COMP90042
Web Search and Text Analysis (Semester 1, 2015)
Workshop exercises: Week 6

Discussion

Pre-workshop

- Revise "relevance feedback", from before the non-teaching period, and how it might be implemented as a kind of "query expansion".
- Revise the basic rules of probability, in particular, Bayes' Rule.

Workshop

1. Consider "relevance feedback":
   (a) What is the difference between "explict" relevance feedback and "pseudo"-relevance feedback? Give an example of an explicit relevance feedback mechanism.
   (b) The Rocchio algorithm is one mechanism for performing pseudo-relevance feedback. Consider the following set of documents:

`baby`	`crawl`	`swim`	`walk`	LEN
5	0	0	2	$\sqrt{29}$
2	2	0	0	$\sqrt{8}$
0	1	5	0	$\sqrt{26}$
1	1	1	1	2

   - For the query `crawl`, find the original document ranking according to TF-IDF, then apply the Rocchio algorithm on the top-ranked document with weights $(\alpha, \beta, \gamma) = (0.5, 0.5, 0)$, and observe how the document ranking changes.
   - Using the same query, consider the following parameter values $(\alpha, \beta, \gamma) = (0.8, 0.4, 0.2)$. Has anything changed?
   - Instead consider the top 3 results as relevant, and apply the algorithm with the two parameter settings above. How does the document ranking look now?

2. With a probabilistic IR model, we are trying to estimate the following probability: $P(R \mid d, q)$.
   (a) What does the above expression mean?
   (b) What assumptions do we need to make, in practice, to estimate this probability?
   (c) How do the different components in the final BIM model relate to a TF-IDF model?

3. Consider the (rather long) formula for BM25.
   (a) What do the variables $N, f_t, f_{d,t}, f_{q,t}, L_d, L_{ave}$ represent?
   (b) What are the parameters k_1, k_3, b attempting to control?

(c) Why isn't there a k_2 parameter? (Just kidding! :-))

Post-workshop

- Read through the derivation of BIM in the accompanying reading (IIR Chapter 11). It is quite detailed, and you don't need to understand every step. For the assumptions discussed above, identify where in the derivation they are applied and why.
- BIM can be used as a relevance feedback mechanism. Identify the step in the process where this happens, and what advantages it gives over the BIM model without the relevance feedback.
- BM25 is often the top-performing TF-IDF model on IR shared tasks. Why might this be the case?

Programming

Pre-workshop

- Implement a BM25 ranked querying engine. (Note that this should only involve changing the weighting calculations in your basic VSM model.)

Workshop

1. Given the above data set, try varying the parameters k_1 and b, each between 0 and 10, in increments of 0.1. Observe how the document ranking chances for the query above.

Post-workshop

- Just discussion this week.