Modelling Search Session

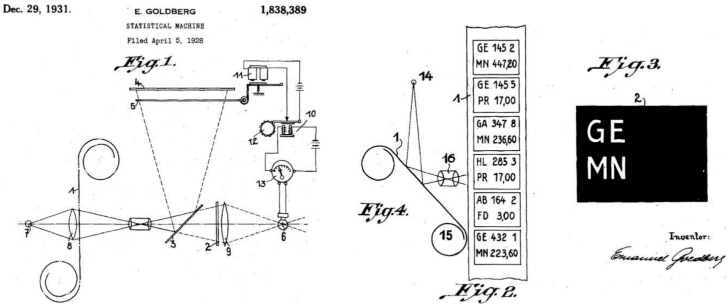
Alfan F. Wicaksono

School of Computing and Information Systems
The University of Melbourne

IR-NLP Talk #2, December 2020

Information retrieval: History

Before the era of electronic computers . . .



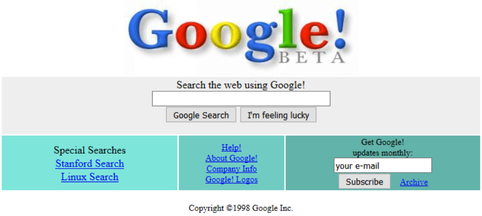
*The History of IR Research, by Sanderson and Croft, 2003

Information retrieval: History

- 1950: The term “information retrieval” was coined for the first time [Mooers, 1950]; IR system using punch cards.
- 1960s - 1980s: The development of ranked retrieval
- 1970s: IDF [Spärck Jones, 1972]; experiments using TF-IDF [Salton and Yang, 1973]
- 1980s - mid 1990s: Retrieval models were extended; BM25 [Robertson et al, 1990s]; Latent Semantic Indexing (LSI), ...

Information retrieval: History

Early Web search engines[1]



[1] https://www.webdesignmuseum.org/uploaded/timeline/google/google-1998.png

Information retrieval: History

Ten blue links ...

ten blue link

All | Images | Videos | Shopping | News | More | Settings | Tools

About 1,320,000,000 results (0.33 seconds)

10 blue links: are they dead or alive in search? – Econsultancy
https://econsultancy.com › 10-blue-links-are-they-dead-or-alive-in-search
Jan 29, 2014 - While many proclaim the death of '**10 blue links**', other experts suggest their own research confirms otherwise. Search is an ever evolving, ...

How Google Search Got Past "10 Blue Links" - Fast Company
https://www.fastcompany.com › how-google-search-got-past-10-blue-links
May 18, 2016 - But at this morning's Google I/O conference keynote, even Google CEO Sundar Pichai made a "**10 blue links**" reference that framed it as an ...

Beyond Google's 10 Blue Links - Ann Smarty
www.seosmarty.com › On-Page SEO › Home
We moved away from **10 blue links** per page long time ago: Google's search results page real estate is much more complicated these days providing us with ...

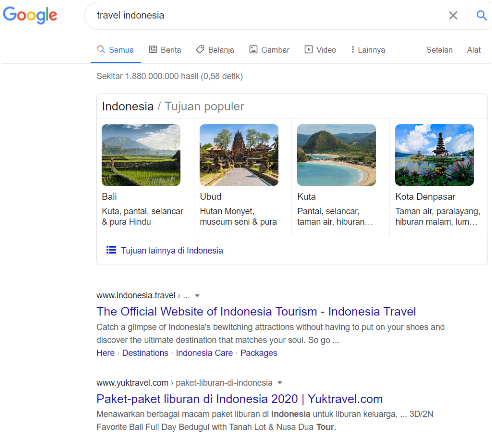
Council Post: Beyond The 10 Blue Links: Optimizing For ...
https://www.forbes.com › sites › forbesagencycouncil › 2017/05/17 › bey...
May 17, 2017 - If you and your brand are only doing SEO for the standard "**10 blue links**" that comprise the web pages, you're losing out. Currently, the ...

[DATA] Ten Blue Links Are Alive and Well in Search - Conductor
https://www.conductor.com › blog › 2014/01 › data-dont-believe-hype-te...
Jan 23, 2014 - **Ten blue links** in search are dead," has been a common refrain in recent years of the search industry. The data, however, tells a different story.

[PDF] **Ten Blue Links on Mars - WWW 2017 Proceedings ...**
papers.www2017.com.au.s3-website-ap-southeast-2.amazonaws.com › pro...
by CLA Clarke - Cited by 1 - Related articles
Apr 7, 2017 - **Ten Blue Links** on Mars. Charles L. A. Clarke, Gordon V. Cormack, Jimmy Lin, and Adam Roegiest. David R. Cheriton School of Computer ...

Information retrieval: Now

More vertical results . . .



