CS674/INFO630: Advanced Language Technologies, Fall 2007, Lecture by Lillian Lee
Lecture 16 Guide: An Introduction to Empirical Evaluations of Clickthrough Data

Alex Chao David Collins

October 25, 2007

1 Introduction

In this lecture we will be examining clickthrough data as a form of relative implicit feedback. Recall that typical search engines, for a given query, return a ranked list of document links to the user, which are further paginated so that only the top ten results are displayed on the first page of results. Often paired with each link is a summary representative of the document. For the purpose of this discussion, the link itself can be considered part of its accompanying summary. The user may click any number of the document links on a single page, and these clicks constitute the clickthrough data in which we are interested.

Recall from the previous lecture what Shen, Ten, and Zhai (SIGIR '05) found to be very valuable to an information retrieval system. Of the two types of implicit feedback that they analyzed (query history and clickthrough data), they found clickthrough data to be more valuable [4]. This finding should motivate us to examine its value in closer detail, as should the mere abundance and overall frequency of clickthroughs in web search today. In general, we want to know how clickthroughs correlate, if at all, with positive relevance feedback.

2 An Experiment

2.1 Setup

Our discussion today will be primarily based on a study conducted at Cornell University by Joachims, Granka, Pan, Hembrooke, and Gay. They authored a report that was published in SIGIR, in August of 2005 [1]. The lecture will also be based on a longer report authored by the same people in addition to Radlinski. This was published in the April 2007 edition of ACM TOIS [2].

In a 2005 experiment conducted by Joachims *et al.*, human subjects were asked to complete predetermined web search tasks using Google. In this study, no pre-specified queries were given to the test subjects; the users were asked to formulate their own queries. Google is both a convenient and practical choice for this study, given that it is ubiquitous in Internet search today.

Before drawing any conclusions about the correlation between clickthrough data and positive relevance feedback, it is important that we remember that the users in this study clicked on links based on the summaries with which they were presented (not based on the documents themselves). We can therefore only realistically think of these clicks as representing relevance feedback for these summaries.

The information retrieval tasks given to the subjects were classified into two types: *navigational* and *informational*. A navigational task involves trying to find a particular website, image, or other such resource, while an informational task involves trying to find some particular bit information, such as a historical fact.

In the Joachims experiment, an entirely different cast of participants, called judges, performed relevance evaluations of the summaries. In particular, for each results page that a subject saw, the ten or so summaries on that page were presented to each judge in some random order. It was the judges' job to rank each page of summaries in order of relevance. The judges were, in effect, giving a partial relevance ranking. The relevance ranking is partial because the judges were allowed to rate two or more summaries as being equally relevant.

2.2 Discussion of Setup

One may ask why the experimenters chose to have the judges rank all of the summaries on a results page rather than have them rank only those documents that were clicked. The primary reason for this choice is that they wanted to evaluate decisions not only to click on a link, but also to *not* click on a link.

Another question one might ask is why the the judges were tasked with ranking the documents rather than making absolute relevance judgments for each of the summaries. The reasons are two-fold. First, one can argue that it is easier for the judges to rank the documents rather than make binary judgments about every document. Second, if we know a ranking threshold above which a summary is determined to be relevant, then the ranking ultimately provides a binary classification anyway.

One may ask why the searchers themselves were not simply assigned to be judges. The main reason for this is that the judges are simply not in the same situation as the searchers. A searcher attempting to rank the documents might have developed presentation and scenario-related biases after having seen Google's ranking on a particular results page. A searcher is also performing an inherently different procedure than a judge, by searching for answers or resources rather than trying to annotate the data. Finally, a searcher attempting to rank the documents would be giving explicit relevance feedback, which defeats the purpose of the experiment to study implicit relevance feedback.

The researchers wanted to quantify any biases on the subjects' behavior from the ranking returned by Google. They considered two different cases:

- The user is biased by Google's ranking (or by the user's own trust in Google) and considered most summaries. In this situation, we want to know whether the user superseded Google's ranking suggestion with self-proclaimed ideas about relevance.
- The user is biased by Google's ranking and did *not* consider most of the summaries. In this case, the searcher has little, if any, intuition that could overcome Google's ranking suggestion.

In order to determine if and how much the searchers were being influenced by Google's presentation (or by their trust in Google), the experimenters secretly altered Google's presentation of the results with a proxy that would change Google's rankings and, for simplicity, remove advertisements. After the experiment, the searchers indicated that they did not notice anything unusual about the experience.

