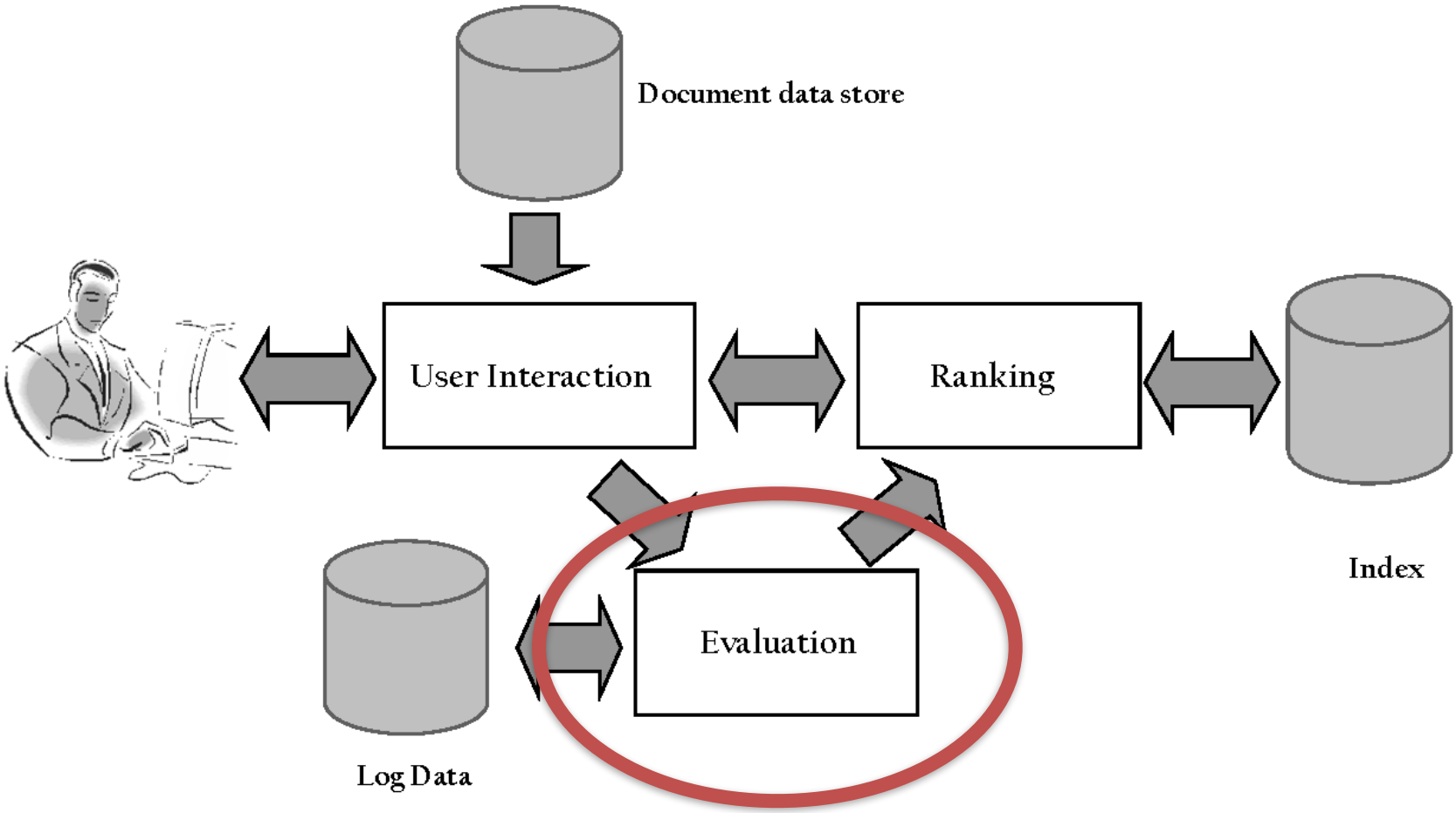
# CS6200
# Information Retrieval

Jesse Anderton
College of Computer and Information Science
Northeastern University

*Based on work by Alistair Moffat and others; see summary*

# Query Process



# Retrieval Effectiveness

- One of the most common evaluation tasks in IR is measuring *retrieval effectiveness* – whether a given ranking helps users find the information they're looking for.