Other vertical search engines: Job search (`Seek.com`), product search, legal search, music search, . . .

Information retrieval: Overarching goals

Overall IR challenges [Moffat, SPIRE 2019]:

- Decide what users are asking for (*query analysis and intent*)
  - query classification & clustering
  - query expansion
  - ...

Information retrieval: Overarching goals

Overall IR challenges [Moffat, SPIRE 2019]:

- Decide what users are asking for (*query analysis and intent*)
  - query classification & clustering
  - query expansion
  - ...
- Decide how to find it (*retrieval heuristics, theories of effectiveness, data structures and algorithms*)
  - language models, learning-to-rank
  - indexing and compression
  - how to efficiently retrieve top-k documents? Ex: WAND algorithm
  - ...

Information retrieval: Overarching goals

Overall IR challenges [Moffat, SPIRE 2019]:

- Decide what users are asking for (*query analysis and intent*)
  - query classification & clustering
  - query expansion
  - ...
- Decide how to find it (*retrieval heuristics, theories of effectiveness, data structures and algorithms*)
  - language models, learning-to-rank
  - indexing and compression
  - how to efficiently retrieve top-k documents? Ex: WAND algorithm
  - ...
- Decide whether you have succeeded (*effectiveness measurement and statistical testing*)

Why do we need measurement?

"When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; . . . ."

– W. Thomson (Lord Kelvin)

Which one is better?

Query: “happy phd student”

System A:

happy phd student

All | Images | Videos | News | Shopping | More | Settings | Tools

About 244,000,000 results (0.64 seconds)

www.reddit.com › AskAcademia › comments › aspnc0

What are happy PhD students doing right? : AskAcademia

Does anyone have any insight into what makes the difference between a **happy grad student** and a miserable one? If you've been through it, what was your ...

ahappyphd.org

A Happy PhD | A Happy PhD

A **Happy PhD**. A blog about doctoral productivity, supervision and wellbeing. ... By now, we have established that PhD **students** (and academics in general) ...

www.quora.com › Are-PhD-students-happy-doing-a-PhD

Are PhD students happy doing a PhD? - Quora

Nov 11, 2017 - As a current (3rd year) **PhD student** at UCLA, I am **happy** doing my **PhD**. In fact, I think I am **happier** than the vast majority of **PhD students**, but even my ...

- What makes you **happy** during your **PhD**? - Quora — 19 Mar 2015
- Is it possible that for average person to get a successful **PhD** ... — 11 Sep 2019
- How can a **PhD student** practice to be **happy**? - Quora — 3 Jun 2017
- How many international **PhD students** are not entirely **happy** ... — 16 Apr 2016

More results from www.quora.com

74 answers

www.nextscientist.com › work-life-balance-in-academia

The Happy PhD Zone: How To Maintain A Work-Life Balance ...

As **PhD students** we're trained to put in a significant amount of time planning our experiments. You need to think of everything in advance: in my biology world it's ...

Images for happy phd student

graduate student | nature | happy birthday | survey | graduate school | grad school | stress

System B:

happy phd student

ALL | IMAGES | VIDEOS | MAPS | NEWS

Sign in to see work results

64,700,000 Results | Date | Language | Region

A Happy PhD | A Happy PhD

https://ahappyphd.org

A **Happy PhD** A blog about doctoral productivity, supervision and wellbeing. What do I write about? 2019 **a-happy-phd** ... By now, we have established that **PhD students** (and academics in general) seem to be at a higher risk to develop mental health problems like depression or chronic stress. But, how can we know if we have one of these mental health problems, right here, right now? In this post ...

Are PhD students happy doing a PhD? - Quora

https://www.quora.com/Are-PhD-students-happy-doing-a-PhD

Jafar Saniie cooperated with his favorite **student** Won-Jae Yi who will buy him gifts regularly assigned me and force me to complete a paper and finished grade 43 **students** homework in two days. To finish these overload works, I didn't sleep in 4 consequent days.

A Happy PhD | Ask A Happy Ph.D. - Student edition

https://ahappyphd.org/posts/ask-happy-phd-student

27/04/2019 · POSTS Ask A **Happy Ph.D.** - **Student** edition April 27, 2019 - 4 minutes read - 673 words. This time I would like to turn the mike over to you, **PhD students**, and let you tell me what you want to hear about in this blog (I will do a separate one for supervisors and supervision later on). What would you like to know about doing a **Ph.D.**, or being a (happier, more productive) **PhD student**?

A Day in the life of a PhD student | PostgradAustralia

https://postgradaustralia.com.au/student-life/a-day-in-the-life-of-a-phd-student

A **PhD** in Australia typically runs between 3 and 6 years, though there is increasing pressure from universities to ensure **PhD students** finish at the three-year mark, or only a little later. The days of **PhD students** lasting for six to seven years are long gone, as funding pressure means that universities need to get **PhD students** graduated and producing papers for the **university**. Typically, a **PhD** ...

