CS 630 Notes: Lecture 14
Lecturer: Lillian Lee

Notes by Matt Connolly and Danish Mujeeb

March 14^{th}, 2006

1 Clickthrough data as an implicit feedback mechanism

Abbreviations: CTD (clickthrough data), IF (implicit feedback)

Our "re-motivation": A paper by **Shen, Tan, Zhai '05**, reported experiments that integrated query history and clickthrough history, using document summaries as their data instead of the actual documents. They observed the following results:

- The original queries improved at each iteration, and modified queries improved at each iteration, leading to an overall improvement as the feedback process was iterated.
- Clickthrough data was much better than query history as an implicit feedback mechanism.
- Only approximately 1/3 of the documents for clicked-through summaries were relevant! This figure is kind of low but it might imply that the document summaries are misleading. It definitely implies that we shouldn't assume that the underlying documents are relevant, which means we can't assume clicks represent IF for the corresponding documents.
- *However*, they still managed to get good performance. Why?
    1. It is possible that summary relevance is not correlated with document relevance, as suggested above.
    2. It may also be possible that the feedback mechanism is so good that it can tolerate a lot of noise ... or is it?

Experiment: As a further step, the researchers retrospectively removed all summaries from the CTD corresponding to relevant documents; thus, the summaries now visible to the users were those of non-relevant documents. They then re-ran the original experiment.

Result: They still observed some improvement, thereby supporting the hypothesis that summary relevance is not (strongly) correlated with document relevance.

Further questions prompted by these experiments:

1. Can we make better use of clickthrough data than what we've mentioned so far?
2. Does a 'click' correspond to the relevance of a *summary*? Here are a couple of cases where it may *not* and thus contaminate the data:

- People may click on a summary for reasons other than "it is relevant". They might find the non-relevant summary interesting enough to click on it anyway, or users may tend to trust the system more than they should to provide accurate rankings.
- People tend to click on the top m summaries regardless of relevance. In fact, browsers with a tabbed interface and a feature to "open in background" actually encourage users to behave this way.

This is the motivation behind the following study ....

2 Evaluating the quality of clickthrough data

Ref. Joachims, Granka, Pan, Hembrooke, Gay '05

2.1 Experiment

The authors conducted a study to evaluate the quality of clickthrough data as an implicit feedback mechanism. It was a fairly large study, which was set up as follows:

- Google was used as the underlying search engine (although extraneous page material such as sidebar ads was stripped from the pages actually viewed).
- Approximately 50 college students total were involved as the subjects during various phases of the project.
- Each student was asked to perform 10 specific tasks using the search engine.
- Students were allowed to come up with and issue whatever queries they liked.
- Clickthrough data was collected while the subjects were performing their tasks.
- There were 5 "judges" collected from a separate pool. For the results of each subject's queries, they created relevance rankings for the generated page summaries and for the underlying documents (even if the subjects didn't look at them).
- We can probably trust the judges' rankings since $80 - 90\%$ of the times the judges were in agreement on the relative ranking of a pair of items.
- Eye-tracking data was also collected to measure the attention focus of subjects (cf. Puolamäki et al. '05, using eye tracking as IF).

2.2 Results

For the following data, "S1" and "S2" correspond to the first and second summaries presented as the results of a search.

- On average, users do in fact look from top to bottom on a page (although they didn't have distracting ads alongside their results).
- Most of the time, S1 and S2 are both actually viewed (looked at, but not necessarily clicked).
  S1 $\sim 68\%$ of pages
  S2 $\sim 61\%$ of pages

- 50% of the time, the summary below the clicked summary is looked at before making the choice.
- S1 gets the 1^{st} click in 40% of pages.
  S2 gets the 1^{st} click in 15% of pages.

Given that users gave almost equal consideration to S1 and S2, why is S1 clicked more? Could this be a presentation bias, or does S1 generally turn out to be more relevant than S2?

2.3 Investigating presentation bias effects

Question: How can we isolate presentation bias? One simple way is to look at cases in which *exactly* one of the two top results (S1 or S2) was clicked (given that we can assume both were looked at). In further investigations, Joachims et al. found that:

- if S1 was more relevant, it was clicked 19/20 times, or 95% of the time.
- if S2 was more relevant, it was clicked 2/7 times, or $\sim 28.57\%$ of the time .

