Dear Emine, Jin, Paul,

This took a bit longer than planned. I found myself doing a more detailed job than we usually do at this stage. This is probably a reflection of the degree to which we care about this topic: we already have three strong evaluation issues and we want the next one to be at least as strong. Of course, your extended outline is only a partial draft, meant to elicit comments… and that’s what it did :-).

There is a lot to like about the draft: the topic, the writing style, the way in which you have organized the material. But there are also many comments, which I have listed below. I have attached a PDF with further notes. I hope this helps. Let me know if you’d like to have a skype meeting.

Maarten

o-o-o-o-o

General:

• Scope statement: the title "offline evaluation in information retrieval” promises more than you aim to deliver. "Offline evaluation of document retrieval" would be more appropriate. Even more appropriate would be “Offline evaluation of document retrieval using human-generated labels”: you are excluding offline evaluation based on interaction data/implicit feedback.

• Implicit in the survey is a fairly narrow concept of what information retrieval is (document retrieval) and of what the basis for offline evaluation is (human generated labels). Both are fine, but your language usage suggests a broader scope.

• Explain how your survey relates to Mark Sanderson’s survey and why we need your survey now.

• While most of the document was written clearly, I found the language usage to be rather vague at various spots. One aspect of this is that some of your statements are made without including arguments or references to back them up. Another is that there are various implicit assumptions that you are making; I mentioned the narrow concept of information retrieval already, but there are others too. E.g., at some point in Ch2, scalability suddenly enters the discourse but you never tell us how many judgments are needed or for how many query document pairs. A third aspect is a certain level of sloppiness in choice of vocabulary such as on p14 where you talk about "new endpoints and scenarios” and then refer back to those as “these services”; the definitions of your key concepts (search task, experiment, etc) aren’t as precise as they should probably be.

• The relation between online and offline evaluation needs a more thorough treatment. The discussion would have to include correlations between offline metrics and online metrics.

• The narrative structure is very good at points, e.g., in Section 1.4, we are offered a very natural structure that actually helps organize the rest of the survey, but it is not always optimal. For instance, in 1.2.2. you sketch recent trends. I agree that it is very useful to have this section early in the survey as it helps explain why your survey is timely, but instead of first listing a number of issues and then listing (as a group) a number of recent publications that address these issues as you do now, I suggest you use more of a “problem, recent advances, plus a pointer to later in the survey in case you discuss this later” kind of structure.

• My most important problem with the current draft is that the survey is not set up yet to deliver on its promise of providing a how-to for setting up an offline evaluation pipeline. In section 1.5 you promise that you "describe each process of offline evaluation in detail so that a reader can design his or her own evaluation pipeline following the flow of this paper”. Each subsequent chapter is then meant to detail one key ingredient of this pipeline. However, what we see in Ch2 is simply too abstract/generic in my view to count as a how-to or as an account of the best practices. We get many general statements, often without references and certainly without technical details or critical assessment of pros and cons. I am not convinced that works as a how-to.

• There is work on the difference between individual and consensus judgments. John Wilbur had a paper on this about 18 years ago with some experimental data that's a useful reference, but of course the issue arises most saliently these days with crowdsourcing. The short version (at least as I understand it) is that individual relevance and consensus relevance are different things, and to the extent that's true it is a point worth making because its easy to implicitly assume otherwise. Likewise, there is some older work by Bill Hersh on the difference between improvements measured in a TREC style setup and perceived by users (even in a user study); that was a paper from the early 2000s. Again, that’s probably a paper you should mention.

Chapter 1

• Chapter 1 about the relation between evaluation paradigms is potentially very useful, but it will need a more thorough discussion and analysis in places instead of merely listing options.

• Section 1.1 is not always as balanced as it should be. E.g, you are right that clicks are not necessarily an indication that a user found a document relevant but neither are expert judgments. I missed a more thorough discussion of the relation between e.g. offline and online measurements and of how differences observed through test collection based measurements do not necessarily imply that you will see the same differences in user studies or online experiments. In general, I would like to see a more thorough discussion of the sequence test collection-based — user studies — user panel — online experiments plus a clear indication of what the ultimate signal is that you believe we should be optimizing for. Surely, that is not expert generated labels.

• Section 1.2 needs a better grounding of statements in the literature, with proper references. It also needs to provide better definitions of key concepts; see my annotations in the PDF.

• Section 1.3 is problematic: what exactly are the scenarios that the section title promises? Some of the contrasts that you are trying to make simply don’t seem to exist (see my comments in the PDF).

• Section 1.4 is really important and very useful, especially 1.4.2. The definitions need sharpening. “User models” appear somewhat unexpectedly. Either drop or integrate better, would be my suggestion. Also, shouldn’t there be a “making inferences from experiments” item somewhere? In my understanding that’s not the same as “How do you evaluate the validity of the outcome?"

• Section 1.5 provides a look ahead, which is useful. What I found a bit surprising is the apparent delta between the lessons we are ging to learn in Ch2-4 on the one hand (“offline evaluation pipeline”) and Ch5 on the other (“evaluation practices from major companies”). Maybe it is just a matter of subtle differences in focus (between Ch2-4 on the one hand and Ch5 on the other), but if the industrial practices don’t follow the pipeline outlined in 1.4, shouldn’t you come up with a different pipeline? And shouldn’t the best practices be discussed with each of the stages of the pipeline, i.e., in Ch2-4?

• Layout/formatting: please use proper latex tables instead of screen dumps of word tables.

Chapter 2

• What exactly is the overall rhetorical goal you are trying to achieve in Ch2? I thought it was to provide a how-to on collecting relevance judgments. However, the chapter needs some work before it succeeds at this. We need to know what the different options are (as you sketch them now), but we need to see more technical details and we need a more explicit story about the implications of the different choices that your readers would be facing if they use this chapter as a how-to. On the whole, it feels a bit too much like you’re just listing things and you’re not doing enough explaining why/when something works and what would be best practices. This obviously limits the potential usefulness of your survey.

• Section 2.1: it would be good if the instructions in 2.1.1 and 2.1.2 could be more hands-on and if they would actually tell the reader how to go about search task creation and/or query sampling. It’s all a bit vague right now and the discussion of the different choices is abstract, not backed by any theoretical or experimental details on how important the differences between the different choices are.

• Section 2.2: this is concerned with judgment interface design, which you split into four questions (at the top of p24), some of which end up in a \subsection{}, some don’t. As before, it is important to tell your readers about the choices they need to make when setting up an offline evaluation pipeline, but the text is too abstract. Also, you need a good discussion about what global criterion it is that you are trying to optimize for, probably already earlier than in Chapter 2 (i.e., relevance vs utility vs satisfaction).

• Section 2.3: what are the scalability issues that suddenly appear here? Please provide details. Why is there no subsection on “Expert labeling” next to the subsection on “Crowdsourced labeling”?

• Section 2.4: I have mixed feelings about this section. Should it be here, as part of a how-to? I agree that the issues raised in 2.4 need to be discussed in the survey, but perhaps in a separate chapter towards the end of the survey? Also, you give so few details about the issues listed that it is hard for anyone who is not already familiar with them to learn much.

• Missing: a look ahead, i.e., a statement on how the choices made in this chapter (setting up the label collection) affects the next two stages in the offline evaluation pipeline, and vice versa.

Chapter 3

• We probably do not need a discussion of the basic metrics; they have been covered extensively in many other places. Something insightful about when to use what + best practices is probably a lot more useful than a list of many metrics compressed down to a few pages.

o-o-o-o-o

---

Maarten de Rijke

University of Amsterdam