Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
3/7/06: **Lecture guides for 2/7/06 (lecture four)**

Authors: Randy Au, Nick Gerner, and Blazej Kot (first seven pages); Peter Babinski and David Lin (subsequent pages).

Some additional questions for this lecture are below. (Not to say that there's anything wrong with the questions posed in the attached lecture guide; these are just some other thoughts I happened to have.)

1. As we discussed in class, Cooper (1995) argues that there is a problem with some independence-related assumptions that have often been used to justify the derivation of the RSJ model.[1] His argument may be paraphased as follows.

   Let X_1, X_2 and R be random variables taking values in $\{y, n\}$, where

   $$P(R = y) = P(X_1 = y) = P(X_2 = y) = .1 \quad \text{and} \quad P(R = y|X_1 = y) = P(R = y|X_2 = y) = .5 \quad (1)$$

   (a useful intuition check is to verify that such a situation is possible). Note that these conditions say nothing explicit about dependencies between X_1 and X_2.

   Now suppose we assume that *both* the following are true:

   $$P(X_1 = y, X_2 = y) = P(X_1 = y)P(X_2 = y) \quad (2)$$
   $$P(X_1 = y, X_2 = y|R = y) = P(X_1 = y|R = y)P(X_2 = y|R = y) \quad (3)$$

   Then, Cooper computes that $P(X_1 = y, X_2 = y) = 0.01$ given Assumption 2, but $P(X_1 = y, X_2 = y, R = y) = 0.025$ given Assumption 3 (be sure you can verify these calculations), which is a logical inconsistency (why?).

   Cooper then proposes that the following assumption be used instead:

   > (*Linked dependence*) There exists a constant α such that *both* the following hold:
   >
   > $$P(X_1 = y, X_2 = y) = \alpha P(X_1 = y)P(X_2 = y) \quad (4)$$
   > $$P(X_1 = y, X_2 = y|R = y) = \alpha P(X_1 = y|R = y)P(X_2 = y|R = y). \quad (5)$$
   >
   > (Cooper writes that "it may be helpful to think of [α] as a crude indicator of degree of departure from independence".)

   Note that *if* (the obvious generalization to many attribute variables of) the linked-dependence assumption holds, then the "factoring" that takes place in the usual derivation of the RSJ model is justified (why?).

   (a) Show that there is a scenario satisfying Cooper's example constraints (given in (1)) in which linked dependence is violated.

   (b) Does there exist a scenario satisfying Cooper's example constraints (given in (1)) in which linked dependence holds? (What would the implications be if the answer were "no"?)

   (c) Answer the analog of these questions for the different independence-related assumption we used in the derivation given in class.

References

Cooper, William S. 1995. Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems*, 13(1):100–111.

Robertson, Stephen E. 1974. Specificity and weighted retrieval. *Journal of Documentation*, 30(1):41–46.

[1] Cooper states that the anomaly presented was observed previously by Robertson (1974).