- The ideal but slow and expensive way is to monitor users directly. What do they really look at, click on, and read? Which documents did they find useful?

- We want to emulate that process in a fast and cheap way for a faster development cycle.

- Many mathematical measures of retrieval effectiveness have been proposed – but are they any good?

# Retrieval Effectiveness

- Last time, we discussed several common measures of retrieval effectiveness, including:
  - Precision of top *k* results:
$$P@k(\vec{r}, k) = 1/k \cdot \sum_{i}^{k} r_i$$
  - Average Precision:
$$AP(\vec{r}, R) = 1/|R| \cdot \sum_{i:r_i \neq 0} P@k(\vec{r}, i)$$
  - Discounted Cumulative Gain:
$$dcg@k(\vec{r}, k) = \sum_{i=1}^{k} r_i / \log_2(i+1)$$
  - Reciprocal Rank:
$$rr(\vec{r}) = 1/i : i = \underset{j}{\operatorname{argmin}}\{j : r_j \neq 0\}$$

# Retrieval Effectiveness

- Today we will learn a common framework for thinking about these measures, and learn a little bit about the process we've gone through to improve on our measures.

- This will allow us to investigate and compare the *user models* the measures assume, and think about how realistic those models may or may not be.

- We will also talk about some suggested properties an ideal user model might have, and compare those properties to actual observations of user behavior.

# A Common Framework For Effectiveness Measures

**Common Framework** | User Models | Observed User Behavior

# Another Look at Precision

$$P@k(\vec{r}, k) = 1/k \cdot \sum_{i}^{k} r_i$$

- This can be interpreted as the probability of a user who selects one of the first *k* documents getting a relevant one: $Pr(relevant|retrieved)$

- Let's support continuous relevance values and rearrange the formula:

$$r_i \in [0, 1]; P@k(\vec{r}, k) = \sum_{i}^{k} 1/k \cdot r_i$$

- We can think of this as expected relevance gained from choosing one of the top *k* documents at random.

# Another Look at Precision

$$P@k(\vec{r}, k) = \sum_{i}^{k} 1/k \cdot r_i$$

- We can consider 1/k to be a *weight vector* for the precision metric:

$$W_{P@k}(i) = \begin{cases} 1/k & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

- This lets us reformulate the metric using the weight and relevance labels:

$$P@k(\vec{r}, k) = \sum_{i}^{\infty} W_{p@k}(i) \cdot r_i$$

# Measures as Weight Functions

- It turns out that we can reformulate all these metrics similarly. Here is a general formula for an effectiveness measure M:

$$M = \sum_{i=1}^{\infty} W_M(i) \cdot r_i; \text{ where } \sum_{i=1}^{\infty} W_M(i) = 1$$

- Most measures can be reformulated in this way.

- These measures can be seen as imposing different probability distributions on an *expected observed relevance* function:

$$M(\vec{r}) = \mathbb{E}_{W_M}[r_i]$$

# Scaled DCG

- Recall the formula for dcg@k:

$$dcg@k(\vec{r}, k) = \sum_{i=1}^{k} r_i / \log_2(i+1)$$

- We can't use $W_{dcg@k}(i) = 1/\log_2(i+1)$ – it doesn't sum to 1.

- Instead, we normalize the DCG by summing over all *k* ranks in the list, creating sdcg@k:

$$W_{sdcg@k}(i) = \begin{cases} 1/S(i) \cdot 1/\log_2(i+1) & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

$$S(k) = \sum_{i=1}^{k} 1/\log_2(i+1)$$

# Probability of Continuing

- Sometimes it's convenient to think not in terms of the weight at each rank, but in terms of the probability that a user will look at document *i+1* given that they just saw document *i:*

$$C_M(i) = \frac{W_M(i+1)}{W_M(i)}$$

- Here are the continuation probabilities for the measures we've seen:

$$C_{P@k}(i) = \begin{cases} 1 & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases} \qquad C_{sdcg@k}(i) = \begin{cases} \frac{\log_2(i+1)}{\log_2(i+2)} & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