Happy graduate students stock photo. Image of diploma ...

https://www.dreamstime.com/stock-photo-happy-graduate-students-image...

Photo about Education theme of **graduate students** lying in the park cheerful and **happy**. Image of diploma, completion, female - 19376220

Search engine evaluation

- Online evaluation
  - A/B testing
  - Interleaving
- Offline evaluation
  - Lab-based user study
  - Test collection-based evaluation

Test collection-based evaluation

Query: "happy phd student"

System A:

D_{46} D_{10} D_{18} D_{21} D_{58} D_{71} D_{99} D_{14} D_{81} D_{29}

System B:

D_{10} D_{18} D_{15} D_{22} D_{59} D_{73} D_{99} D_{14} D_{81} D_{29}

Relevance Judgement:

$\langle (D_{10}, 2), (D_{14}, 0), (D_{15}, 0), (D_{18}, 2), (D_{21}, 1), \ldots \rangle$

The judgements generate a gain vector that is aggregated by an effectiveness metric to yield a score.

Metrics: Graded relevance

The judgements might be multi-graded, for example, categories such as "non-relevant", "somewhat relevant", "relevant", and "highly relevant".

In this case, a gain mapping is required to convert grades ($r_i \in \{0, 1, 2, \dots\}$) to gains, for example,

$$g(r_i) = \frac{2^{r_i} - 1}{2^{\max(r_i)} - 1} .$$

Metrics

System A:

D_{46} D_{10} D_{18} D_{21} D_{58} D_{71} D_{99} D_{14} D_{81} D_{29}

System B:

D_{10} D_{18} D_{15} D_{22} D_{59} D_{73} D_{99} D_{14} D_{81} D_{29}

Question: What score would you give to each of them?

Metrics

System A:

D_{46} D_{10} D_{18} D_{21} D_{58} D_{71} D_{99} D_{14} D_{81} D_{29}

System B:

D_{10} D_{18} D_{15} D_{22} D_{59} D_{73} D_{99} D_{14} D_{81} D_{29}

Question: What score would you give to each of them?

Consider "fasilkom ui" vs "general relativity"

	Reciprocal Rank	Precision
A	0.50	0.50
B	1.00	0.25

Metrics: Why do users search?[2]

Economics says people act when they can exchange effort for utility; and that if they have a choice of alternatives and all other factors are equal, they will favor the option with the best conversion rate.

For search, utility is measured as relevance, or gain; possibly fractional, possibly context dependent, and possibly personal.

Effort is measured in seconds or minutes (or perhaps brain-Watts); or approximated by surrogate units called documents inspected.

[2] Credit: Alistair Moffat, NTCIR'16

Metrics: Why do users search?[3]

If effort can be represented by documents inspected,

and if all other things are equal,

then users will prefer the search service with the greatest expected gain per document inspected.

Because that is the best conversion rate between effort and utility.

[3] Credit: Alistair Moffat, NTCIR'16

Metrics: Expected rate of gain (ERG)

To compute an "expected value", a probability distribution is needed.

Metrics: Expected rate of gain (ERG)

To compute an "expected value", a probability distribution is needed.

Let $W(i)$ be the fraction of user attention paid to rank i, with $\sum_i W(i) = 1$. Then, the rate at which the user gains relevance is:

$$M_{\text{ERG}}(\text{r}) = \sum_{i=1}^{\infty} W(i) \cdot g(r_i)\,,$$

where $0 \leq g(r_i) \leq 1$ and $g(.)$ is the gain mapping function.

The units for $M_{\text{ERG}}(\text{r})$ are "expected gain per document inspected."

$W(i)$: How do users search?

$W(i)$ reflects how users interact with Search Engine Results Page (SERP).

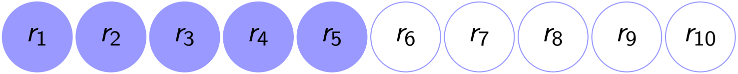
For example, Precision@K assumes that users always inspect items from rank 1 to K:

$$W(i) = 1/K \text{ for } 1 \leq i \leq K, \text{ otherwise } W(i) = 0\,.$$

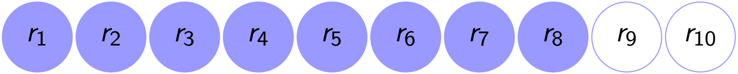
r_1 r_2 r_3 r_4 $\cdots$ r_k r_{k+1} r_{k+2} r_{k+3} $\cdots$

$W(i)$: A model for user search behaviour

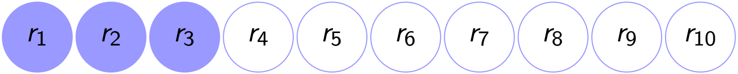
User A:



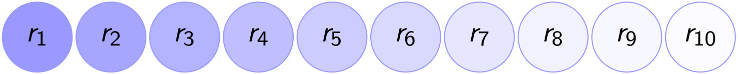
User B:



User C:



A population of users:



What should the $W(i)$'s be?

