Reviewer's report

Title: Prediction of relevant biomedical documents: A human microbiome case study

Version: 2

Date: 1 July 2015

Reviewer: Karin Verspoor

Reviewer's report:

The authors have improved the manuscript substantially and I believe I now understand the research itself far better based on this presentation.

It is notable that the main body of the manuscript is 9 pages long, and the specific research question that the authors investigate is not introduced in detail until page 6 (with the methods/results/discussion therefore forming less than 1/3 of the overall text). This balance seems off; while the broader context of IR systems and relevance feedback is interesting, I was left wondering, where are we going with this? At least, the paper needs a proper Introduction that introduces the basic research question and sets the context in the biomedical domain ("Biomedical searchers" are mentioned in a single sentence on page 2; although PubMed methods are discussed, all of the Background section is really generic to any "user" and generic information needs). Perhaps the section "A Role for Relevance Feedback ..." can be moved to the front of the paper to address this.

Major Compulsory Revisions

Given the authors' interest in supporting relevance feedback, and the discussion of the use of relevance judgements in an implicit feedback scenario, it would be meaningful to provide a learning curve (a plot of the performance of the system as a function of the number of training points -- i.e. a classifier trained with relevance judgements for only 10, 20, 40, etc. documents) for the performance of the classifier, showing the impact of more training data/relevance judgements. This allows to explore the question of how much feedback is needed to produce a meaningful relevance classifier. How much feedback (implicit or explicit) about a given query could you expect to receive in a realistic scenario?

Minor Essential Revisions

Results section: Please summarize the quantitative results of the experiments, rather than simply stating the algorithms "do (reasonably) well" -- cite actual numbers. 'Well' compared to what/by what relevant notion of 'good' results?

Could you please provide descriptive statistics on the user's judgements of the data sets (i.e., how many in each labelled data set are "yes", "no", "maybe"? This provides important context for interpreting the results of the 'mixed' classifiers that you have implemented (a confusion matrix, as I suggested in my original...
review, would be informative as well as provide more basis for understanding the impact of folding in "maybe" judgements with either "no" or "yes" judgements).

Typos:
PubMedCentral -> PubMedCentral
"noes" -> "nos"

Discretionary Revisions
I'm still not in love with the words "nos" "maybes" "yeses". These aren't very nice-looking words to read. Couldn't you use e.g., "... when "maybe" judgements are merged with "no" judgements". There is still one in Table 1 as well: "Maybes + no"

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:** I declare that I have no competing interests.