National University of Singapore
School of Computing
CS3245: Information Retrieval
Tutorial 5

IR Evaluation, XML Retrieval and Relevance Feedback

Readings: IIR Chapters 8 (excluding 8.5), 9 & 10

1. **Non-ranked Evaluation.**

   A related problem to information retrieval is *information extraction (IE)*, where automated programs pore over natural language text to try to identify certain key information, and extract it for use downstream.

   Let us use an example task of an IE system that identifies people and organization names from news text, such as from the Reuters newswire. In evaluating such a system we have three labels that can be assigned to a word: {(part of) person name, (part of) organization name, neither}.

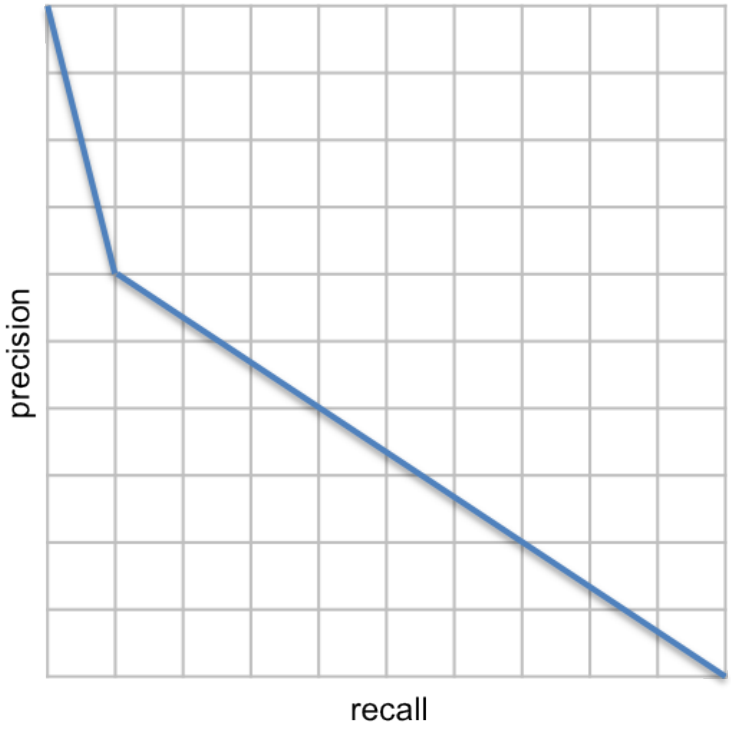
   (a) Two competing companies, A and B, are bidding for a contract to create a production system for the above task to showcase for further funding, and have produced two prototypes. *System A*'s claims 95% accuracy in its system's judgment, whereas B claims only 90% accuracy. Which is the better system?

   (b) Your friend, Shuhrakh Kan, states that he can build a system with 95% accuracy on the task, within a few minutes. Is he kidding? If not, explain how he might accomplish his task.

   (c) Your other friend, Sylvia Taronel, suggests that the evaluation metric of accuracy isn't appropriate. Is she right? She suggests using F_1 instead. Will that solve the problem?

2. **Ranked Evaluation.**

   (a) The below line plot shows the precision-recall curve for a ranked retrieval system. At what point is F_1 maximized? What about $F_{0.5}$? F_2?



   We have seen that interpolated precision curves generally look like a concave curve connecting the upper left corner (high precision, low recall) with the bottom right (high recall, low precision), such as the above plot in (a). Let's look into this in a bit more detail.

   (b) Do we always have a point at the upper left hand corner (that is, at the $y = 1, x = 0$) on every precision and recall curve?

   (c) Do we always have a point at the lower right (that is, at the $y = 0, x = 1$ point) on every precision and recall curve?

   (d) Is precision always a non-increasing function with respect to larger recall values on an interpolated precision-recall curve? What about on a non-interpolated curve?

3. **XML Retrieval.**

   In the VSM method for retrieving XML as introduced in the textbook we consider structural terms, which are paths that end in a single vocabulary term.

   (a) Should we also include paths that do not end in vocabulary terms at all?

   (b) Would it make sense to have queries that only give structural constraints and no content constraints?

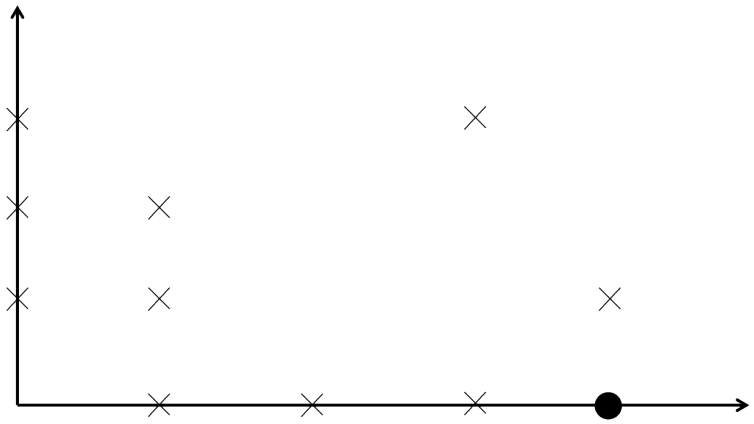
   (c) The matching method includes a calculation for *context resemblance*, which is a function that reflects how similar the two contexts are compared with each other. However, this is not a symmetric function: it states that contexts are similar only if c_q can be transformed to c_d by adding additional nodes. Why shouldn't this be a symmetric function? Can you give a counterexample where symmetry would be helpful, if possible?

   (d) Context matching, in the form of context resemblance, examines the path of the element to be matched. However, sometimes we may want to match elements that have structural constraints beyond the text to be matched. Give an example of such a query and describe how we could modify the context resemblance algorithm to cater for these types of queries.

4. **Relevance Feedback in Vector Space.**

   The below diagram show documents as 'X's in a two-dimensional vector space. A query given by the user is shown as a filled circle.

   (a) Show the ranking for the documents under the VSM for the original query, by numbering any relevance documents with their rank ("1" for 1st ranked, "2" for 2nd ranked, etc.) If two or more documents are tied in rank, use the same number; and if a document is not relevant, mark it with a "0".



   (b) Now say that the user has ranked documents marked as A and B as relevant. Given the modified Rocchio formula for determining the modified query as defined below:

$$\vec{q_m} = \alpha q_0 + (1-\alpha)\frac{1}{|D_r|}\sum d_j$$

   Show the approximate position of the new query $\vec{q_m}$, if $\alpha = 0.25$ Give the rankings of all documents with respect to $\vec{q_m}$, following the same instructions as in (a).