# User Models, Metrics and Measures of Search: A Tutorial on the CWL Evaluation Framework ACM CHIIR UMMMS 2021

by
Leif Azzopardi, Alistair Moffat, Paul Thomas and Guido Zuccon

University of Strathclyde
THE UNIVERSITY OF MELBOURNE
Microsoft
THE UNIVERSITY OF QUEENSLAND AUSTRALIA

# Section Two: What is C/W/L?

Presenter: Alistair Moffat

# Measuring the usefulness of a ranking?

Let's suppose that a numeric **gain** can be attached to each document in the ranking, and that $0 \leq r(i) \leq 1$ is the gain attached to the document at rank i.

A gain of **zero** means "useless", and a gain of **one** means "fully useful".

**How do we measure the usefulness of the ranking as a whole**??

# When one user looks at a ranking

First document

Second document

Third document

Fourth document

*[and then stops looking]*

# When a second user looks at the same ranking

First document

Second document

*[and then stops looking]*

# When a third user looks at the same ranking

First document

Second document

Third document

Fourth document

Fifth document

*[and then stops looking]*

# When a fourth user looks at the same ranking

First document

Second document

Third document

Fourth document

Fifth document

Sixth document

[*and then stops looking*]

# When a fifth user looks at the same ranking

First document

Second document

Third document

Fourth document

[*and then stops looking*]

# When a sixth user looks at the same ranking

First document

[*and then stops looking*]

# When a seventh user looks at the same ranking

First document

Second document

[*and then stops looking*]

# When an “average” user looks at the same ranking

**First document**

Second document

Third document

Fourth document

Fifth document

Sixth document

Seventh document

# When an "average" user looks at the same ranking

**First document**

Second document

Third document

Fourth document

Fifth document

Sixth document

Seventh document

C(1): probability of continuing from doc 1 to doc 2

C(2): probability of continuing from doc 2 to doc 3

C(3): probability of continuing from doc 3 to doc 4

C(4): probability of continuing from doc 4 to doc 5

C(5): probability of continuing from doc 5 to doc 6

C(6): probability of continuing from doc 6 to doc 7

# Huh? What is C(i)?

Define C(i) to be:

- the **conditional continuation probability** that a randomly selected user will proceed from document i in the ranking to document i+1
- **given** that they have just looked at document i, and
- **assuming** that users always start at the top of the ranking at the first document (rank position 1).

# Huh? What is C(i)?

Clearly, $0 \leq C(i) \leq 1$ for each depth i in the ranking.

And C(k+1) onward are immaterial if C(k)=0 occurs for some k.

**What factors might affect C(i)??**

# Huh? What is C(i)?

*Example*: suppose that users are modelled as **always** looking at the first five documents in the ranking, and **never** going beyond those five.

Then C(1)=C(2)=C(3)=C(4)=1.0, and C(5)=0.0.

If this pattern of behavior makes you think about the metric **precision at depth five**, P@5, your instincts are working well.

And if it doesn't, well, you'll find out why it should have in just a minute!

# Huh? What is C(i)?

*Example*: suppose that users are modelled as **always** continuing from depth i to depth i+1 with some **constant probability** $\phi$, that is, $C(i)=\phi$ for all i.

Now what? Now there will be non-zero "probability of being viewed" that can be calculated for every position in the ranking.

For each different function C(i) a **weight** can be derived and associated with the document at rank i.

# Huh? What are these "weights"?

We can compute the corresponding W(i) function for any C(i) function.

It captures the **fraction of all user attention associated with the document in the i'th place of the ranking:**

$$W(i) = \frac{\prod_{j=1}^{i-1} C(j)}{\left(\sum_{k=1}^{\infty} \prod_{j=1}^{k-1} C(j)\right)}.$$

*Example*: C(1..4)=1.0, C(5..)=0; then W(1..5)=0.2, W(6..)=0.0.

*Example*: $C(i)=\phi$; then $W(i)=(1-\phi)\phi^{(i-1)}$.

# Huh? What are these "weights"?

Can now compute the **expected rate of gain** version of the "metric" defined by the values associated with the function C(i):

$$M_{\mathrm{ERG}}(\mathrm{r}) = \sum_{i=1}^{\infty} W(i) \cdot r(i)$$

*Example*: C(1..4)=1.0, C(5..)=0; then W(1..5)=0.2, W(6..)=0.0.

