CS 630 Lecture 14: 03/14/2006

Lecturer: Lillian Lee Scribes: Asif-ul Haque, Benyah Shaparenko

---

1 Clickthrough Data (CTD) as Implicit Feedback

In a study by Shen, Tan, Zhai '05, it was shown that clickthrough data is much more effective than query history (at least with respect to the models they proposed) in determining the user's information needs. Interestingly, this was true even though only one-third (29 out of 91) of the documents that the clicked summaries linked to were relevant. So the obvious question that comes to mind is, why is CTD so effective even though such a small fraction of the clicked summaries corresponds to documents that were relevant? One explanation could be that the process is highly noise tolerant and even a small amount of accurate information is enough for relevant documents to come up. A more realistic explanation would be that the summary relevance is not correlated with document relevance. To prove this claim the authors retrospectively removed the summaries of documents that were relevant from the CTD and checked the performance. The results showed that there were still some improvement although not as much as while including those summaries. If the summary relevance had strong correlation with document relevance and the improvements were due to this correlation (rather than the effectiveness of CTD), there should have been no improvement when the clicked summaries linking to relevant documents were removed. So the summary relevance is presumably not correlated with document relevance.

2 CTD as a source of Implicit Feedback (IF)

In another study by Joachims, Granka, Pan, Hembrooke, and Gay '05, the authors questioned how we can interpret CTD accurately. They tried to understand if clicked summaries always meant summary relevance or is there some bias originating from the reliability and presentation of the search engine that makes CTD an unreliable source of implicit feedback. In the experiment, a number of subjects were asked to find the answer to one of ten selected search tasks. They used Google as the search engine. There were also independent judges who ascertained the relative relevance of all the summaries as well as the corresponding documents for those search tasks. (The assumption here is that human ability to judge relevancy is very accurate. In $80 - 90$ % cases the judges agreed on the relevance of pairs of summaries and documents). Eye tracking was used to get data on users' viewing patterns of search results. Tests were also performed where the subjects were provided with the search results from Google with the top two documents swapped or the ranks of the first ten documents reversed. This was done without the subjects knowing or realizing it; the goal was to discover whether any presentation bias effects exist.

2.1 Findings

- The users do look over the page of search results from top to bottom. (Note: However, in this study, advertisements and other information besides the search results were removed from the results page that Google generated before the pages were presented to the users.)

- The first summary, s_1, and the second summary, s_2, were mostly both viewed (though not necessarily clicked on). Summary s_1 was looked at 68% of the time. Summary s_2 was looked at 61% of the time. The time spent viewing s_1 and s_2 was roughly equal. More importantly, what should jump out at you is that even the 68% number is very low. This number says that quite a bit of the time, the user never even looks at the first search result (at least as measured by the particular eye-tracking equipment in question).

- Do you think summaries below a clicked-on summary are viewed? As it turns out, 50% of the time, the summary right below the clicked-on summary is viewed. But then, 50% of the time, it is not viewed.

- Summary s_1 gets the first click 40% of the time, and s_2 only gets it 15% of the time. Since we also have the eye-tracking results to see exactly what the user looked at, we know that the users looked at both s_1 and s_2, but purposely selected s_1.

2.2 Presentation Bias

Perhaps s_1 gets most of the clicks because of a presentation bias meaning that since s_1 is presented before s_2 by a "reliable" system like Google, people may have the tendency to click on the first document irrespective of the relevance of the summary. To test this hypothesis, we look at the case where exactly one of s_1 and s_2 were clicked. If s_1 was more relevant, it was clicked on 19 out of 20 times. If s_2 was more relevant, it was clicked on only 2 out of 7 times. So, there is a presentation bias.

However, notice that s_1 was more relevant 20 times, while s_2 was more relevant only 7 times. The most likely possibility is that Google is doing a good job, and the more relevant of s_1 and s_2 appears first most of the time. However, in the few cases where this is not the case, perhaps this indicates an odd situation, such as where s_1 and s_2 are very similar, so that the CTD still isn't "incorrect".

To remove the "odd situation" possibility as a potential explanation, some subjects (unknowingly) were shown pages with s_1 and s_2 swapped (so that in this setting most of the time the second link presented is truly more relevant than the first). As it turns out, users still prefer clicking on the top link, though not as much as before. The conclusion is that CTD for IF is probably problematic unless there is some way to compensate for the presentation bias.

3 "Preference" Information vs. Relevance

We will now move away from the assumption that a click implies that the summary is relevant. Work by Joachims in '02 proposed that a click gives "preference" information instead of relevance. CTD will only be used for relative judgments between summaries in this scenario.

3.1 Relative Judgments

Assume that ordered document summaries are returned to the user: s_1, ..., s_i, ... s_j. Suppose the user clicks on the summary s_j, but does not click on the summary s_i. In this case, we assume that the user's click represents the fact that s_j is more relevant than s_i. Presentation bias is not a factor in the quality of such relative judgment inferences: if users make it past the earlier summary s_i to

click on the later summary s_j, then they have overcome the presentation bias. However, presentation bias is likely to reduce the amount of data available.

3.2 Using Relative Judgments

A naive idea (but relatively effective) is that every clicked link should be ranked above the previous unclicked links. A better idea is to do the same inference, but only if summary s_j is *temporally* the last clicked summary.

To make this idea even more useful in practice, we will try to convey RF across users and queries (since the lack of such generalization seemed problematic in our previous discussions of feedback-based approaches). However, there is a problem. If we have this preference information, how can we use RF in the probabilistic retrieval model? Recall that we had probabilities of the form $P(A_j = 1|R = y)$. So, we now have quality feedback that the methods we have presented so far don't know how to use.

4 Problems

1. In the preference ranking scenario, what is an argument for ranking only the temporally latest clicked summary above the previously unclicked links? What is an argument for ranking all the clicked summaries above the previously unclicked links?

2. Personal clickthrough and summary investigation exercise: When searching the Internet today, pay attention to the links that you click on. Instead of just clicking on a link, think about why you might have clicked that link. Was it the summary information, the title of the website, or the position on the page? Read all the summaries on a page and think about whether you would still have selected the same summary as the first one you clicked on. Pay attention to the summaries that you click on and count how many are relevant to your query. What conclusions can you draw about the quality of clickthrough data based on your personal experiences?

Answers

1. An argument for selecting only the temporally latest clicked-on summary is that the user's information need was not satisfied until the user clicked on this summary, otherwise the user would have stopped searching earlier. Therefore, the latest summary is relevant, while the previous clicked-on summaries probably are not relevant. An argument for ranking all the clicked-on summaries above the previous not-clicked summaries is that while the documents underlying these summaries may not have the complete answer, their summaries still look better than the other summaries, since the user clicked on these summaries while bypassing others.