CS 674/INFO 630: Advanced Language Technologies Fall 2007

Lecture 8 (Part 2) — September 20, 2007

Prof. Lillian Lee Scribe: *Vasumathi Raman + additional edits by L. Lee*

Note Additional edits were made by L. Lee to correct a lecturer-induced problem, and to make Part 2 consistent with Part 1.

4 Approximation

Given the exponentially decreasing numerator and denominator and the upper and lower limits proved, and assuming that $T_j = y$ and $R_q = y$ are positively correlated, the function under consideration increases monotonically with $d[j]$ to an asymptote at the *idf*. But our function still has too many unknowns, so the easiest way out is to try and approximate it. Do we know a function that has this curve? Robertson and Walker did [RW94]. They used the function

$$\left(\frac{d[j]}{k + d[j]}\right) \times idf \tag{1}$$

for some tunable constant k. This function behaves correctly in the "right-hand" limit - it increases monotonically to an asymptote at *idf*. While this may seem somewhat like a magic trick, it does work in practice.

Notice that this scoring function has the form *tf* × *idf*. Thus we have a theoretically (arguably) justified derivation of a score function that has both *tf* and *idf* terms. Cynics may complain about the hand-waving and the ad hoc assumptions that were made along the way. Nevertheless, this function has been used in practice and shown to work, and is a good example of practice motivated by theory, in which a relatively sophisticated mixture-of-Poissons model was found to support the *tf* × *idf* intuition.

4.1 Normalization

Recall that we have thus far assumed that documents have the same length due to our use of Poisson distributions. We still need some sort of normalization to account for varying document lengths. If we suppose that the constant k in (1) is "appropriate" for documents of average length, then we can use

$$\left(\frac{d[j]}{k \times \frac{\text{length of document } d}{\text{average document length}} + d[j]}\right) \times idf \tag{2}$$

4.2 BM25

There are several more sophisticated variations of this score function – some of them can been found in [Sin01]. Robertson et al [RWHB$^+$92] introduced a family of such scoring functions called

"Best Match ij" or "BMij" where $i, j \in \{0, ..., 9\}$, which differ slightly from each other in their parameters and constants. BM25, which stands for "Best Match, version 25", is the most famous of these functions, and very widely used. It is defined as follows:

$$\sum_{t \in Q,D} ln\frac{N - df + 0.5}{df + 0.5} \cdot \frac{(k_1 + 1)tf}{k_1((1-b) + b\frac{dl}{avdl}) + tf} \cdot \frac{(k_3 + 1)qtf}{k_3 + qtf} \tag{3}$$

Where

tf is the term's frequency in the document
qtf is the term's frequency in the query
N is the total number of documents in the collection
df is the number of documents that contain the term
dl is the document length (in bytes), and
$avdl$ is the average document length

The document length normalization term in BM25 is $k_1((1-b) + b\frac{dl}{avgdl})$. This looks a lot like pivoted document length normalization [SBM96]. Selecting the average old normalization as the pivot, the final normalization factor reduces to $(1.0 - slope) + slope \times \frac{\text{old normalization}}{\text{average old normalization}}$. While in pivoted document length normalization we use the old and average norm to get the new one, here we use the current and average document lengths.
k_1 controls the influence of tf on the overall function. As k_1 grows, the tf term in the function grows, influencing the ranking more. b controls length normalization. As b grows larger, the normalization depends more on $\frac{dl}{avgdl}$. Thus, the larger the value of b, the more we penalize documents of longer length. These parameters can be tuned based on properties of the corpus. b is usually set at 0.75 and k_1 is between 1.0 and 2.0.
The odd-looking last term with the qtf seems to be a result of treating the query as a document. The parameter k_3 plays the same role for the query as k_1 does for the document. If k_3 is higher, qtf (the frequency of term t in the query q) has greater influence on the ranking. We can thus tune k_3 based on how much importance we want to give to repetition of terms in the query.

