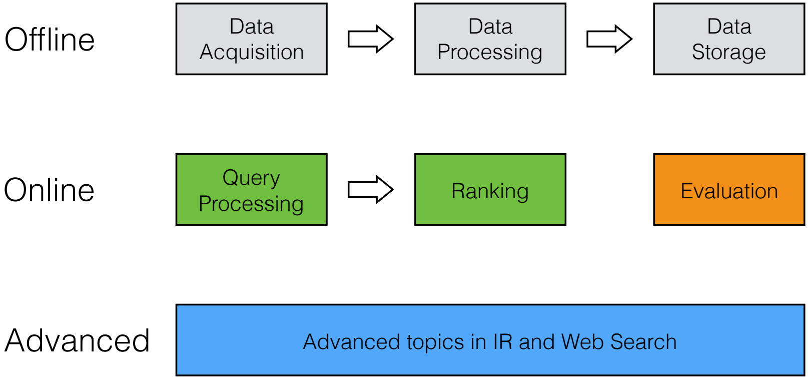
Information Retrieval

Online Evaluation and Statistical Testing

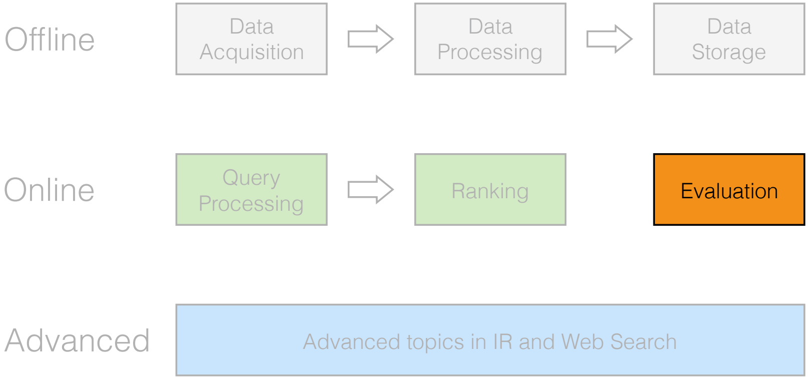
Ilya Markov
i.markov@uva.nl

University of Amsterdam

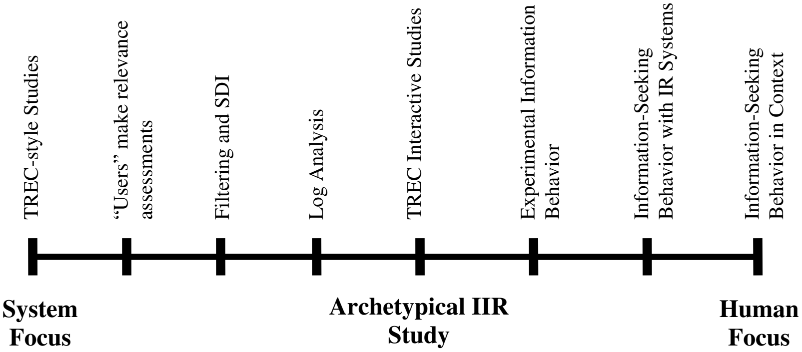
Course overview



This lecture

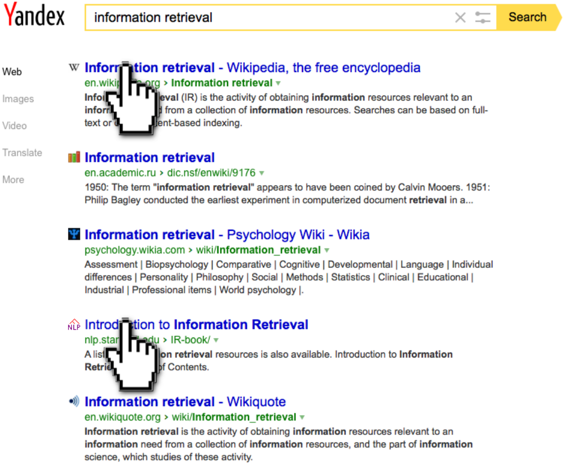


Taxonomy of evaluation approaches



Diane Kelly, "Methods for Evaluating Interactive Information Retrieval Systems with Users"