Can we take it as axiomatic that $W(i) \geq W(i+1)$?

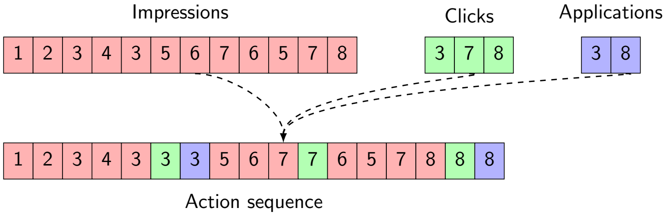
What should the $W(i)$'s be?

Can we take it as axiomatic that $W(i) \geq W(i+1)$?

Let's see empirical evidence from search interaction logs!

Interaction logs from `Seek.com`

A collection of action sequences



	iOS/Android	browser
Users	5,003	5,107
Queries or Action Sequences	74,475	54,341
Page Size	unlimited	paginated, 20

What should the $W(i)$'s be?

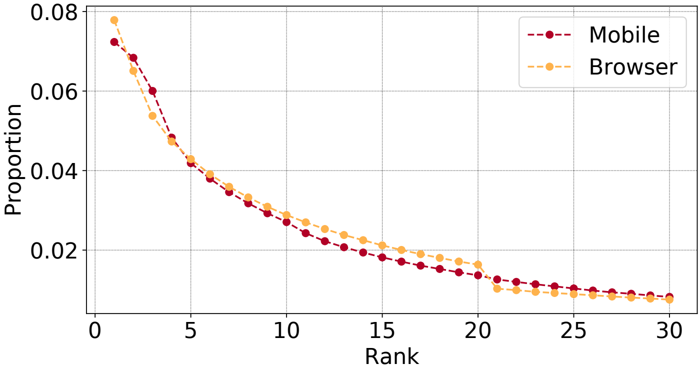
Can we take it as axiomatic that $W(i) \geq W(i+1)$?

What should the $W(i)$'s be?

Can we take it as axiomatic that $W(i) \geq W(i+1)$?

Yes, for the most part

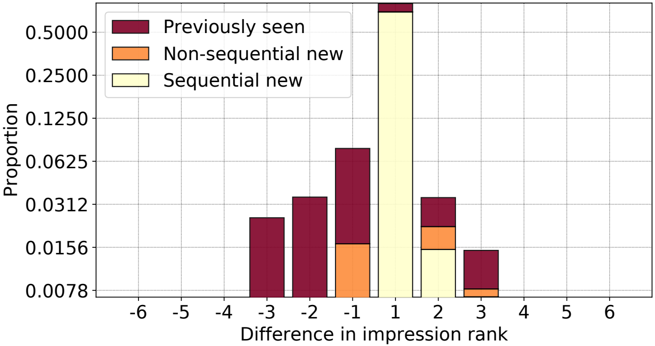
Empirical $\hat{W}(i)$:



Cascade model

Assume that the user starts at rank 1, and sequentially inspects the ranking until they stop.

Distribution of impression jumps:
$\langle 1, 2, 1, 3, 4, 7, 6, 8 \rangle \rightarrow \{-1 : 2, +1 : 2, +2 : 2, +3 : 1\}$



[Wicaksono and Moffat, CIKM 2018]

Continuation probability at rank i, $C(i)$

In a cascade user model, we can alternatively consider $C(i)$:

$$C(i) = \frac{W(i+1)}{W(i)} \, .$$

That is, the **conditional** continuation probability at rank i.

Continuation probability at rank i, $C(i)$

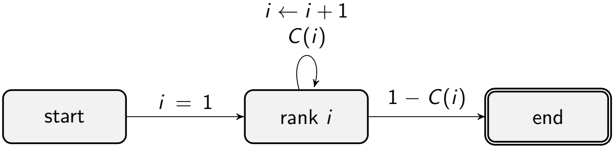
In a cascade model, $C(i)$ is

$$P(\text{ inspection at rank } i+1 \mid \text{ inspection at rank } i\ ).$$

$$r_1 \rightarrow r_2 \rightarrow r_3 \rightarrow r_4 \rightarrow r_5 \rightarrow \cdots$$

Given that a user has inspected the item at rank 4, they have a probability of $C(4)$ to continue to rank 5, and alternatively $1 - C(4)$ to stop.