So there's definitely a bias! But because the total numbers of clicks - 7 versus 20 - are so different, it's possible that the second set of cases was somehow "odd". Perhaps they were incidents in which S1 and S2 were too close to reliably judge between them, for example.

2.4 Additional research

In a further experiment, the researchers showed some subjects modified Google pages in which the S1 and S2 results were swapped (without the subjects' knowledge). However, the first link was still preferred (although the ratios were not as distinct).

3 What to do with CTD?

The moral of the story is that we have something of a problem if we want to use CTD with the assumption that a click implies relevance: it's inherently biased!

However, what if we say that CTD implies something different? **(Ref: Joachims '02)** In particular, instead of trying to form absolute statements about relevance, perhaps we can use the click through data simply to form relative judgements: even if a summary S_j may be ordered lower than S_i on a page of search results, we can say that S_j is *more* relevant than S_i if S_j is clicked and S_i is not. Presentation bias may still exist, but it is not a factor in the quality of this inference (though it does affect the amount of data).

Using this premise in experiments yielded a relevance comparison accuracy of $\sim 80\%$, compared to the judges' agreed judgements of about 90%.

A further enhancement of this method that works slightly better in practice is to make the above judgement only if S_j is also the temporally last click (i.e., the last link clicked by the user).

Next time: Can we use this information to improve retrieval for a specific query? and can we even transfer this information about user preference across queries?

4 Questions

1. The experiments by Joachims et al. imply that the visual presentation of results to a query on a search page biases them in a user's mind. In this exercise we consider the implications of this idea.

   (a) Is this necessarily a negative result?

   If the summaries or documents presented are ranked, then there is an underlying assumption that the first result should be the most relevant, and thus the first one to be chosen. However, as we've seen, a user's tendency to prefer the first link is not always correlated with its relevance.

   (b) Are there ways of presenting search results that eliminate user bias while still preserving ranking information?

   Any number of visual or geometric schemes could be posited. The limiting factor in the presentation style is the desire to offer a summary of the underlying document so that the user can judge its true relevance before clicking a link; if the summary is physically connected to the link itself, then the possible visual arrangements of results on a computer screen are necessarily restricted. However, we can eliminate this problem by dissociating the document summaries from the links. One way of doing this would be to present one summary at a time in a distinct screen region activated as a rollover action when the user moved his mouse across the different result links.

   Freed from this geometric limitation, we can imagine a number of presentation styles:

   - A horizontal list of links with the most relevant (S1) at the center, but the remainder stretching outwards in both directions:

     10 — 8 — 6 — 4 — 2 — 1 — 3 — 5 — 7 — 9 — 11

   - A flattened pyramid (Why flattened? Because an isolated "point" link would introduce the same bias toward the "topmost" link that we are trying to eliminate!)

     1 2 3 4

     5 6 7 8 9 10

     11 12 13 14 15 16

   - A three-dimensional cube, with the top-ranked links situated at the corners and lower links arranged on the faces.

   The problem with all of these schemes is that they combat bias by reducing usability. They are "cute" but unintuitive, forcing the user to think about where her most relevant results can be found - something that is likely to induce extreme frustration with the system! Can we find a scheme that is both bias-free (or at least bias-limited) and user-friendly?

   One trendy concept on the Web these days is the idea of a "word cloud" (see one working example at http://www.snapshirts.com). A word cloud is a picture or block of common words from a document. Individual words in the cloud are printed in different font sizes

proportional to their document frequencies. The most commonly used words are the largest and the first to catch your eye; rare words are smooshed into such tiny type that they can hardly be read at all.

Suppose we apply this concept to a set of search results? Instead of document frequency, type size would represent ranked relevance. Of course, we don't have an infinite number of font sizes available to us, nor would that be desirable. Instead, we can bin groups of results into different sizes, with, say, four to a bin. Thus the top four "most relevant" results would be displayed by document title in the largest type size, then the next four in a slightly smaller font, and so on. Again, rollovers would provide document summaries in a different window region. Items in a word cloud are typically presented in alphabetical order, but there is no reason that they could not be randomized for further "unbiasing".