Online scenario

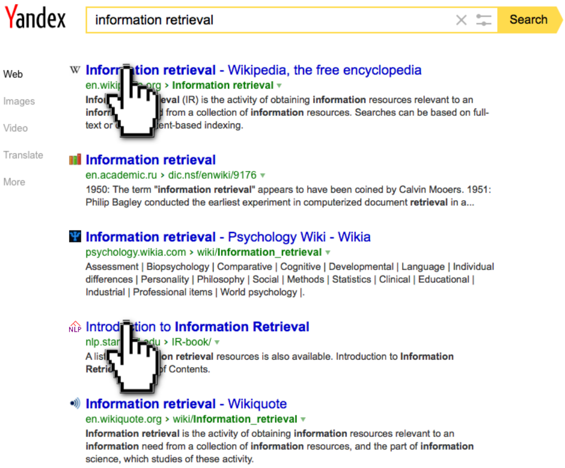


Outline

1. Online evaluation
   - Online metrics
   - Between-subject experiments
   - Within-subject experiments
   - Summary
2. Hypothesis testing

Outline

Which user-generated signals indicate search quality?



Classification of online measures

Evaluation Method	Absolute	Relative
Item Level	Click-through rate, ...	Click-skip, ...
SERP Level	Abandonment rate, ...	A/B Testing, Interleaving, ...

Evangelos Kanoulas, "A Short Survey on Online and Offline Methods for Search Quality Evaluation"

Online metrics

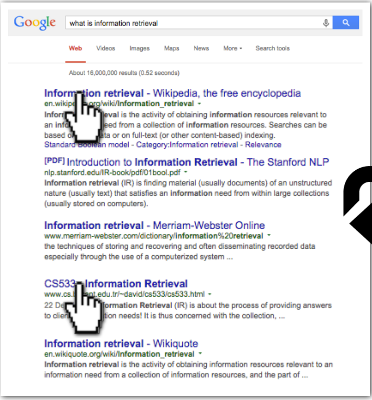
Type of interaction	Metric	Good	Bad
Clicks	Click-through rate	$\uparrow$	$\downarrow$
	Click rank (reciprocal rank)	$\downarrow$	$\uparrow$
	Abandonment	$\downarrow$	$\uparrow$
Time	Dwell time	$\uparrow$	$\downarrow$
	Time to first click	$\downarrow$	$\uparrow$
	Time to last click	$\uparrow$	$\downarrow$
Queries	Number of reformulations	$\downarrow$	$\uparrow$
	Number of abandoned queries	$\downarrow$	$\uparrow$

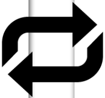
Outline

1. Online evaluation
   - Online metrics
   - Between-subject experiments
   - Within-subject experiments
   - Summary

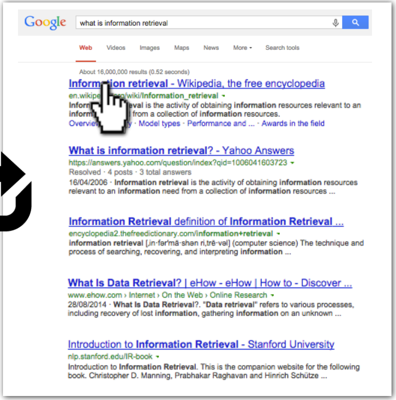
A/B testing

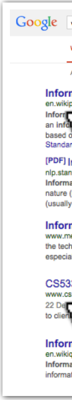
Control





Treatment





A/B testing

1. Set the current search system as control
2. Set an alternative search system as treatment
3. Assign 0.5%-1.0% of users to each of these systems
4. Record user interactions with these systems during time period T
5. Compare the systems using online metrics
6. Choose a winner based on one or several metrics

A/B testing, practical considerations

- Choosing metrics
- Control extraneous factors
- Estimate adequate sample size
- Novelty impact
- Etc.