User Model



$C(i)$, $W(i)$, and ... $L(i)$

C/W/L ("cool") Framework

One family of metrics is described via three inter-connected functions, and the premise that users scan the ranking sequentially from the top until they exit:

- $C(i)$, the conditional continuation probability of the user shifting their attention from the i th document in the ranking to the $i+1$ th
- $W(i)$, the fraction of user attention paid to the document at rank i in the ranking
- $L(i)$, the probability that the i th document in the ranking will be the one last one viewed.

Dualism between metrics and user models

C/W/L Framework describes relationship between metrics and user models.

Dualism between metrics and user models

C/W/L Framework describes relationship between metrics and user models.

$C(i)$ defines how the users interact with search engine results pages.

Dualism between metrics and user models

C/W/L Framework describes relationship between metrics and user models.

$C(i)$ defines how the users interact with search engine results pages.

Once $C(i)$ has been defined, metric scores can be computed via $W(i)$:

$$\sum_{i=1}^{\infty} W(i) \cdot g(r_i) \, .$$

What should the ~~$W(i)$'s~~ $C(i)$'s be?

Existing user models with static $C(i)$

Precision@K: $C(i) = 1$ for $i < K$ and 0 otherwise.

$$r_1 \xrightarrow{C(1)} r_2 \xrightarrow{C(2)} r_3 \xrightarrow{C(3)} r_4 \xrightarrow{C(4)} r_5 \xrightarrow{C(5)} r_6 \xrightarrow{C(6)} \cdots$$

What should the $C(i)$'s be?

Existing user models with static $C(i)$

Scaled DCG, SDCG@K [Järvelin and Kekäläinen, 2002]:

$$C(i) = \frac{\log(i+1)}{\log(i+2)},$$

when $1 \leq i < K$, and 0 when $i \geq K$.

$$r_1 \xrightarrow{C(1)} r_2 \xrightarrow{C(2)} r_3 \xrightarrow{C(3)} r_4 \xrightarrow{C(4)} r_5 \xrightarrow{C(5)} r_6 \xrightarrow{C(6)} \cdots$$

What should the $C(i)$'s be?

Existing user models with static $C(i)$

Rank-biased Precision, RBP [Moffat and Zobel, 2008]:

$$C(i) = \phi\,,$$

where $0 \leq \phi \leq 1$.

high ϕ: patient users
low ϕ: impatient users

$$r_1 \xrightarrow{C(1)} r_2 \xrightarrow{C(2)} r_3 \xrightarrow{C(3)} r_4 \xrightarrow{C(4)} r_5 \xrightarrow{C(5)} r_6 \xrightarrow{C(6)} \cdots$$

Hang on . . .

Let T be the user goal, the number of relevant documents the users initially hoped to see.

Do users really have the same $C(i)$, regardless of T?

Hang on . . .

Let T be the user goal, the number of relevant documents the users initially hoped to see.

Do users really have the same $C(i)$, regardless of T?

Wouldn't it be better to take T into account when determining $C(i)$?

Hang on . . .

Let T be the user goal, the number of relevant documents the users initially hoped to see.

Do users really have the same $C(i)$, regardless of T?

Wouldn't it be better to take T into account when determining $C(i)$?

Example:

	Navigational	Exploratory
Query	"fasilkom ui"	"general relativity"
Target	$T = 1$	$T = 5$

Do you think that $C(i; T = 1) = C(i; T = 5)$?

What should the $C(i)$'s be?

Existing user models with static $C(i)$

INSQ [Moffat et al, 2012]:

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

All others being equal ...
(1) $C(i)$ increases with rank i (sunk cost investment)
(2) $C(i)$ is positively correlated with T (goal sensitive)

$$r_1 \xrightarrow{C(1)} r_2 \xrightarrow{C(2)} r_3 \xrightarrow{C(3)} r_4 \xrightarrow{C(4)} r_5 \xrightarrow{C(5)} r_6 \xrightarrow{C(6)} \cdots$$

Hang on . . .

Do users really have the same $C(i)$, regardless of what they have already seen?

Hang on . . .

Do users really have the same $C(i)$, regardless of what they have already seen?

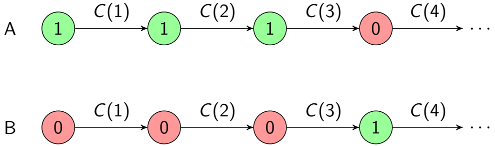
Wouldn't it be better to take r_i into account when determining $C(i)$?

Hang on . . .

Do users really have the same $C(i)$, regardless of what they have already seen?

Wouldn't it be better to take r_i into account when determining $C(i)$?

Query: "fasilkom ui"



Do you think that $C(1; A) = C(1; B)$?

What should the $C(i)$'s be?

