Project 2

Shun Zhang (sz4554)

October 9, 2012

1 Algorithms and Implementation

I didn't make any change to the original codes. Here are the classes I've added.

`ir.eval.ExperimentRelFeedback`. This inherits from `Experiment`. I mainly overrode the method `processQuery`, in which after the original retrieval, I immediately clear the first M documents, which are used for feedback. Then, I send the retrieval result and good document references (which we already know) to a `FeedbackSimulation` object. In this object, I set its `goodDocRefs` and `badDocRefs`, and run its `newQuery` method to get the revised retrieval result. This result is returned to the `ExperimentRelFeedback` object. Then, it is used to plot the PR curve.

`ir.vsr.FeedbackSimulation`. This class inherits from `Feedback`. I mainly edited its constructor, in which I set its `goodDocRefs` and `badDocRefs` attributes. So, the `newQuery` method can be called later.

`ir.vsr.InvertedIndexFactory`. I didn't find a good place to put the method of eliminating a `InvertedIndex` object's documents references which are used in training set - certainly it cannot be a method of `ExperimentRelFeedback`. So I make this class to hold the handler of a `InvertedIndex` object and do operation on it.

2 Questions

1. Does using feedback improve retrieval accuracy? Why or why not?

   Yes. Compared to the curve of N = 0 (N is the number of documents providing feedback), the precision-recall curves of other N values are outside this curve. So, when a precision value is fixed, when N increases, higher recall is achieved. Similarly, when a recall value is fixed, when N increases, higher precision is achieved.

2. How does the amount of feedback affect retrieval accuracy? Why?

   The more amount of feedback is used, the more accurate the retrieval is. As shown in the graph, when the value of N increases, the curve moves in a direction opposite to the axises.

3. Does using feedback affect the ability to achieve 100% recall (with ¿ 0 precision)? Why or why not?

   No. By returning all the documents in the corpus, 100% recall can always be achieved. Precision now is exactly $\frac{Number\ of\ positive\ results}{Number\ of\ all\ results}$.

4. Does using feedback noticeably affect retrieval time? Why or why not?

   It certainly affects the running time, but not in a noticeable way. Any way, using feedback means that we need to run at least two times of retrieval. We run the first time to get the initial result, then according to the feedback by human or a table to look up, we run the second time.