \approx_{rank} \sum_{i=1, d_i=q_i=1}^{|V|} \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

This summation can now be performed quickly at retrieval time as we are only summing over the matched query terms (which will typically be a very sparse set).