4.3 Critique of the Probabilistic Model of IR

Let us now take a step back to evaluate the Probabilistic Model (PM) approach. The probabilistic framework is flexible. It allows for different levels of mathematical sophistication depending on how much information is available. For example, we were able to model term frequency in different ways (binary vs. following a Poisson distribution) depending on how much information we made available to the model (or made assumptions about). We were able to introduce new information if we had some knowledge about the relevance variable R_q through human evaluations, for example. We pay for this flexibility with the presence of more unknowns in the scoring function as the model gets more sophisticated. We also had to make a lot of rather precarious approximations to simplify the RSJ model.
Note also that the query q was not handled explicitly within the probabilistic framework. Our next approach differs from it in this respect.

5 Language Modeling Approach – the Third Framework of IR

This approach is due to [PC98]. The motivation here can be found in [LZ03], and involves explicitly modeling both document selection and query generation.
We define:

Q :	a random variable over queries	(based on the user)
D :	a random variable over documents	(based on the author/corpus)
R :	a (binary) relevance variable	

Rewrite the PM initial score function, i.e. $P(R_q = y|\vec{A} = \vec{d})$ as

$$P(R = y|D = d, Q = q) \tag{4}$$

Here we are treating R as a random variable, but it may be argued that, for a given document and query, the relevance of the document to the query is either "yes" or "no". We take the "variance", i.e. what makes this an interesting probability distribution (not just 0 or 1) to be due to the user(s). This differentiates the score function from attribute binning.
Now we can get Q and R on the same side of the conditional by performing a Bayes flip on D and R. This allows us to rewrite (4) as

$$\frac{P(D = d|R = y, Q = q)P(R = y|Q = q)}{P(D = d|Q = q)} \tag{5}$$

Notice that $P(R = y|Q = q)$ is document independent and can be taken out of the function. If we assume D to be independent of Q, we can replace $P(D = d|Q = q)$ with $P(D = d)$. We claim that this is a reasonable assumption, as it amounts to stating that the document generation/selection process is separate from query generation. We can now rewrite (5) as

$$\frac{P(D = d|R = y, Q = q)}{P(D = d)} \tag{6}$$

This is exactly what we had for our scoring function in the probabilistic model, i.e. $\frac{P(\vec{A}=\vec{d}|R_q=y)P(R_q=y)}{P(\vec{A}=\vec{d})}$ – just the notation has changed.
We could alternatively have tried Bayes flipping Q and R in (4) to get

$$\frac{P(Q = q|R = y, D = d)P(R = y|D = d)}{P(Q = q|D = d)} \tag{7}$$

Since Q is independent of D, $P(Q = q|D = d) = P(Q = q)$. So the above is equal under ranking to

$$P(Q = q|R = y, D = d)P(R = y|D = d) \tag{8}$$

We still do not know $P(R = y|D = d)$. Notice that this is simply the *a priori* probability that document d is relevant before we even see the query q. We are allowed to define this *a priori* probability distribution as we please. For example, we could set it to be a constant or assume that R and D are independent to make it disappear altogether under ranking. We could alternatively use document information to define it – relevant information might include document length, the PageRank or URL (if dealing with webpages), and the hub or authority score. For example, for

most queries, a document from the World Book Encyclopedia is more likely to be relevant than the musings of some unknown author on his or her blog.
The remaining term $P(Q = q|R = y, D = d)$ seems to be modeling the probability that query q was issued given document d and the judgment that d is relevant. In other words, we are considering the probability of the user generating q given that his/her information need is satisfied by d. This seems counterintuitive, and in reverse to what we have been doing so far – assuming the query to be given and choosing a document based on the query. We now seem to be modeling the user's information need in terms of the documents considered relevant rather that modeling documents in terms of relevance to the query.

References

[LZ03] J. Lafferty and C. Zhai. Probabilistic Relevance Models Based on Document and Query Generation. In Bruce Croft and John Lafferty, editors, *Language Modeling and Information Retrieval*, volume 13. Kluwer International Series on Information Retrieval, 2003.

[PC98] Jay M. Ponte and W. Bruce Croft. A Language Modeling Approach to Information Retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, New York, NY, USA, 1998. ACM Press.

[RW94] S. E. Robertson and S. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[RWHB$^+$92] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC. In *Text Retrieval Conference*, pages 21–30, 1992.

[SBM96] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted Document Length Normalization. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, New York, NY, USA, 1996. ACM Press.

[Sin01] Amit Singhal. Modern Information Retrieval: A Brief Overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.