- This already reveals something of the user models

# Rank-Biased Precision

- Suppose we want something softer than "@k" to decide when a user stops. What if we just pick a constant probability *p*?

$$C_{rbp}(i) = p$$

- This implies the following weights:

$$W_{rbp}(i) = (1-p)p^{i-1}$$

- This has an expected number of documents examined of $1/W_{rbp}(1) = 1/(1-p)$

# Rank-Biased Precision

$$rbp(\vec{r}) = \sum_{i=1}^{\infty} W_{rbp}(i) \cdot r_i$$

- Rank-Biased precision is suggested as an improvement to P@k because it is still top-heavy, but admits some probability of users viewing any document in the ranking.
- However, it has its own flaw: it supposes that users will proceed with the same probability at rank 100 as at rank 2. Do we really believe this?

# Inverse Squares

- Using a constant continuation probability doesn't allow for different behavior in different types of queries. *Inverse Squares* instead uses a parameter *T*, the number of relevant documents a user wants to find.

  ➡ For a navigational query, $T \approx 1$

  ➡ For an informational query, $T \gg 1$

- This metric has associated probabilities:

$$C_{insq}(i) = \frac{(i+2T-1)^2}{(i+2T)^2}; W_{insq}(i) = \frac{1}{S_{2T-1}} \cdot \frac{1}{(i+2T-1)^2} \text{ where } S_m = \frac{\pi^2}{6} - \sum_{j=1}^{m} \frac{1}{j^2}$$

# Inverse Squares

- This measure is more flexible to different query types than RBP.
- It has an expected number of documents viewed of approximately 2T + 0.5, expressing the belief that users will be more patient if they're looking for more documents.
- However, P@k, sdcg@k, rbp, and insq all have a common flaw: they assume that user behavior does not change as the user reads through the list. They all have *static user models.*

# Reciprocal Rank

- Reciprocal Rank is an example of a measure with an *adaptive user model*. It can be expressed in terms of its continuation probability:

$$C_{rr}(i) = \begin{cases} 1 & \text{if } r_i < 1 \\ 0 & \text{if } r_i = 1 \end{cases}$$

- The idea is that the user examines each document in the list, top to bottom, and stops at the first (fully) relevant document.

- This is the first continuation probability function we've seen which takes *document relevance* into account.

# Probability of Being Last

- Many of these measures vary mainly by the way they think about when a user stops. It can simplify things to express our models using the probability that a given item is the last one the user examines:

$$L_M(i) = \frac{W_M(i) - W_M(i+1)}{W_M(1)}$$

- This is a probability distribution as long as we are careful to pick $W_M(i)$ so that it never increases.

- This does not fully specify a model, however: we can't, in general, find W from L.

# Probability of Being Last

- Here are the probabilities of being last for many of the measures we've seen so far:

$$L_{P@k}(i) = L_{sndcg@k}(i) = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$$

$$L_{rr}(i) = \begin{cases} 1 & \text{if } i = \text{argmin}_j \{r_j : r_j = 1\} \\ 0 & \text{otherwise} \end{cases}$$

# Average Precision

- Average precision can easily be defined using L:

$$L_{ap}(i) = \begin{cases} r_i/R & \text{if } R > 0 \\ 0 & \text{otherwise} \end{cases}$$

- By this interpretation, the user will select a relevant document uniformly at random and read all documents in the ranking until the selected one is reached.

- Defining this in our model framework takes a little more work, and is omitted here.

# Summing Up

- We have seen a lot of measures, and discussed a way to view them all as various ways to calculate the *expected relevance* a user will gather from a ranked list.

- The measures generally assume the user will scan the list from top to bottom, and are mainly concerned with specifying when the user will stop, and how to change the probability for items further down the list.

- Let's state a little more clearly what these assume about users.

# User Models

Common Framework | **User Models** | Observed User Behavior

# User Model for P@k

$$C_{P@k}(i) = \begin{cases} 1 & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

- P@k imposes a uniform distribution over the top k documents, and puts zero probability on further documents.
- The user model here assumes that the user will read all of the top k documents, gain whatever relevance is there, and then stop.
- A relevant document at position (k-1) is equivalent to a relevant document at position 1: the user is equally likely to observe both.
- Observation: we want our distribution to be *top-heavy*, with higher probabilities for smaller ranks.