Existing user models with adaptive $C(i)$

Let $T_i = T - \sum_{j=1}^{i} r_j$ be the unmet volume of relevance.

$$r_1 \xrightarrow{C(1)} r_2 \xrightarrow{C(2)} r_3 \xrightarrow{C(3)} r_4 \xrightarrow{C(4)} r_5 \xrightarrow{C(5)} r_6 \xrightarrow{C(6)} \cdots$$

What should the $C(i)$'s be?

Existing user models with adaptive $C(i)$

Let $T_i = T - \sum_{j=1}^{i} r_j$ be the unmet volume of relevance.

$$r_1 \xrightarrow{C(1)} r_2 \xrightarrow{C(2)} r_3 \xrightarrow{C(3)} r_4 \xrightarrow{C(4)} r_5 \xrightarrow{C(5)} r_6 \xrightarrow{C(6)} \cdots$$

INST [Moffat et al, 2017]:

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

All others being equal ...

(1) $C(i)$ increases with rank i (sunk cost investment)
(2) $C(i)$ is positively correlated with T (goal sensitive)
(3) $C(i)$ reacts to relevance found (adaptive)

What should the $C(i)$'s be?

Existing user models with adaptive $C(i)$

Other adaptive models . . .

Average Precision:

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

What should the $C(i)$'s be?

Existing user models with adaptive $C(i)$

Other adaptive models . . .

Average Precision:

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

"clairvoyant users !" Is this plausible?

What should the $C(i)$'s be?

Existing user models with adaptive $C(i)$

Other adaptive models ...

Average Precision:

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

"clairvoyant users !" Is this plausible?

The first version of Information Foraging Model [Azzopardi et al, 2018]:

$$C(i) = 1 - \left(1 + b_1 \cdot \exp^{(T_i \cdot R_1)}\right)^{-1} .$$

R_1 and b_1 are additional parameters.

Inferring $\hat{C}(i)$

These are what we believe about $C(i)$:
(1) $C(i)$ increases with rank i (sunk cost investment)
(2) $C(i)$ is positively correlated with T (goal sensitive)
(3) $C(i)$ has a positive relationship with T_i (adaptive)

Inferring $\hat{C}(i)$

These are what we believe about $C(i)$:
(1) $C(i)$ increases with rank i (sunk cost investment)
(2) $C(i)$ is positively correlated with T (goal sensitive)
(3) $C(i)$ has a positive relationship with T_i (adaptive)

These are hypotheses!

Inferring $\hat{C}(i)$

These are what we believe about $C(i)$:
(1) $C(i)$ increases with rank i (sunk cost investment)
(2) $C(i)$ is positively correlated with T (goal sensitive)
(3) $C(i)$ has a positive relationship with T_i (adaptive)

These are hypotheses!

Empirical $\hat{C}(i)$ is needed to develop evidence for or against these hypotheses.

RQ 1: How to infer empirical $\hat{C}(i)$ from search interaction logs ?

Inferring $\hat{C}(i)$ from logged behaviours

Three operational definitions of *continuation*:

- Rule L assigns non-continuation to the final impression.
- Rule M assigns all occurrences of the maximum rank as being non-continuations.
- Rule G assigns continuation to any impression that is succeeded by one at a higher ranking position.

Example: Consider the impression sequence

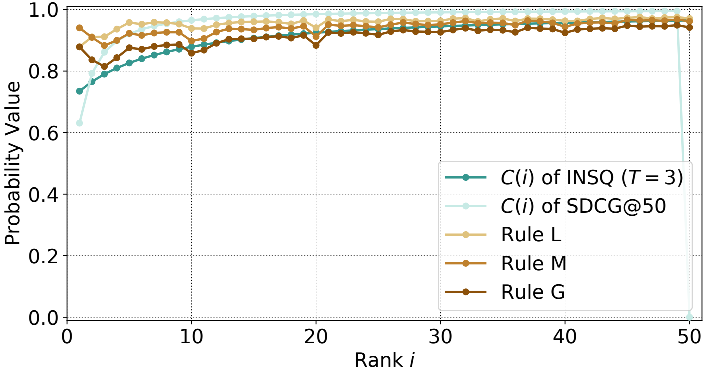
$$\langle 1, 2, 1, 4, 5, 6, 1, 3, 4, 6, 5 \rangle .$$

Can you spot all continuations (for rule L, M, and G)?

$\hat{C}(i)$ is computed using maximum likelihood estimation via these three rules.