A/B testing discussion

- Pros
  - Can evaluate anything
  - Using any online metric
- Cons
  - High variance between users
  - Not very sensitive
  - Needs lots of observations

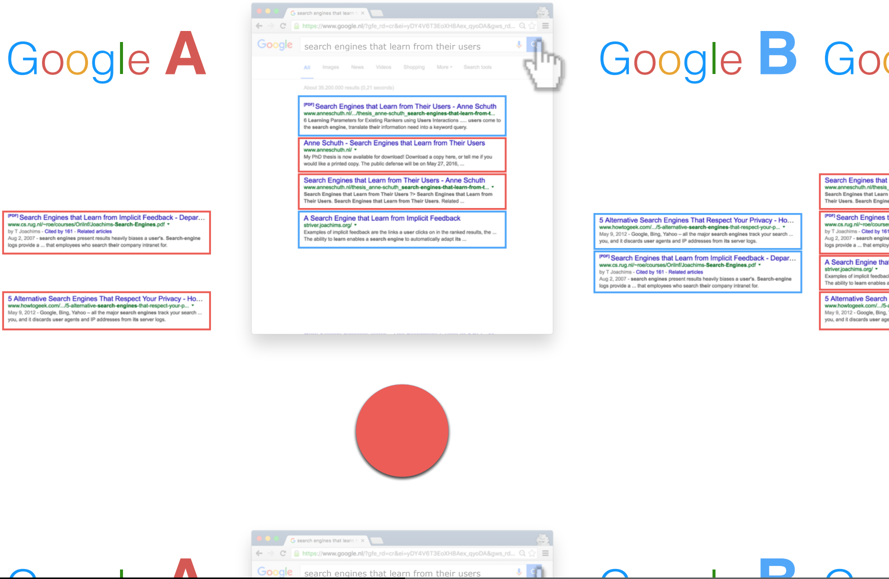
Outline

1. Online evaluation
   - Online metrics
   - Between-subject experiments
   - Within-subject experiments
   - Summary

Interleaving

1. Given a user's query, produce two rankings (current and alternative)
2. Merge the rankings into a single ranking using a **mixing policy**
3. Present the merged ranking to a user and collect interactions (see online metrics)
4. Choose a winning ranking using a **scoring rule**
5. Repeat steps 1–4 until a clear winner is identified

Team draft interleaving



Team draft interleaving

- **Mixing policy:** each ranker selects its highest ranked document that is not yet in the combined list
- **Scoring rule:** a ranker is preferred if its results get more clicks

Other interleaving methods

- Probabilistic interleaving
- Optimized interleaving
- Multileaving

Interleaving discussion

- Pros
  - No variance due to different users
  - Highly sensitive
  - Needs much fewer observations compared to A/B testing
- Cons
  - Can only use document-level metrics

Outline

1. Online evaluation
   - Online metrics
   - Between-subject experiments
   - Within-subject experiments
   - Summary

Online evaluation summary

- Online metrics
- Between-subject experiments – A/B testing
- Within-subject experiments – interleaving

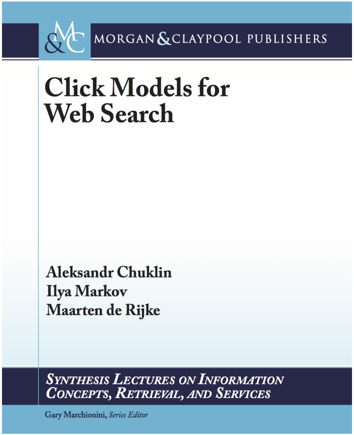
What are the advantages of online evaluation?

- Based on real users
- Cheap as it uses a running search system

What are the disadvantages of online evaluation?

- Online metrics are difficult to interpret
- May disturb users
- Cannot run too many experiments in parallel
- User search interactions are biased

Click models



http://clickmodels.weebly.com/the-book.html

Materials

K. Hofmann, L. Li, F. Radlinski
Online Evaluation for Information Retrieval
Foundations and Trends in Information Retrieval, 2016

Evaluating efficiency

Metric name	Description
Elapsed indexing time	Measures the amount of time necessary to build a document index on a particular system.
Indexing processor time	Measures the CPU seconds used in building a document index. This is similar to elapsed time, but does not count time waiting for I/O or speed gains from parallelism.
Query throughput	Number of queries processed per second.
Query latency	The amount of time a user must wait after issuing a query before receiving a response, measured in milliseconds. This can be measured using the mean, but is often more instructive when used with the median or a percentile bound.
Indexing temporary space	Amount of temporary disk space used while creating an index.
Index size	Amount of storage necessary to store the index files.