# User Model for sdcg@k

$$C_{sdcg@k}(i) = \begin{cases} \frac{\log_2(i+1)}{\log_2(i+2)} & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

- sdcg@k puts higher probability on smaller ranks. It supposes a user is more likely to stop as they move further down the list.
- This particular discount function might not be the right one: for k=100, the probability of continuing at rank 100 is about 1/7th that at rank 1.
- Worse, the probability is suddenly 0 at rank 101. Given that a user reads the document at rank k, will they really *always* stop there?
- Observation: We want our probabilities to drop smoothly, and perhaps to fall off more steeply than this.

# User Model for RBP

$$C_{rbp}(i) = p$$

- Rank Biased Precision uses a geometric distribution to put a probability of the user visiting any document in the list.
- The probability of visiting a document does decrease as you move down the list, and more sharply than does sdcg@k.
- However, it assumes that a user is equally likely to continue very early and very late in the list. This doesn't seem to be true: if you just read the 47th document you seem more likely to read "just one more."
- Observation: We may want our probability of continuing to *increase* deeper in the list.

# User Model for Inverse Squares

$$C_{insq}(i) = \frac{(i + 2T - 1)^2}{(i + 2T)^2}$$

- Inverse Squares uses the number of relevant documents the user expects to find, T, to choose a probability of continuing.
- The probability of continuing decreases as you move further down the list. It applies fairly good weights both to the top and the bottom of the list.
- However, it is still static – the probability of continuing does not depend on whether the user found what they were looking for.
- Observation: we want to use an adaptive user model.

# User Model for Reciprocal Rank

$$C_{rr}(i) = \begin{cases} 1 & \text{if } r_i < 1 \\ 0 & \text{if } r_i = 1 \end{cases}$$

- Reciprocal Rank supposes that the user reads every document from the top of the list to the first fully-relevant document.
- It does not assign any likelihood to the event that the user gives up early.
- It does, however, take into account whether the user gathered the information they were looking for, for some simple definition of an information need.

# User Model for Average Precision

$$L_{ap}(i) = \begin{cases} r_i/R & \text{if } R > 0 \\ 0 & \text{otherwise} \end{cases}$$

- AP supposes that the user selects a relevant document uniformly at random, and then reads all documents from the top of the list to the selected document.
- It is adaptive, in that the user stops based on finding a relevant document.
- It is also highly unrealistic, in the sense that it assumes the user knows which documents are relevant before they begin.
- Further, to calculate it we need to know how many total relevant documents are in the collection, whether they were retrieved or not.
- Observation: we (may) want our measure to be calculated based only on the retrieved documents.

# Modeling Expected Relevance

- If a measure is a way to estimate the expected relevance a user observes, its assumptions about user behavior should closely match real user behavior.

- This provides a suggestion for comparing the utility of these measures: the measure which more closely fits actual user behavior is to be preferred.

- What do we think users do?

# Observed User Behavior

# Expected User Behavior

1. Users engage in different search tasks – sometimes users just want one document, and sometimes they want to read many.

   ➡The parameterless RR and AP fail here, but the others can be adapted to handle it.

2. Users may look arbitrarily deeply in the ranked list, though the probability of looking deeper should be relatively small. That is, C(i) should never be zero.

   ➡P@k and sdcg@k fail here.

3. If users have already looked deeply into the list, they are more likely to continue. All else being equal, C(i) should increase.

   ➡P@k and RBP both fail here.

# Expected User Behavior

4. Users may change their behavior based on the documents they have already seen.

   ➡The static models fail here.

5. Users may exit the query without being satisfied.

   ➡The dynamic models – RR and AP – fail here.

# Actual User Behavior

- Moffat et al [2013] performed a user study to determine whether users actually exhibit these expected behaviors.
- Their findings confirmed many of them, but included some surprising aspects.
- The remainder of this lecture will describe what they did, and what they found.

# Actual User Behavior

- Users were given a list of six information needs and a proposed starter query, and asked to use a custom search engine to find documents to adequately satisfy those needs and to mark them as relevant.