As an aside, this setup allows for a comparison between the judgments of an IR system and those of the users (essentially, a *what-Google-thought vs. what-people-thought* paradigm), though it should be noted that Google's ranking is not based on the provided summary, but rather on the content of the document itself (an observation noted in class by Nikos Karampatziakis). For each results page returned by Google, the experimenters compared the ranking of the summaries returned by Google with the ranking supplied by the judges. Let s_i represent the i^{th} ranked summary on a particular results page returned by Google. This means that s_1 and s_2 would be the summaries corresponding to the documents ranked first and second by Google, respectively. In one experiment, of 85 results pages, the judges found that the relevance of s_1 was greater than the relevance of s_2 36 times (roughly 42% of the time). This number may be alarming until one considers a second finding, in which the relevance of s_1 was found to be less than the relevance of s_2 20 times (roughly 24% of the time). Thus, about 76% of the time, s_1 was either more relevant than or as relevant as s_2.

2.3 Findings

In order to take measurements of what part of a results page a searcher was reading at any given time, the experimenters fitted the the subjects with eye-tracking equipment. From the analysis of the eye-tracking data, the experimenters found, not surprisingly, that the searchers tended to analyze a results page by starting at the top summary and working their way downward. They also found that when searchers were presented with a results page, they tended to look at s_1 and s_2 almost immediately, with a noticeable pause before looking at s_3. In general, summaries with a better Google ranking were more likely to be read. The experimenters also found that 45% of the time, the searchers viewed the summary above and below the one they clicked. Other findings are as follows:

- s_3 was viewed less than 50% of the time.
- s_1 was viewed about 70% of the time and was clicked 40% of the time. One should find it odd that s_1 was viewed considerably less than 90% of the time, and, in fact, the experimenters concluded that this statistic was due to error in the eye-tracking equipment.
- s_2 was viewed about 60% of the time and was clicked 15% of the time. The click rate here is much smaller than that of s_1 and begs the question of whether this disparity denotes a rational decision on the part of the user. Such a decision would be rational if Google, over the course of the experiment, correctly placed the most relevant document above the second most relevant document.

In regards to using clickthrough data as a source of feedback, there are several potential reasons why, when we analyze the clickthrough data for a particular results page, we cannot interpret the absence of a click on a summary as negative feedback. First, the searcher examines the items top down, so it is very possible that an assigned task will be satisfied

before all of the summaries on results page are examined. Second, the user might trust Google's ranking in a way that tends towards picking documents closer to the top.

In order to determine how Google's presentation of the summaries influenced the searchers' behavior, the experimenters secretly swapped the first and second-ranked summaries returned by Google. They found that the searchers tended to prefer the new top summary, even after this swap, showing that the searchers' decisions were, in fact, being heavily influenced by Google's presentation. It is important to note that the experimenters were not testing why the searchers' behavior was changed by Google's presentation, but rather how it was changed. This part of the study very clearly showed that Google's ranking heavily influences user behavior, which is an observation made especially salient by the fact that the subjects did not notice anything wrong or different about their experience.

It follows from these findings that the utility of implicit relevance feedback can be improved if one removes those elements of presentation that are prone to biased judgments.

3 An Alternate Interpretation

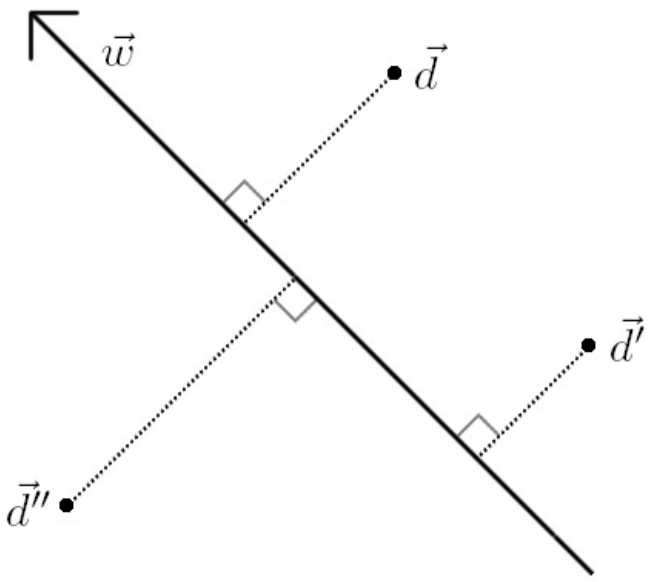
This section is based on a 2002 study published in SIGKDD in which Joachims examined clicks on summaries not as binary feedback, but rather as relative judgments [3]. Joachims reasoned that if s_1 is not clicked but s_2 *is* clicked, then we can reasonably conclude that the relevance of s_2 is greater than that of s_1. More generally, if the i^{th} summary is not clicked, and the $(i+k)^{th}$ summary is clicked, then we can conclude that the $(i+k)^{th}$ summary is more relevant than the i^{th} summary, because the searcher consciously overcame any bias toward Google's ranking and presentation. However, note that if the i^{th} summary was clicked and the $(i+k)^{th}$ summary was not clicked, we cannot conclude that the i^{th} summary is more relevant than the $(i+k)^{th}$ summary, because this may very well be a result of Google's presentation.