Croft et al., "Search Engines, Information Retrieval in Practice"

Outline

Example

Query	A	B	B-A
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

- How can we be sure that B is better than A?
- Test statistical significance (hypothesis testing)

Croft et al., "Search Engines, Information Retrieval in Practice"

Outline

2 Hypothesis testing
- Basics
- Hypothesis testing in IR

Test procedure

1. Set null hypothesis H_0
2. Set alternative hypothesis H_1
3. Collect sample data $\mathbf{X} = \{X_1, \ldots, X_n\}$
   - $\mathbf{X}$ is **unlikely** under $H_0 \implies$ reject H_0 in favor of H_1
   - $\mathbf{X}$ is **not unlikely** under $H_0 \implies$ no evidence against H_0
   - . . . but this is not an evidence in favor of H_0!

Example

1. $H_0 : p_{head} = 0.8$
2. $H_1 : p_{head} \neq 0.8$
3. Perform 10 tosses, observe 4 heads

$$\frac{n!}{h!(n-h)!} p^h (1-p)^{n-h} = \frac{10!}{4!6!} 0.8^4 0.2^6 = 0.005$$

4. Perform 10 tosses, observe 7 heads

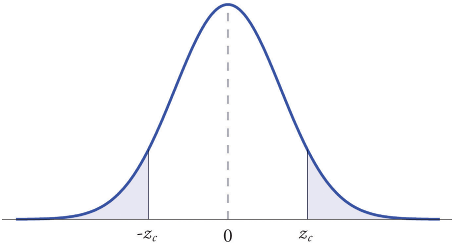
$$\frac{10!}{7!3!} 0.8^7 0.2^3 = 0.201$$

Test procedure (cont'd)

1. Consider a statistical model
2. Set H_0 and H_1
3. Choose a test statistics $T(X_1, \ldots, X_n)$
4. Choose a critical region C (discussed next)
5. Decision rule
   - $T \in C \implies$ reject H_0 in favor of H_1
   - $T \notin C \implies$ fail to reject H_0

Example (cont'd)

- $H_0 : p_{head} = 0.8$
- $H_1 : p_{head} \neq 0.8$
- Sample mean
  $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- For large enough n
  $\overline{X} \mid H_0 \sim \mathcal{N}(0.8, \sigma^2/n)$
- Test statistics
  $T = \frac{\overline{X} - 0.8}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$
- Critical region
  $C = (-\infty, -z_c] \cup [z_c, \infty)$



- $P(T \in C | H_0) \leq \alpha$
- α – size, usually $\in \{0.01, 0.05\}$
- For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$

Picture taken from http://2012books.lardbucket.org/books/beginning-statistics/s09-04-areas-of-tails-of-distribution.html

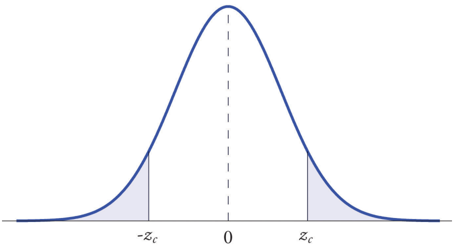
Example (cont'd)

Reject H_0 if

- $T \leq -z_{\alpha/2}$, $T \geq z_{\alpha/2}$
- $\overline{X} \leq -z_{\alpha/2} \cdot \sigma/\sqrt{n} + 0.8$,
  $\overline{X} \geq z_{\alpha/2} \cdot \sigma/\sqrt{n} + 0.8$

Alternatively, reject H_0 if

- $p = P(|T| > T_{obs} \mid H_0) \leq \alpha$
- p-value



Test statistics and p-value can be calculated using any statistical software, e.g., R