The advantage of this scheme is that the user is attracted to the (hopefully) most relevant items regardless of their positioning, but the relative rankings are still mostly preserved (although the binning eliminates some granularity).

One further idea: perhaps it would be possible to work with an existing word cloud generator on a term-by-term basis. If, for example, we generate a cloud for a page of Google query results, then we should get a picture of term frequency within those documents. For example, the query "white house" returns for the three top hits:

- **Welcome to the White House**
  Official site. Features a virtual historical tour, history of American presidents and their families, and selected exhibits of art in the White House.

- **Welcome to the White House**
  Parody of official White House web site. Includes spoof news and gossip.

- **Welcome to the Office of National Drug Control Policy - ONDCP**
  ONDCP features White House Drug Policy initiatives, programs, and publications. Find testimony and press releases. Outlines national drug control strategy ...

If we generate a word cloud for each of these, then we are left with "links" that arguably provide a more representative and more engaging look at the relevance of a result (see Figures 1 - 3).

2. It was noted by Shen et al. that the relevance of document summaries is only weakly correlated with the relevance of the documents themselves. One could argue that this is due to the way that most search engines summarize documents. Google and Yahoo!, for example, both often (but not always) present summaries by excerpting a line or two of text containing the first instance or instances of the query terms. However, it is quite possible for this strategy to misrepresent the overall content of the page. Here we look at alternate strategies for creating summaries of documents.

   An approach that seems reasonable is to try to summarize a search result as an "ordinary" document without giving the query terms special treatment. Various automated services ex-

ist for doing this: Microsoft Word contains an AutoSummarize tool, and Apple's Mac OS X provides a Summarize "service" built into the OS. How do these tools compare to Google's summarization techniques?

To answer the question, we entered three different queries into Google. For each query, we recorded the top five result summaries. For each of these top five results, we then followed the link to the document itself and created a summary of the entire page's text using Apple's Summarize tools:

Query 1: "information retrieval"

Result 1: Information Retrieval

Google: An online book by CJ van Rijsbergen, University of Glasgow.

Apple: Finally, I am grateful to the Office of Scientific and Technical Information for funding most the early experimental work on which the book is based; to the King's College Research Centre for providing me with an environment in which I could think, and to the Department of Information Science at Monash University for providing me with the facilities for writing.

Result 2: Information Retrieval

Google: Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in information retrieval.

Apple: C.J. van Rijsbergen B.Sc., Dip. NAAC, Ph.D., F.B.C.S., F.I.E.E., C.Eng., F.R.S.E.

Result 3: Modern Information Retrieval

Google: A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web retrieval.

Apple: Full text of Chapters 1 (Introduction) and 10 (User Interfaces and Visualization) are available on-line, as well as a table of contents, exercises and resources for other chapters.

Result 4: UMass Amherst: Center for Intelligent Information Retrieval

Google: University of Massachusetts research lab focused on efficient access to large, heterogeneous, distributed, text and multimedia databases.

Apple: CIIR accomplishments include significant research advances in the areas of distributed information retrieval, information filtering, topic detection, multimedia indexing and retrieval, document image processing, terabyte collections, data mining, summarization, resource discovery, interfaces and visualization, and cross-lingual information retrieval.

Result 5: SIGIR

Google: Addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, retrieval, and distribution ...

Apple: ACM SIGIR addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, retrieval, and distribution of information..

Query 2: "new york income tax"

Result 1: New York State Department of Taxation and Finance

Google: The New York State Department of Taxation and Finance web-site provides ... New Electronic Service. You may now make estimated income tax payments by ...

Apple: Income Tax Extension of Time to File You may file and pay your six-month income tax extension online or by paper.

Result 2: New York State Income Tax Inforamtion (sic)

Google: New York State Income Tax Information ... Taxable Income Is:. The Amount of New York State Tax Withholding Should Be: ...

Apple: Determine the exemption allowance by applying the following guideline and subtract this amount from the result of step 5 to compute the taxable income.

Result 3: Individual Taxpayer Home Page

Google: skip banner navigation, new york state banner - this will open a new window ... Important Information for Income Tax Taxpayers and Preparers ...