Inferring $\hat{C}(i)$ from logged behaviours

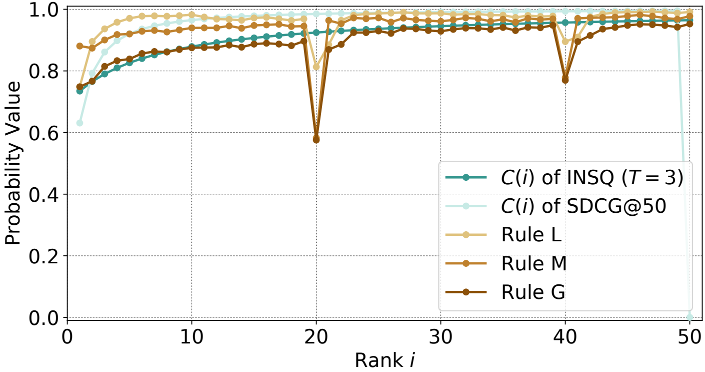
Inferred $\hat{C}(i)$ for mobile-based queries (infinite scrolling)



[Wicaksono and Moffat, CIKM 2018]

Inferring $\hat{C}(i)$ from logged behaviours

Inferred $\hat{C}(i)$ for browser-based queries (pagination)



[Wicaksono and Moffat, CIKM 2018]

Factors affecting $C(i)$

Potential factors for $C(i)$:

(1) Rank i,

(2) The user's target T,

(3) The unmet volume of relevance, $T_i = T - \sum_{j=1}^{i} r_j$

Factors affecting $C(i)$

Potential factors for $C(i)$:

(1) Rank i,

(2) The user's target T,

(3) The unmet volume of relevance, $T_i = T - \sum_{j=1}^{i} r_j$

How to infer T and r_i from the SEEK dataset?

(1) "job application" at rank i is observed $\rightarrow r_i = 1$

Factors affecting $C(i)$

Potential factors for $C(i)$:

(1) Rank i,
(2) The user's target T,
(3) The unmet volume of relevance, $T_i = T - \sum_{j=1}^{i} r_j$

How to infer T and r_i from the SEEK dataset?

(1) "job application" at rank i is observed $\rightarrow r_i = 1$
(2) T can be inferred from the number of job applications in an action sequence

Impressions may not always be observable

Suppose we want to compute $\hat{C}(i)$ from other resources ...

Impressions may not always be observable

Suppose we want to compute $\hat{C}(i)$ from other resources ...

We have an access to two commercial Web search logs:

- A sample of 105,000 queries from `Bing.com` search logs (Thanks to Paul Thomas from Microsoft),
- A sample of 1 million queries from `Yandex.ru` search logs.

Impressions may not always be observable

Suppose we want to compute $\hat{C}(i)$ from other resources ...

We have an access to two commercial Web search logs:

- A sample of 105,000 queries from `Bing.com` search logs (Thanks to Paul Thomas from Microsoft),
- A sample of 1 million queries from `Yandex.ru` search logs.

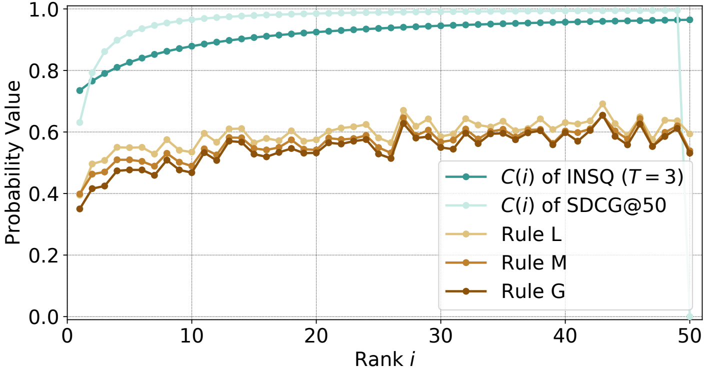
Neither of them has impressions!

All they have are click sequences.

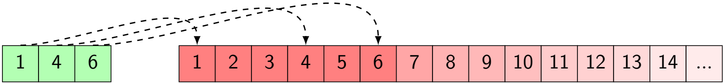
RQ 2: How to compute empirical $\hat{C}(i)$ from clicks?

Can't we just use click sequences?

Clicks are not a direct surrogate for impressions.



Impression model
[Wicaksono et al., ECIR 2019]



Click sequence

Inferred impression distributions

Suppose $\hat{V}(i \mid u, q)$ is the probability that user u viewed the item listed at rank i for query q.

Using click-through data, this can be estimated as:

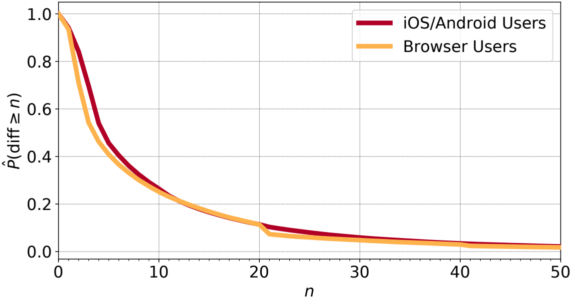
$$\hat{V}(i \mid u, q) = \begin{cases} 1 & i \leq \mathrm{dc} \\ P(\mathrm{diff} \geq i - \mathrm{dc}) & \text{otherwise}, \end{cases}$$

where *diff* is the difference between the deepest click rank (dc) and the deepest impression rank.