The corresponding metric is **Precision at Depth Five**, P@5.

# Huh? What are these "weights"?

$$M_{\mathrm{ERG}}(\mathrm{r}) = \sum_{i=1}^{\infty} W(i) \cdot r(i)$$

*Example*: if C(i)=$\phi$; then W(i)=$(1 - \phi)\phi^{(i-1)}$.

The corresponding metric is **Rank-Biased Precision**. When $\phi$=0, the user is completely impatient, matching P@1. When $\phi$=0.5, the user is somewhat impatient, expected search depth is two.

When $\phi$=0.95, the user is relatively patient, and expected search depth = 20.

# ERG versus ETG metrics

Can also compute the **expected total gain**:

$$M_{\mathrm{ETG}}(\mathbf{r}) = \sum_{i=1}^{\infty} \left( r(i) \times \prod_{j=1}^{i-1} C(i) \right)$$

This is the total “usefulness” derived by the average user when viewing the SERP in question.

# ERG versus ETG metrics

Simple algebra then gives the **expected viewing depth** (the average number of documents viewed by users) as 1/W(1).

ERG metrics measure systems based on the rate at which their users acquire "usefulness", and have units of "rels/document".

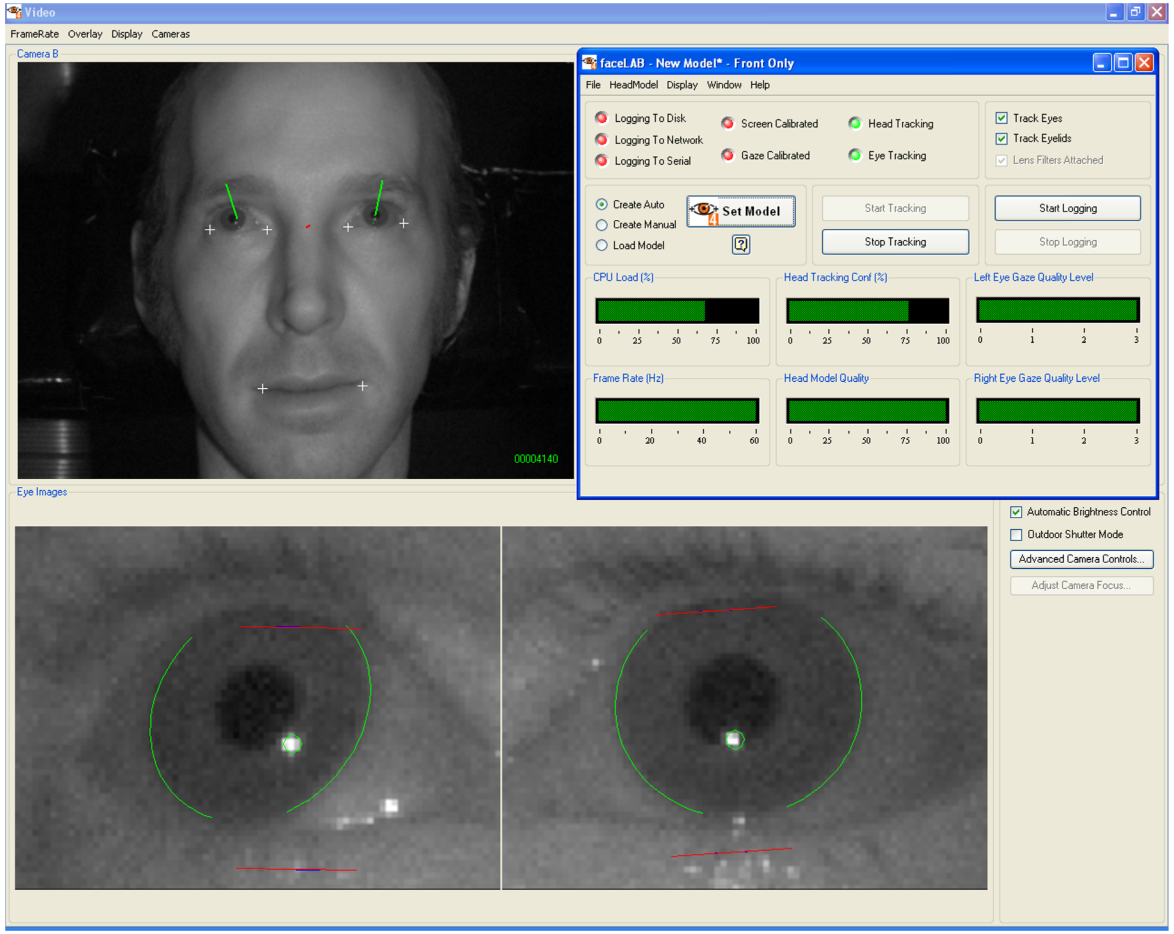
ETG metrics measure systems based on the total "usefulness" acquired by users, and have units of "rels".

