Reviewer’s report

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

Version: 1 Date: 16 August 2014

Reviewer: Karin Verspoor

Reviewer’s report:

The authors have described a study in which they explore the use of a relevance classifier to support an information retrieval task. They have identified a real user with a specific information need, and have trained a classifier on the basis of that user’s judgements. They have used a simple initial model, and applied standard machine learning strategies, with reasonable (but not great) initial results for their approach. It represents a step forward towards retrieval systems that incorporate user-specific judgements of relevance. While the manuscript describes methods that are appropriate, some restructuring would improve the clarity of the manuscript, and some additional experiments and discussion would enhance the contribution of the research.

__Major Compulsory Revisions__

The underlying assumption of the classification experiments presented in this manuscript is that there is a single model of relevance for a given domain or task (e.g. retrieval of microbiome articles); the authors build a model that aims to determine the relevance of a document in the absence of a specific query. This contrasts with the normal notion of query-relevance feedback that is explored in much of the information retrieval literature, where it is the relevance of a document to a particular query that is scored. The authors have not identified this distinction, nor justified their choice. Given that the authors have explored six queries (or possibly five, as it appears one was discarded as too similar to another query), it would be reasonable to ask whether this generic notion of relevance is consistent across the queries. It could (and should) in fact be directly tested by training the model with the results of 4 of the queries, and testing on the 5th, in a cross-validation scenario that would reveal possible differences between the notion of "relevance" for each query.

The authors should discuss the generalisability of their approach to real-world searching contexts, since they specifically wish to support real users with real information needs. How would the use of a relevance classification model be applied in practice? Presumably a different classifier would need to be trained for each information need. Is this practical or realistic?

The authors have not adequately explained their experimental context. I assume, though it is not stated explicitly, that they train binary classifiers, and that the categories they refer to, such as "maybes to yes" imply a binary (yes/no)
classifier for the "maybe + yes" class, i.e., a classifier which decides whether or not a document is a member of the "maybe + yes" class. (In contrast to the alternative of a multi-label or multi-class classifier). The authors should more clearly explain this. Figure 1 would be more helpful if it truly captured the experimental context and the classification scenario, e.g. how the data was organized and split for the study, rather than what it currently does, which is to summarize the processing tasks (incompletely, as it does not include the PubMed and full text processing, or document transformation steps). The figure states "Run Weka" on "Features extracted from Documents" but it does not define what labelled data or classification task Weka is being asked to learn. The text in the methods section also does not explicitly address this, just stating "our binary categorization results". Remind the reader what the categorization task is and how you structured it.

Assuming that I have understood this experimental context correctly, why have the authors not built 3 binary classifiers, one for each class "yes", "maybe", "no"? Then a confusion matrix could be constructed to more directly explore the overlaps between the three classes and more specifically explore the issues around the 3-way classification that based on their current results appear to be somewhat unresolved.

__Minor Essential Revisions__

In the results, the authors state "these results are worse than random". Can you demonstrate this directly? What is the distribution of the documents in the gold standard across yes/no/maybe? This is needed to establish "random" or baseline performance.

Some references to foundational summaries of IR methodology (e.g., Manning et al 2008), and biomedical IR (Lu 2011, http://database.oxfordjournals.org/content/2011/baq036.full) should be added. Note the latter refers to the RefMed and MiSearch systems which aim to incorporate relevance feedback.

__Discretionary Revisions__

In the Background section, the authors mention that "metadata ... allow[s] documents to be retrieved". The metadata itself doesn't allow retrieval per se. It allows documents to be retrieved more precisely, or on the basis of general characteristics, or something else ... i.e. it impacts how retrieval might happen, not that it is possible in the first place.

An information need is defined as "a mental state representing a gap in the searcher's knowledge". There are other information needs, e.g. re-finding (known item search). The authors may wish to be somewhat more explicit.

Some more minor issues:
* Where the "six PubMed queries" are mentioned, Appendix 1 where they are listed should be mentioned.
* The second-to-last paragraph of the Background could benefit from re-structuring. The first sentence introduces the task. The second section introduces the data representation. Then the goal of the study is introduced. This seems to mix different levels of content; I suggest the authors consider moving the bag-of-words/document representation discussion elsewhere.

* The Methods section could benefit from small structural changes for clarity. The paragraph starting "For this study ..." could be split before "To assemble ...". At that point, the focus of the paragraph shifts from data collection to processing of the documents. The final paragraph "Most research ..." does not fit in the Methods but rather should be moved to the Discussion or Future Work.

* Please rephrase "maybes" and "yeses" as "maybe" judgements' and "yes" judgements' for clarity. Note headings such as "Maybes to yes" are hard to interpret.

* Note typo: PubMedCenteral -> PubMed Central

**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.