Apple: Please see our Tax Relief for Victims of Terrorist Attacks page which explains the tax relief provided under the New York State Tax Law for victims of the September 11, 2001 terrorist attacks.

Result 4: Military Main Page

Google: Publication 361 - New York State Income Tax Information for Military Personnel and Veterans, Provides information about NYS resident and nonresident status ..

Apple: Effective for tax years beginning on or after January 1, 2004, in determining New York adjusted gross income, an individual who is a member of the New York State organized militia must subtract from federal adjusted gross income compensation received for performing active

service within New York State due to state active duty orders issued pursuant to section six of the New York Military Law.

Result 5: New York Life Insurance Company

Google: ... Long Term Care Insurance, and Lifetime Income from New York Life Insurance. ... What was New and Newsworthy, Tax Center A Resource for Tax Planning ...

Apple: New York Life makes that option available again today...learn more Opportunity for All: A Career as a New York Life Agent At New York Life, agents of all different backgrounds have found success in helping protect families and individuals from financial hardships...read more Who Needs Lifetime Income?

Query 3: "tom baker"

Result 1: Welcome to the Official Tom Baker Website

Google: Offers interviews, photographs, merchandise, a list of stage, film and television credits, a biography, and special pages devoted to his television work ...

Apple: FEATURED SHOP PRODUCT City of Death Special DVD Package See the DVDs Section in the Shop for more details.

Result 2: Tom Baker (I)

Google: Tom Baker (I) - Filmography, Awards, Biography, Agent, Discussions, Photos, News Articles, Fan Sites.

Apple: More photos Date of birth (location) 20 January 1934 Liverpool, England, UK Mini biography The British character actor Tom Baker, best known as the fourth incarnation...

Result 3: Tom Baker - Wikipedia, the free encyclopedia

Google: Article providing a biography and career overview.

Apple: Baker has played character parts on television (including Captain Redbeard Rum in the second series Blackadder episode "Potato" and Puddleglum in the BBC's production of The Chronicles of Narnia: The Silver Chair) and radio (including John Mortimer Presents the Trials of Marshall Hall in which Baker plays Britain's most celebrated criminal barrister, Sir Edward Marshall-Hall).

Result 4: Tom Baker Says...

Google: I am Tom Baker and I will be saying things to you by the magic of SMS. ... It all started when I (Mark Murphy) sent a Tom Baker message to a few friends. ...

Apple: One of the first things we came up with was getting Tom to say some classic (and not so classic, but rude) movie quotes.

Result 5: The History of Tom Baker's Scarves

Google: THE HISTORY OF TOM BAKER'S SCARVES. as presented by Bill "the Doctor" Rudloff. Tom Baker's original multicolored scarf was 13 1/2 ft. long in Season 12. ...

Apple: It has the pattern of the previous scarves (execpt it's missing the area of blue-gray that can be seen in Figure F on the right side beneath the purple, red, and yellow stripes) and had more length added to the red/yellow end.

All in all, the results are inconclusive. Although in many cases the summary provided by Google seems more representative, or more informative, than Apple's Summarize service, that is not true in all cases. In particular, Query 1 Result 5 and Query 2 Result 3 are better represented by Apple's summary than Google's. The discrepancies, though, point to the need for more sophisticated summary techniques. In several cases Apple's summarization was inhibited by the embedding of text on a web page into non-text elements such as images or Flash items; the text could not be readily extracted from such elements, so the summaries naturally suffered.

A further summary method that might prove fruitful is the simple compilation of the initial sentence from each paragraph on a page. Of course, for certain pages this technique would lead to very lengthy summaries! However, this points to a problem inherent in all summarization methods: the cost of a tradeoff between brevity and information content.

3. (Open-ended) The study presented by Joachims et al. used eye-tracking equipment to obtain additional information about what test subjects were considering on a page. Of course, for most practical purposes such equipment is impractical because it is expensive, cumbersome, and annoying to the wearer.

   Would it be possible to obtain similar data by tracking the motions that a subject makes with his mouse (i.e., not "clicks", but movements and gestures)? There are some studies that have investigated things such as scrolling time. See Kelly/Belkin survey for a starting point.

Figure 1: Cloud for Link 1

Figure 2: Cloud for Link 2

Figure 3: Cloud for Link 3