# What factors could/should/might affect C(i)?

Some “directional” hypotheses, assuming a user who searching for, and hoping to acquire, a total of T units of gain:

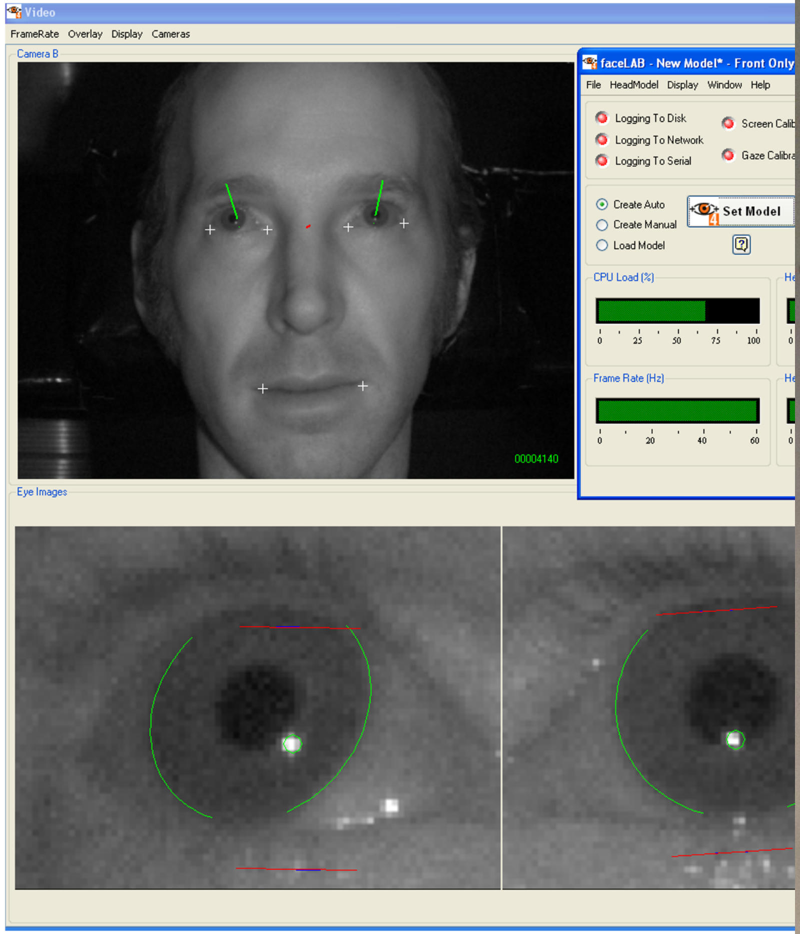
- When T is larger, C(i) is larger, AOTBE
- When i is larger, C(i) is larger, AOTBE
- As the relevance collected gets larger, C(i) gets smaller, AOTBE.