- The computer they used employed the Yahoo search API with a non-branded interface. Users could formulate a query, read document URLs and snippets, view documents (in a popup), and then mark them as relevant or non-relevant.

- Users eye-movements were also tracked, to investigate the order in which users actually considered documents.

# Information Needs

- Users were given the following tasks. The idea was that the tasks should become progressively harder.

| Information Need | Starter Query |
|---|---|
| (remember) You recently watched a show on the Discovery Channel, about fish that can live so deep in the ocean that they're in darkness most or all of the time. This made you more curious about the deepest point in the ocean. What is the name of the deepest point in the ocean? | deepest ocean point |
| (remember) You recently attended an outdoor music festival and heard a band called Wolf Parade. You really enjoyed the band and want to purchase their latest album. What is the name of their latest (full-length) album? | wolf parade |
| (understand) Your nephew is considering trying out for an Australian Rules football team. His parents are supportive of the idea, but you think the sport is dangerous and are worried about the potential health risks. Specifically, what are some long-term health risks faced by football players? | australian rules football health risks |
| (understand) You recently became acquainted with one of the farmers at the local farmers' market. One day, over lunch, they were on a rant about how people are ruining the soil. They were clearly upset, so you're interested in finding out more. What are some human activities that degrade soil fertility? | damage soil fertility |
| (analyze) Your sister is turning 25 next month and wants to do something exciting for her birthday. She is considering some type of extreme sport. What are some different types of extreme sports in which amateurs can participate? What are the risks involved with each sport? | extreme sport |
| (analyze) You recently heard someone claim that identity theft in Australia is on the rise. This has made you concerned about protecting your own identity. How easy or difficult is it for a stranger to open a credit card under your name? What essential information about you is needed and what are some effective ways in which you can protect your identity in the future? | identity theft and credit cards |

# Data Collected

- Users answered a brief demographic survey before the work (results: 8 female, 26 male; mean age 26; all fluent in English, but for half it was not their first language; all pursuing degrees in CS, math, or engineering)

- For each user and query, the researchers collected:

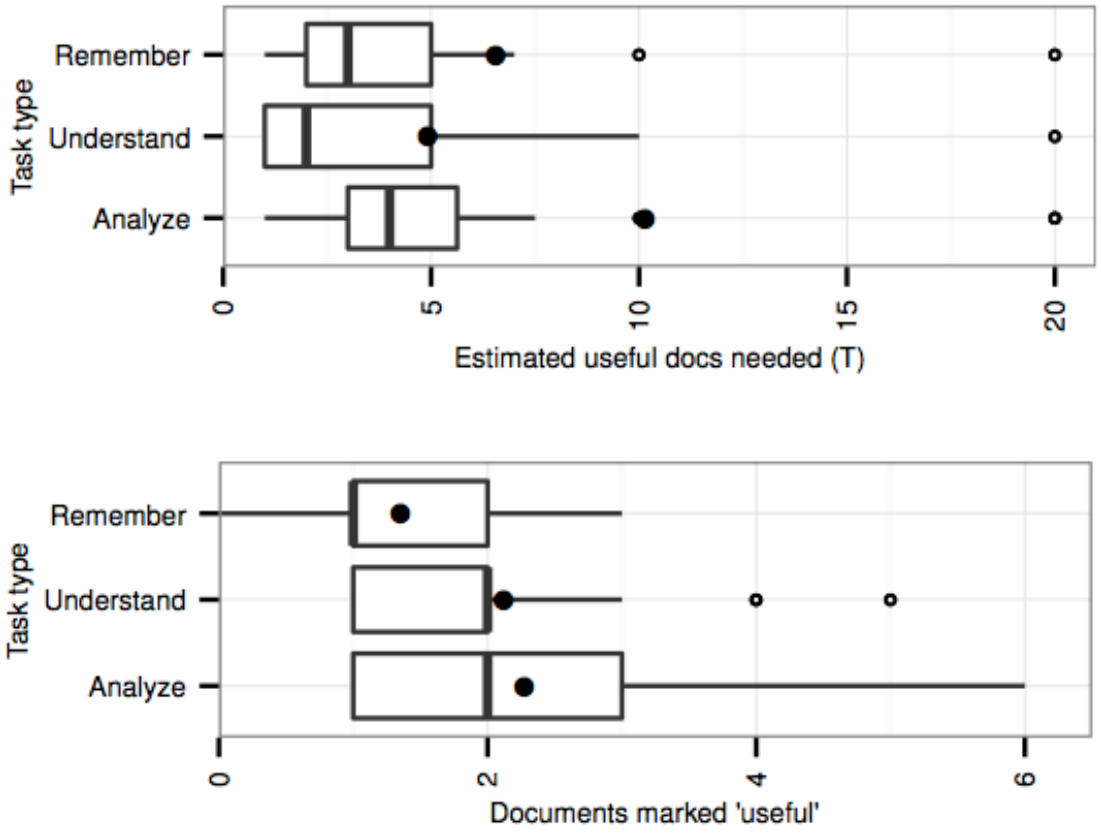
  ➡ The user's reported number T of pages they expect to need to read to answer the query