Impression model

[Wicaksono et al., ECIR 2019]

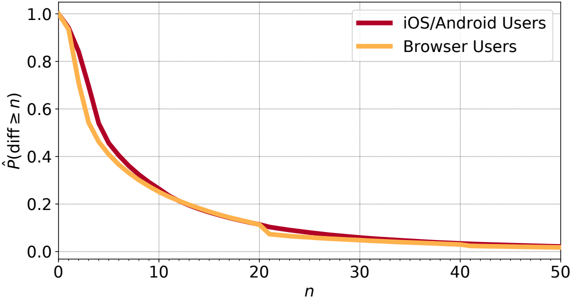
Empirical $\hat{P}(\text{diff} \geq n)$ observed from the data:



Impression model

[Wicaksono et al., ECIR 2019]

Empirical $\hat{P}(\text{diff} \geq n)$ observed from the data:



$$P(\text{diff} \geq n) = e^{-n/K}$$

Impression model
[Wicaksono et al., ECIR 2019]

Impression Model 1: K is a single variable that needs to be estimated.

Impression model
[Wicaksono et al., ECIR 2019]

Impression Model 1: K is a single variable that needs to be estimated.

Further investigation also suggests that diff is:
(1) positively correlated with the deepest click rank (dc), and
(2) is negatively correlated with the number of clicks (nc).

Impression Model 2:

$$K = g(w_0 + \mathrm{dc} \cdot w_1 + \mathrm{nc} \cdot w_2)\,,$$

where $g(.)$ is a "softplus" function, and $\{w_0, w_1, w_2\}$ is a set of parameters.

Using impression models to infer $\hat{C}(i)$

[Wicaksono et al., ECIR 2019]

Weighted mean squared error (WMSE) between the "true" $\hat{C}(i)$ values computed from impression sequences and the $\hat{C}(i)$ values estimated using impression models.

Model	WMSE (top-20)		WMSE (top-50)	
	Micro	Macro	Micro	Macro
Clicks	172.5×10^{-3}	179.1×10^{-3}	169.3×10^{-3}	175.4×10^{-3}
ZPM	5.7×10^{-3}	4.1×10^{-3}	4.5×10^{-3}	3.3×10^{-3}
AWTC	4.1×10^{-3}	2.5×10^{-3}	3.4×10^{-3}	2.1×10^{-3}
Model 1	4.0×10^{-3}	2.5×10^{-3}	3.1×10^{-3}	2.0×10^{-3}
Model 2	2.2×10^{-3}	1.2×10^{-3}	1.8×10^{-3}	1.0×10^{-3}

Model 2 significantly outperformed the other approaches (Wilcoxon signed-rank test, $p < 0.01$).

Multi-query sessions

A user may submit more than one query to address a single information need.

Finding a job as a teacher:
"teacher"

Multi-query sessions

A user may submit more than one query to address a single information need.

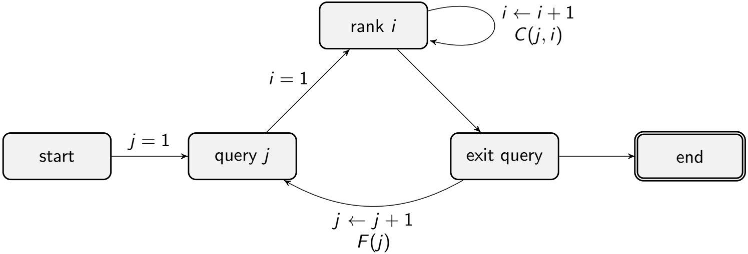
Finding a job as a teacher:
"teacher" $\rightarrow$ "teacher science"

Multi-query sessions

A user may submit more than one query to address a single information need.

Finding a job as a teacher:
"teacher" $\rightarrow$ "teacher science" $\rightarrow$ "teacher high school"

Session-based C/W/L
[Moffat et al., CIKM 2013]



Summary

We proposed a method for inferring empirical $C(i)$ from logged behaviours.

We have developed session-based C/W/L and proposed a new session-based metric (& user model) under this framework.

Acknowledgements

Thank you Seek.com!

We also used datasets generated by others, and we thank those people for their open approach to research.

Thanks are also due to Alistair, Justin, Damiano (RMIT), Sargol (SEEK), Paul Thomas (Microsoft), ...

This work was supported by seek.com and by the Australian Research Council (LP150100252 and DP190101113).