# Is there any evidence??

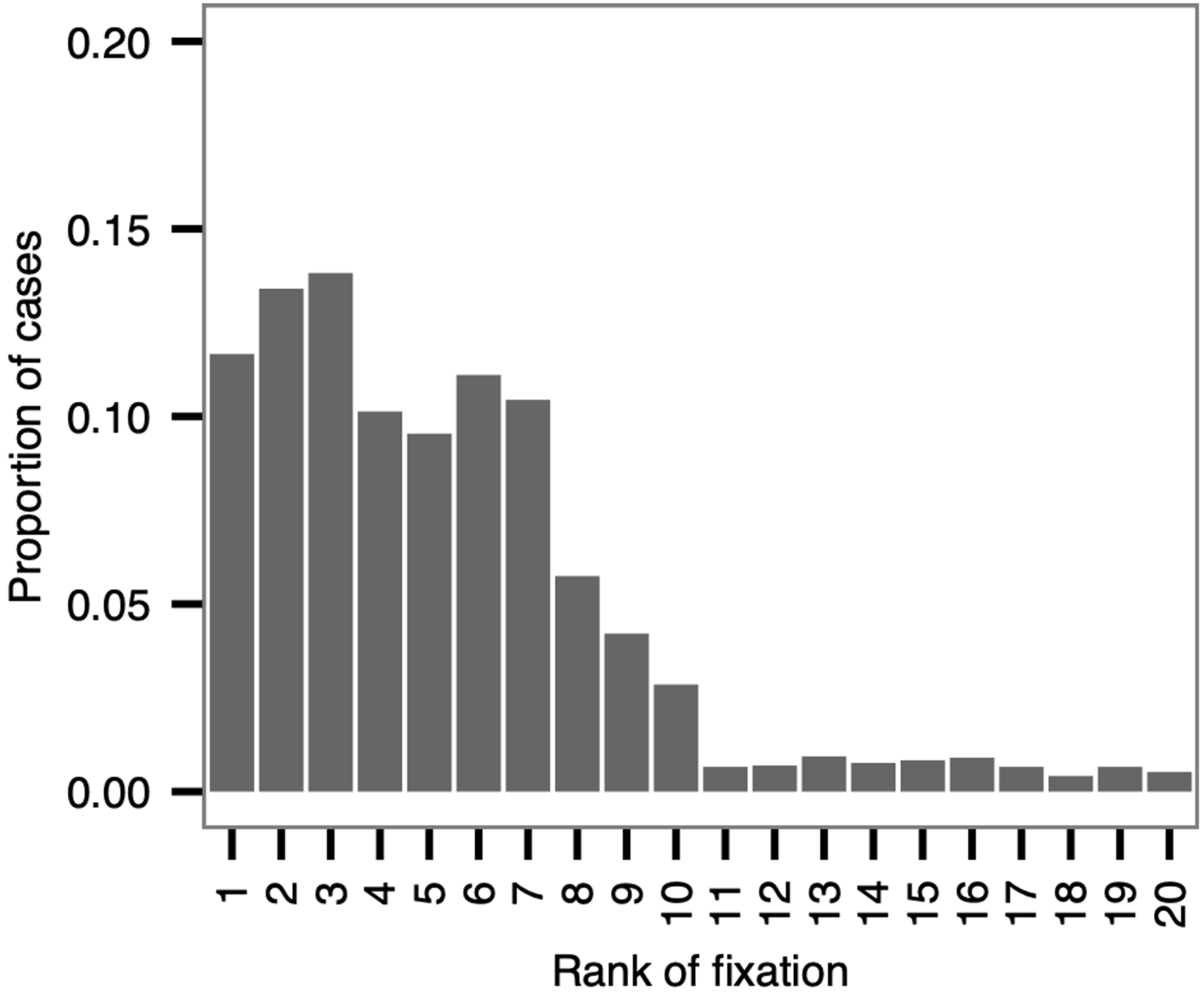



Source: Paul Thomas.