  ➡ The order in which the user's eyes scanned the results list

  ➡ Whether each visited document was marked relevant or non-relevant by the user

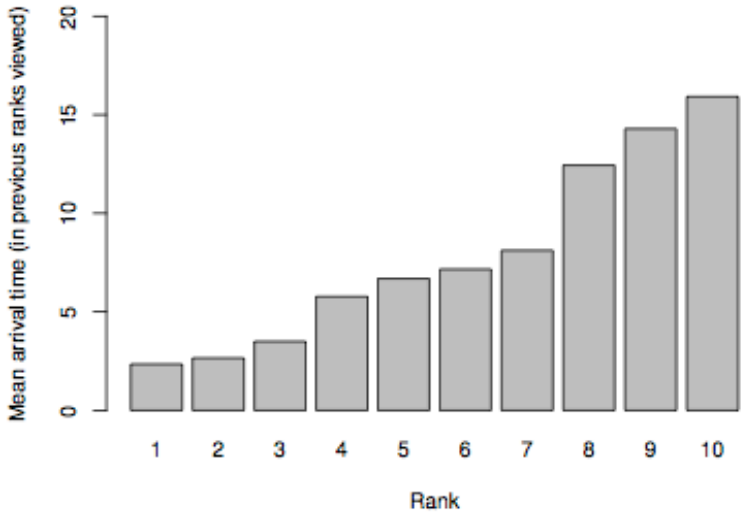
# Estimating T

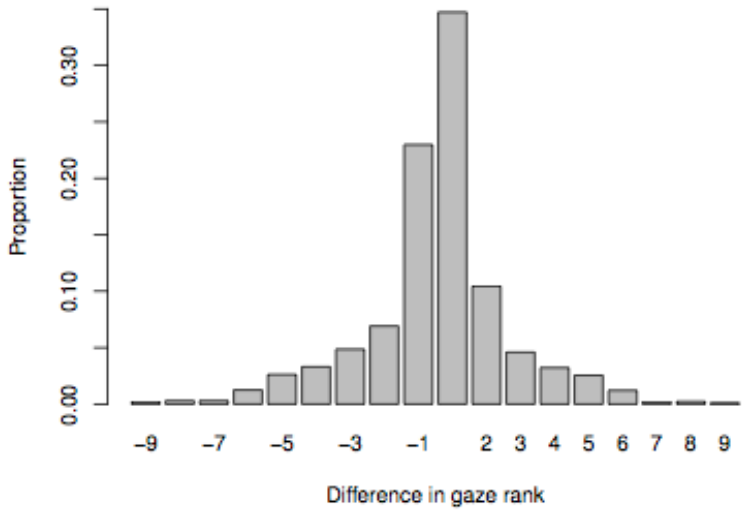
- The number T of documents the users expected to need are compared here with the number actually marked as useful.
- Contrary to expectations, the only significant difference in estimates was between *understand* and *analyze* tasks.
- While users did need more documents for *analyze* tasks, the numbers were much smaller than expected.