A major downside of taking this approach to extracting relevance data from clickthrough data is that a lot of the data goes unused. In particular, a click only provides information about the relative relevances of other links above the clicked link; no conclusions may be drawn from the absence of clicks below the clicked link. On the other hand, a major advantage to taking this approach is that the information that is gained is believable, even in light of presentation and scenario biases.

3.1 Extracting Pure Relevance Data from Preference Data

One may now be wondering what one can do with this preference data, since it only provides relative relevances, or preferences, and does not seem to exactly translate into pure relevance data. Joachims proposes an approach to this problem that incorporates the pairwise orderings of summaries as special relationships, given a query q.

Suppose, for example, that $rel_q(d) > rel_q(d')$ and $rel_q(d) > rel_q(d'')$, and suppose that each of these documents is represented by a point in space as follows:

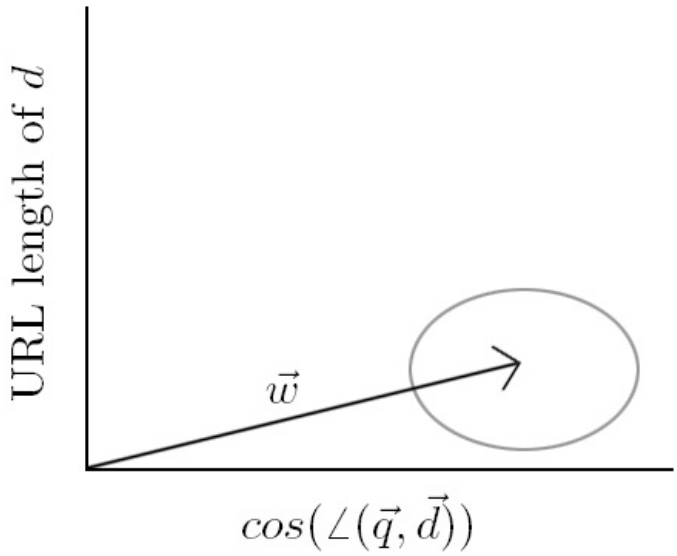


The vector $\vec{w}$ in the figure is chosen such that the orthogonal projection of each of the points onto $\vec{w}$ respects the order of preferences as much as possible. However, this procedure assumes a single query q, and so a problem arises once a new query is introduced. Namely, we must find a whole new $\vec{w}$, since we cannot generalize across all queries.

In order to deal with the new queries, we must treat q as part of the input. We put the query characteristics into feature vectors that may, for example, encode the closeness of the query to the documents in geometric terms. In such a case, we might define $\vec{\phi}_{q,d}$ such that $\phi_{q,d}[i] = cos(\angle(\vec{q}, \vec{d}))$. An alternate and perhaps more crude example would be the following:

$$\phi_{q,d}[i] = \begin{cases} 1 & \text{if } d \text{ is a Wikipedia page} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In this latter case, the feature does not explicitly take $\vec{q}$ into account. Indeed, the features could be of both $\vec{d}$ and $\vec{q}$ or simply of just $\vec{d}$. Consider another type of feature: the URL length of the document. To decide on an appropriate $\vec{w}$ in this case, we might plot two features against each other in the following way:



The region enclosed by the ellipse represents the feature set where documents have relatively short URL's and are close to the query vector (i.e. large $cos(\angle(\vec{q}, \vec{d}))$). This is a reasonable place to put $\vec{w}$, as the query-document vector $\vec{\phi}_{q,d}$ of a document that exhibits these characteristics will project "high" onto $\vec{w}$, thus tending to preserve the preference relations. This is one way in which carefully selected document and document-query features can be combined with implicit relevance feedback, in the form of user preference relations.

4 Questions

4.1 Question 1

The judges in Joachims' experiment were giving relevance feedback for summaries instead of the documents themselves. We are interested in evaluating the relevance of documents. How do we reconcile these seemingly conflicting ideas?

4.2 Question 2

The following question was posed in class: with the rise of sites like Wikipedia, would Joachim's 2005 experiment have been practical to perform in today's world? Think about what precautions the researchers might have needed to take.

4.3 Question 3

Although the judges used in Joachims 2005 experiment were different from the subjects, they were not specially trained relevance judges (see Joachims [1, p.156]). Why would it not have been advantageous to use specially trained judges?