# Is there any evidence??

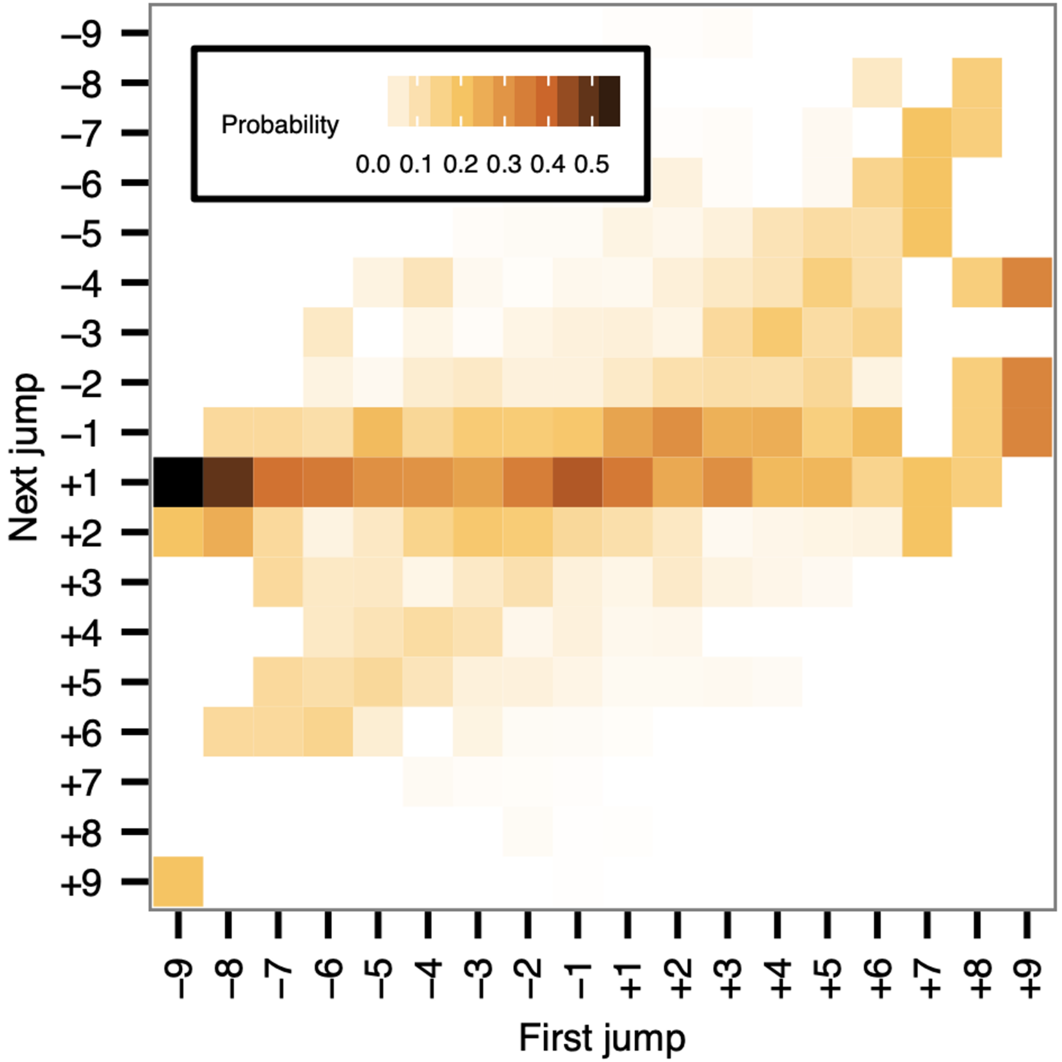



# Is there any evidence??




Graph: Thomas et al., AIRS 2013.