# Tracking Gaze

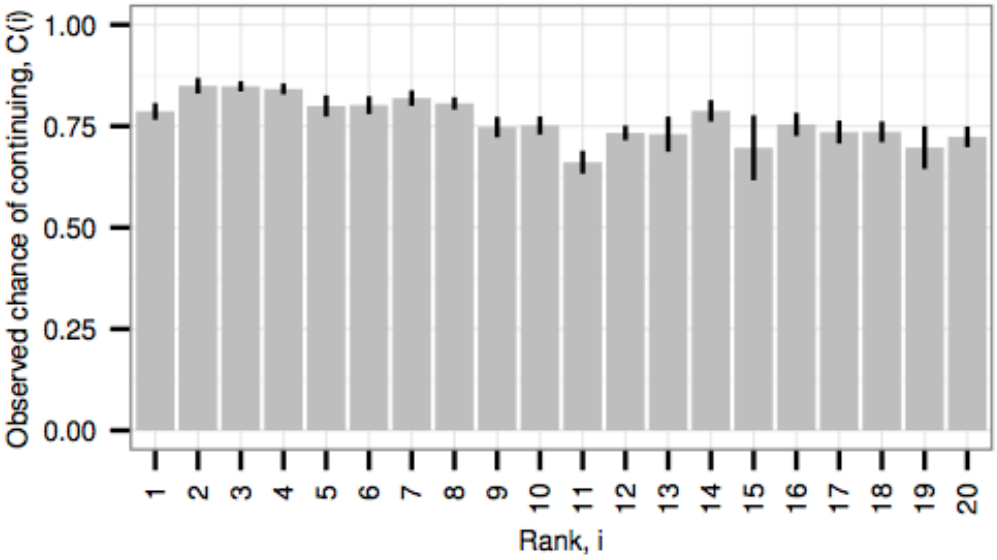
- Most models assume users scan documents from top to bottom.
- The first chart shows that as an overall trend, this is true.
- However, the second chart shows that the story is more complex: users often skip ahead by two or more document, or skip backwards in the list.
- Also, notice the jump from rank 7 (the last visible on the screen) to 8.



# Estimating C(i)

- The data were also used to estimate an empirical probability of continuing C(i).
- The results, averaged across all users and queries, are shown to the right.
- Although this appears linear, or even constant, this is a combination of several factors.



# Estimating C(i)

- The authors fit a model to estimate C(i) based on a variety of factors. The model selected the most parsimonious model (based on AIC) to explain the data.
- The effect size is a multiplicative factor that changes the value of C(i+1) given C(i).
- The user's identity was the greatest source of variance, but the probability also heavily depends on other factors.

| Factor | Effect |
|---|---|
| (intercept) | 11.70 |
| User | 0.11–10.95 |
| Proportion of $T$ collected | 0.34 |
| Proportion of docs viewed that are relevant | 0.50 |
| Gaze sequence | 0.97 |
| Rank $i$ | 1.06 |
| Query count, in task | 1.10 |

# Estimating C(i)

- Comparing "proportion of T collected" to "proportion of docs views that are relevant" – both matter, but the user's prior expectations have a larger effect on choosing to stop.
- Reading a document at a higher rank does correspond to higher probability of reading "just one more."

| Factor | Effect |
|---|---|
| (intercept) | 11.70 |
| User | 0.11–10.95 |
| Proportion of $T$ collected | 0.34 |
| Proportion of docs viewed that are relevant | 0.50 |
| Gaze sequence | 0.97 |
| Rank $i$ | 1.06 |
| Query count, in task | 1.10 |

# Conclusions

- The authors conclude that most of the points in “Expected User Behavior” are supported by their data.
- A measure which effectively models user behavior must take into account rank, relevance, and also user’s expectations and relevance obtained so far.
- It would be ideal to also include per-user parameters – that was the highest variance parameter – but this is probably not realistic in most settings.

# Summary

- Many of these measures were created without expected relevance in mind. The framework described here was created later, and seems to be a useful way to compare them.

- It is an open research question which particular aspects of user behavior need to be modeled to effectively evaluate effectiveness, and how to best model them.

- Citations:

  - Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and metrics: IR evaluation as a user process. In Proceedings of the Seventeenth Australasian Document Computing Symposium (ADCS '12).

  - Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: what observation tells us about effectiveness metrics. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (CIKM '13).