5 Sample Answers

5.1 Answer 1

We cannot simply assume that relevance feedback for summaries implies relevance feedback for the documents themselves. In their 2005 experiment, Joachims *et al.* not only asked the judges to give relevance feedback for the summaries on the results pages, but they asked them to give relevance feedback for the corresponding documents. The experimenters compared the preferences derived from clickthrough data with the explicit relevance feedback for the summaries, and they compared these same preferences with the relevance feedback for the documents corresponding to the summaries. They found that there was only a 3 percent average drop in performance from the first comparison to the second comparison. This was used to show that the clickthrough data could, in fact, be reasonably used as implicit relevance feedback for the actual documents (see Joachims *et al.* [1, p. 160]). This conclusion may be surprising when one considers that Shen, Tan, and Zhai found that approximately thirty percent of all clicked-on summaries were for non-relevant documents in their 2005 study (which used a TREC data set in which a binary relevance judgment was made for each query and document pair) [4, p. 47]. However, it is important to note that Joachims *et al.* were interested in the relative relevance among documents, rather than any binary relevance evaluation of the documents for a given query. In the study by Shen *et al.*, perhaps some of the nonrelevant documents corresponding to clicked summaries were still more relevant than the non-relevant documents corresponding to better-ranked summaries that were not clicked.

5.2 Answer 2

The prominence of Wikipedia today certainly would have added a few complications to the experiment. The issue is that searchers have prior knowledge about the features of Wikipedia. This could potentially make the clickthrough data somewhat noisy, because it seems very possible that users would select a Wikipedia summary in the results page without thoroughly studying the relevance of the summaries above. It is even possible that the searcher would enter a navigation query, by typing 'Wikipedia' in the Google search box to reach the Wikipedia homepage, where they would then attempt to perform their initial informational or navigational search task.

One potential way the researchers could account for Wikipedia is to be extra careful about the types of informational and navigational search tasks they give the subjects, trying to only give tasks that Wikipedia likely would not handle well. If the researchers do this, though, it is possible they would not be observing real-world behavior for some of the more common types of search tasks. Another possible solution would be to use the same proxy, the one that changed Google's ranking of the summaries and removed the ads, to block any Wikipedia results. Again, it would be important that the subjects not observe anything wrong or different about their experiences.

Another potential way to to account for this bias toward Wikipedia, assuming it does, in fact, exist, is to mitigate it in a fashion similiar to how Joachims handled the Google presentation bias. For instance, suppose a Wikipedia summary s_a is earlier in the ranking of Google's results page than a second non-Wikipedia summary s_b. If s_b is clicked and s_a is not clicked, then, as in Joachims' experiment, we can reasonably conclude that s_b is more relevant than s_a. On the other hand, suppose that the Wikipedia entry s_a is later in the ranking of Google's returned results than summary s_b. If s_a is clicked and s_b is not

clicked, then we would not conclude that s_a is more relevant than s_b, because of the user's assumed bias toward clicking on Wikipedia summaries. Other than this type of example, the relevance preferences can be extracted from the clickthrough data in the same manner as in Joachims' study. Relevance preferences, for instance, would still be extracted between two Wikipedia summaries or two non-Wikipedia summaries since we do not need to account for a bias toward Wikipedia in either case.

As Wikipedia is very prominent in today's world of internet search, it does not seem appropriate to avoid it when studying search. In fact, it would seem to be very valuable to thoroughly study just what effect Wikipedia has on search. However, when empirically testing clickthrough data as implicit relevance feedback, it was probably most appropriate to treat Wikipedia entries separately, since it does seem that users may have developed a trust in Wikipedia entries.

5.3 Answer 3

Joachims *et al.* reasoned that since their subjects, the searchers, were everyday users, they should have the relevance evaluations be everyday users as well. This ensures that the clickthrough data and explicit relevance assignments are both are compatible and similar to what they would observe in the real world (see Joachims [1, p. 156]).

References

[1] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. *Accurately interpreting clickthrough data as implicit feedback.* SIGIR, pp. 154-161 (2005).

[2] Thorsten Joachims, Laura Granka, Bin Pan, Helene Hembrooke, Filip Radlinski, Geri Gay. *Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search.* ACM Transactions on Information Systems (TOIS), volume 25, issue 2, April 2007.

[3] Thorsten Joachims. *Optimizing search engines using clickthrough data.* KDD, 2002.

[4] Xuehua Shen, Bin Tan, and ChengXiang Zhai. *Context-sensitive information retrieval using implicit feedback.* SIGIR, pp. 43-50 (2005).