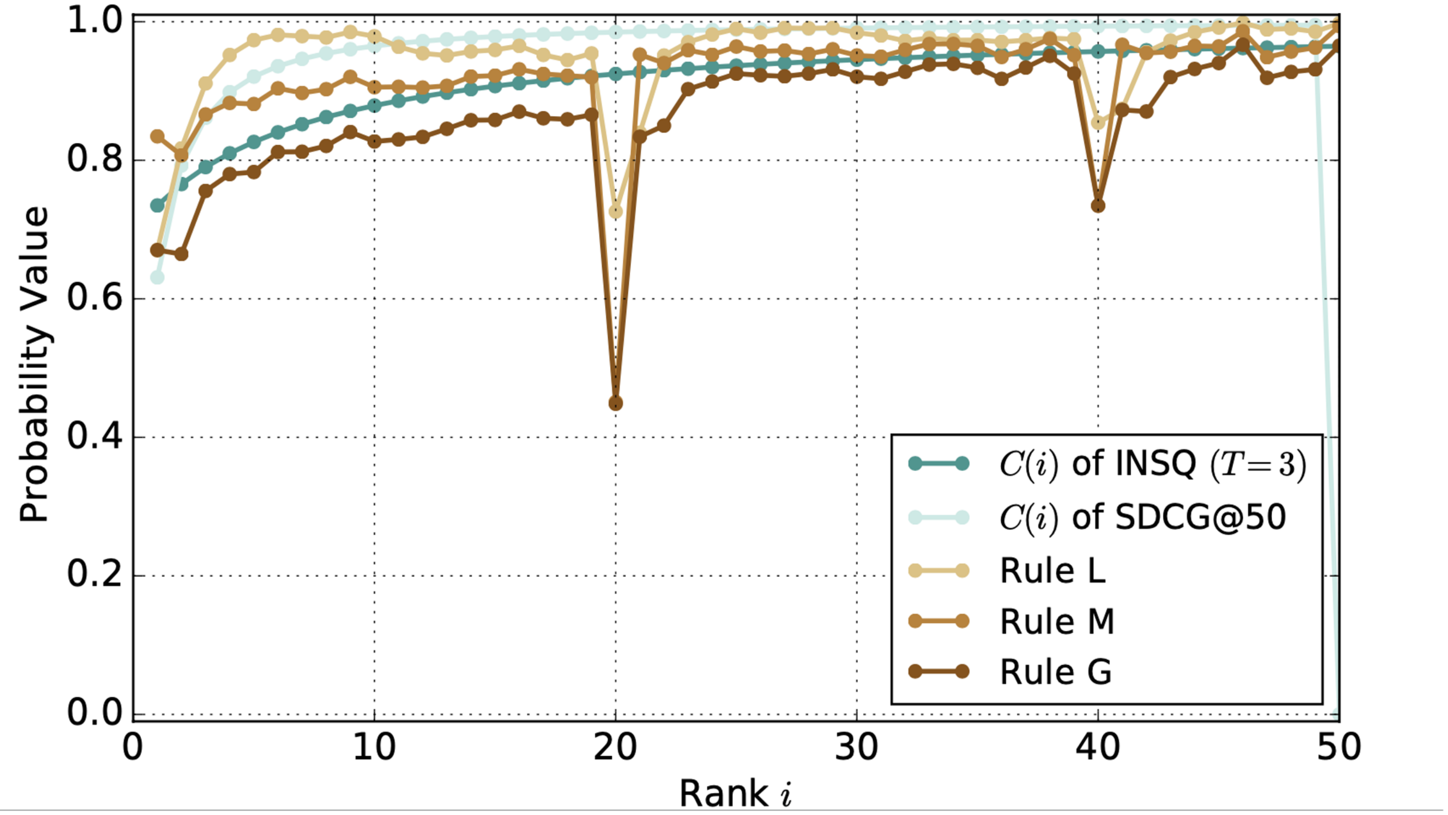
# Is there any evidence??




Graph: Thomas et al., AIRS 2013.