Errors

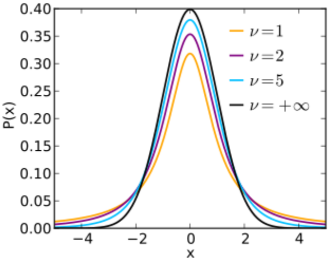
	H_0 is false	H_0 is true
Reject H_0	power	type I error (α)
Not reject H_0	type II error	

Outline

2. Hypothesis testing
    - Basics
    - Hypothesis testing in IR

T-test

1. Get measurements for systems A and B
$M(A) \sim \mathcal{N}(\mu_A, \sigma^2)$, $M(B) \sim \mathcal{N}(\mu_B, \sigma^2)$
2. $H_0 : \mu_A = \mu_B$
3. $H_1 : \mu_A \neq \mu_B$
4. $T = \frac{\overline{A} - \overline{B}}{\hat{\sigma}/\sqrt{n}} \sim \mathcal{T}^{(n-1)}$ – Student's t-distribution
5. Use standard hypothesis testing procedure with $\alpha \in \{0.01, 0.05\}$



Picture taken from https://en.wikipedia.org/wiki/Student%27s_t-distribution

T-test example

Query	A	B	B-A
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

- $\overline{B} - \overline{A} = 21.4$
- $\hat{\sigma} = 29.1$
- $T = \frac{21.4}{29.1/\sqrt{10}} = 2.33$
- $p = P(|T| > 2.33 \mid H_0) = 0.02$
- If $\alpha = 0.05$, reject H_0
- If $\alpha = 0.01$, do not reject H_0

Croft et al., "Search Engines, Information Retrieval in Practice"

Wilcoxon signed-ranks test

1. Get measurements for systems A and B
2. For each item i, compute $|m_{A,i} - m_{B,i}|$ and $sgn(m_{A,i} - m_{B,i})$
3. Exclude items with $|m_{A,i} - m_{B,i}| = 0$
4. Order the remaining N_{nz} items based on $|m_{A,i} - m_{B,i}|$
5. Assign ranks R_i from smallest to largest
6. Compute the test statistics

$$W = \sum_{i=1}^{N_{nz}} [sgn(m_{A,i} - m_{B,i}) \cdot R_i]$$

7. For large N_{nz}, $W \sim \mathcal{N}$
8. Use standard hypothesis testing procedure with $\alpha \in \{0.05, 0.01\}$

Wilcoxon test example

Query	A	B	B-A
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

- Ranked non-zero differences
  2, 9, 10, 24, 25, 25, 41, 60, 70
- Signed ranks
  -1, +2, +3, -4, +5.5, +5.5, +7, +8, +9
- $W = 35$
- $p = P(|W| > 35 \mid H_0) = 0.025$
- If $\alpha = 0.05$, reject H_0
- If $\alpha = 0.01$, do not reject H_0

Croft et al., "Search Engines, Information Retrieval in Practice"

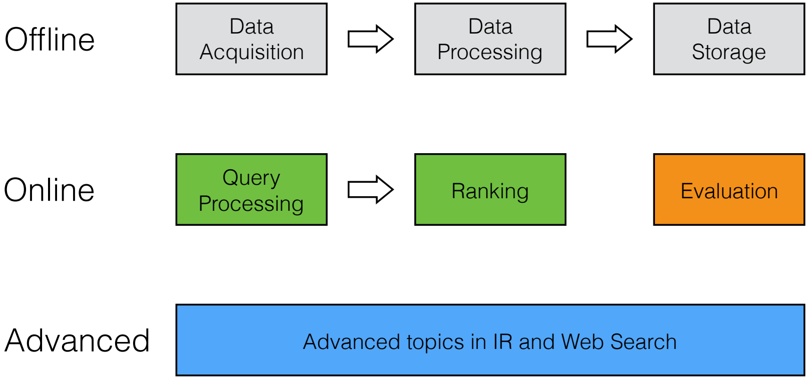
Hypothesis testing summary

- IR must use statistical testing
- The most common and one of the most powerful is the **paired t-test**

Materials

M. Smucker, J. Allan, B. Carterette
A Comparison of Statistical Significance Tests for Information Retrieval Evaluation
Proceedings of CIKM, pages 623–632, 2007

Course overview



See you tomorrow at 14:30