Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
2/14/06: **Lecture 6 handout**

---

*This handout is a version that was updated on 3/7/06 for posting to the course homepage.*
**Announcement:** I've begun updating the course webpage with links to references and the like. See http://www.cs.cornell.edu/courses/cs630/2006sp .

**I. RSJ model** (Our version of) the Robertson and Spärck Jones (1976) scoring function (aka the RSJ model) for a document $d$:

$$\prod_{j:q_j>0,a_j(d)>0} \frac{P(A_j = a_j(d) \mid R = \mathrm{y})}{P(A_j = a_j(d))} \cdot \frac{P(A_j = 0)}{P(A_j = 0 \mid R = \mathrm{y})} \tag{1}$$

which is equivalent for ranking purposes to the log thereof:

$$\sum_{j:q_j>0,a_j(d)>0} \log \left( \frac{P(A_j = a_j(d) \mid R = \mathrm{y})}{P(A_j = a_j(d))} \cdot \frac{P(A_j = 0)}{P(A_j = 0 \mid R = \mathrm{y})} \right). \tag{2}$$

**II. Comparison of scoring functions** The table below is taken from Singhal, "Modern information retrieval: A brief overview" (2001), and shows state-of-the-art probabilistic-retrieval (in the classic sense) and VSM-inspired scoring functions. **Note** that there is a typo in the Okapi function: the $k_1$ in the denominator should be outside of the parenthesis.

| | |
|---|---|
| *tf* | is the term's frequency in document |
| *qtf* | is the term's frequency in 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 |

Okapi weighting based document score: [23]

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

$k_1$ (between 1.0–2.0), $b$ (usually 0.75), and $k_3$ (between 0–1000) are constants.

Pivoted normalization weighting based document score: [30]

$$\sum_{t \in Q,D} \frac{1 + ln(1 + ln(tf))}{(1-s) + s\frac{dl}{avdl}} \cdot qtf \cdot ln \frac{N+1}{df}$$

$s$ is a constant (usually 0.20).

Table 1: Modern Document Scoring Schemes

Reference [23] is Robertson, Walker, and Beaulieu, "Okapi at TREC-7: automatic ad hoc, filtering, VLC and filtering tracks" (1999). Reference [30] is Singhal, Choi, Hindle, Lewis, and Pereira, "AT&T at TREC-7" (1999).