# Is there any evidence?

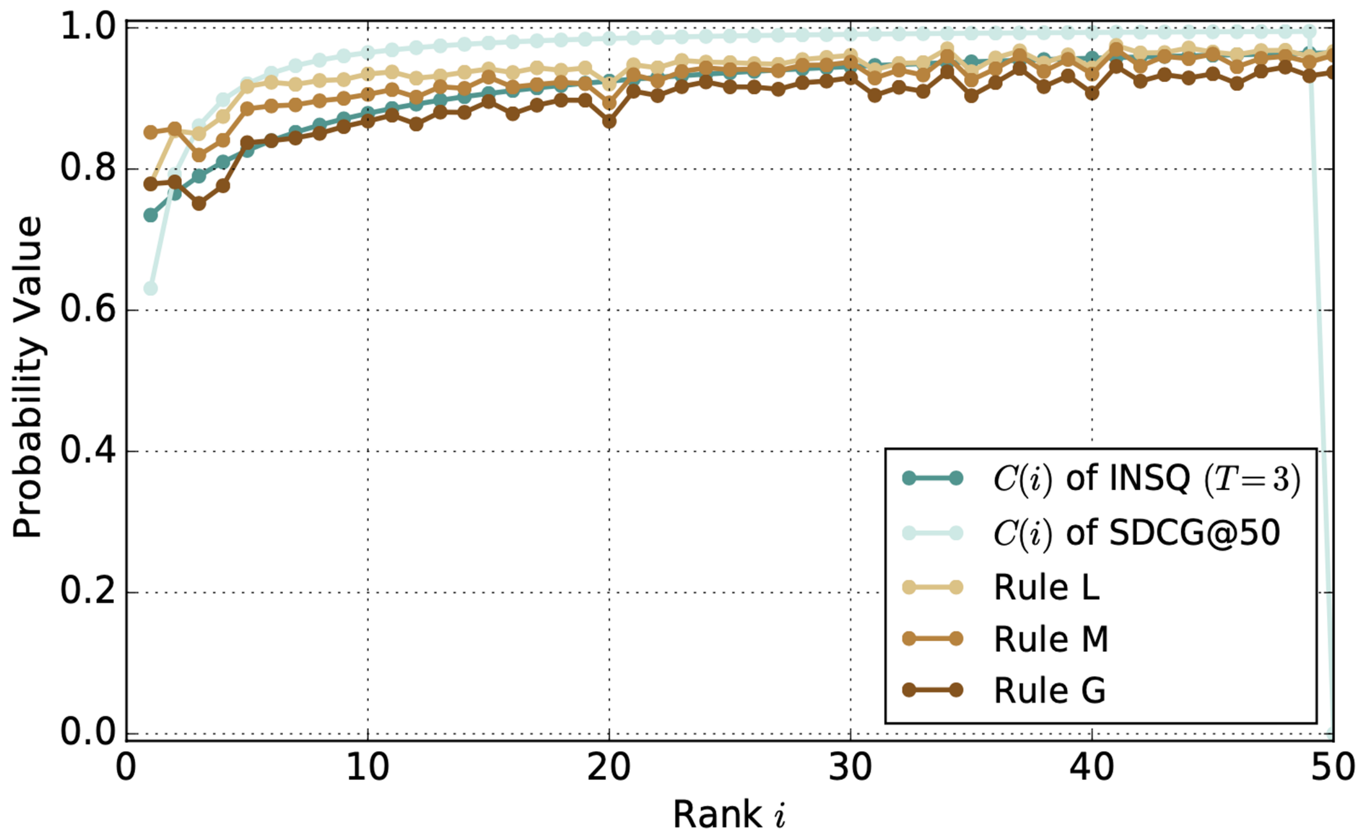
Inferred C(i) for job search users, web browser, pages of 20.




Graph: Wicaksono & Moffat, CIKM 2018, with thanks to Seek.com.

# Is there any evidence?

Inferred C(i) for job search users, phone app with continuous scroll.




Graph: Wicaksono & Moffat, CIKM 2018, with thanks to Seek.com.

# Formulating metrics (1)

Already covered in examples: if $C(i)=1$ for $1 \leq i < k$, and $C(k)=0$, then the CWL metric is **P@k**.

And $W(i)=1/k$ for $1 \leq i \leq k$, with $W(i)=0$ for $i > k$.

The expected viewing depth is k.

That was an easy one, to get started.

# Formulating metrics (2)

Also already introduced: if C(i)=$\phi$ for all i, then

- W(1) = (1 – $\phi$),
- W(i+1) = $\phi$ W(i), and
- expected viewing depth is 1/(1 – $\phi$).

Rank-biased precision assigns non-zero weight to every document in the ranking.

But unless $\phi$ > 0.95, the actual weight assigned at ranks > 50 is negligible.

# Formulating metrics (3)

What about a user who seeks one useful document, and stops if they find it?

Set C(i) = 1 – r(i), and suppose that r(i) is binary, either zero or one. The metric is now **adaptive**, in that user behavior depends upon **what they have seen**. (Wow!)

Then W(i) = 1/d, where d is the rank of the first relevant document.

This is **Reciprocal Rank**!

(Could also have non-binary C(i) = 1 – r(i), but does not equate to ERR.)

# Formulating metrics (4)

What about the metric defined by this function?

$$C(i) = \frac{\sum_{j=i+1}^{\infty}(r(j)/j)}{\sum_{j=i}^{\infty}(r(j)/j)}$$

The user is modeled as deciding what to do now (at rank i) based on relevance values they have not yet seen (from ranks j>i).

This is the definition of **Average Precision**. Yes!

# Formulating metrics (5)

Suppose the user starts their search with the hope of acquiring T units of “usefulness”. And suppose that by rank i, they have acquired R(i) units.

Define T(i) = T – R(i) as the “unmet requirement” at depth i. Then take

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

This is the definition of a metric named **INST**.

# Formulating metrics (5) – Huh?

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

INST has these properties (AOTBE):

- When T is larger, C(i) is larger – **goal sensitive**
- When i is larger, C(i) is larger – **sunk effort**
- As R(i) gets larger and T(i) gets smaller, C(i) gets smaller – **adaptive**.

# Formulating metrics (6)

You don’t have to use any of those metrics!

If **you** have an understanding of **your** users and how they interact with **your** SERPs, you can define **your own** C(i) function, and use it to measure the effectiveness of **your** system as to strive to provide a better search experience.

# Hey, hang on! What about L?

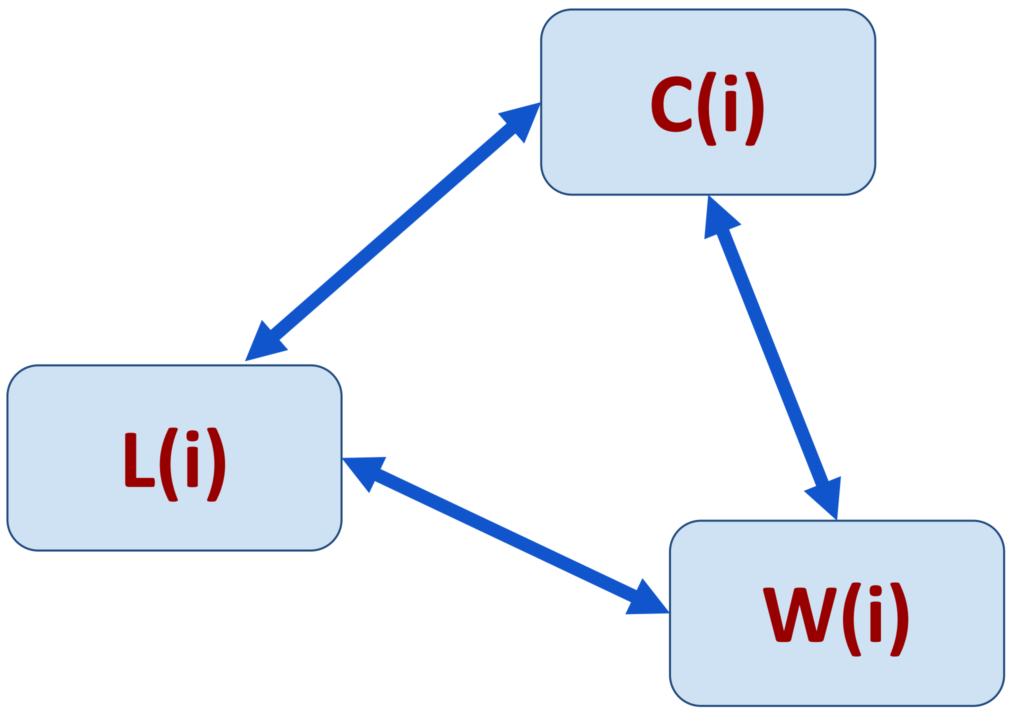
We have talked about C(i). And about W(i). What happened to L(i)?

It is the "last" function, the probability that the document at rank i will be the last one inspected by the user.

It can be computed from either C(i) or from W(i):

$$L(i) = \frac{W(i) - W(i+1)}{W(1)}$$

# See, Double You, 'Ell!



# Something's missing: Residuals

To completely evaluate a metric, need relevance judgments.

What if full judgments are not available?

Compute a **residual** by calculating two scores:

- first, assuming all unjudged document have r(i)=0
- then, assuming all unjudged documents have r(i)=1

True score is between these extremes. The residual is the width of the interval.

If the residual is large, your experiment **may have a problem**.

# Summary of Section II

C/W/L metrics are constructed by hypothesizing behavior over a population of users.

The critical component is C(i), the conditional continuation probability. But they can also be defined via W(i) and/or L(i), each leads to the other two.

From any of C/W/L, ERG and ETG metrics are available, including ones that are goal sensitive and/or adaptive.

Residuals should be monitored, and not ignored.

# Summary of Section II

C/W/L metrics are constructed by hypothesi [illegible] ver a population of users.

The critical component is C(i), t [illegible] continuation probability. But they can also be defined v [illegible] (i), each leads to the other two.

From any of C/W/ [illegible] metrics are available, including ones that are goal sens [illegible] ptive.

Residua [illegible] monitored, and not ignored.

C/W/L Spells "Cool"