11-442 / 11-642 / 11-742:
Search Engines

Evaluating Search Effectiveness

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

1

Overview of the Evaluation Unit

Introduction to evaluation

The Cranfield methodology

- Overview and introduction
- Test collections
- Metrics

Creating test collections

- Cranfield @ TREC and other evaluation forums
- Cranfield @ work

Evaluation in a dynamic environment

2

The Cranfield Methodology: Creating Test Collections

Two methods of creating test collections are common

1. **Developed by a community (e.g., a research community)**
   – Usually designed to be useful for a long time ("reusable")
   – Must accommodate today's system(s) and future systems
   – Higher effort, higher expense

2. **Developed by an organization (e.g., a company)**
   – Usually designed to address specific needs
   – Usually lower effort, lower expense
   – Usually a short lifespan

Cranfield@TREC

The U.S. National Institute of Standard and Technologies (NIST) supports scientific and commercial progress by defining state-of-the-art measurement capabilities

In 1992, NIST began providing resources for large-scale evaluation of text retrieval

- Annual production of tasks and test collections
- The Text REtrieval Conference (TREC)
  – An annual forum for comparison of methods and results
- Most TREC evaluation is based on the Cranfield methodology

Cranfield@TREC

Each year, TREC defines a set of tasks ("tracks")

TREC 2019 tracks

- **Complex answer:** Integrate info from multiple sources
- **Conversational assistance:** Search in dialogue systems (e.g., Siri)
- **Decision:** Search that helps people make decisions (e.g., health)
- **Deep learning:** Train your neural system with a lot of data
- **Fair ranking:** Relevance + fairness (representativeness)
- **Incident streams:** Analyze social media in emergencies
- **News:** Search of news (Washington Post)
- **Precision medicine:** Link oncology patients to clinical trials

5

5

Cranfield@TREC: Creating Test Collections

Most TREC tracks produce test collections

- The research community defines a task
  - E.g., Microblog retrieval
- NIST works with researchers to obtain a document collection
- NIST defines information needs and queries
  - Sometimes in collaboration with industry or other groups
- The research community identifies documents to be judged
  - Pooling: Run your favorite technique, submit your results
- NIST employees and/or participants judge the documents

6

6

Cranfield@TREC: International Siblings

TREC is a community-driven approach to creating datasets

- NIST enables creation, but does not do all of the work itself

Other regions have adopted this approach to creating data

- CLEF (Europe)
  - Originally cross-lingual retrieval, now many other topics
- NTCIR (Japan)
  - Originally Asian languages, now also other topics
- FIRE (India)
  - Originally Indian languages, now also other topics

7

7

Cranfield@TREC: Summary

TREC test collections are designed to be reusable

- The pool of judged documents is large and diverse
- Why is this important?
  - It enables accurate measurements for techniques that were not in the assessment pool
- (Most) TREC collections accommodate today's system(s) and future systems
- Reusability is an essential property of TREC collections

The lifespan of a typical TREC test collection is 5-10 years

- Some datasets have been used for 20+ years

8

8

Overview of the Evaluation Unit

Introduction to evaluation

The Cranfield methodology

- Overview and introduction
- Test collections
- Metrics

Creating test collections

- Cranfield @ TREC and other evaluation forums
- Cranfield @ work

Evaluation in a dynamic environment

9

Cranfield@Work

TREC collections may not cover your particular needs

- E.g., because you use proprietary information
- E.g., because the source of information is new

You may need to create your own test collection

- This happens all the time
  - In industry
  - In research environments (such as ours)

What factors must you consider?

10

Cranfield@Work: How Many Information Needs Are Needed?

Suppose that you are building your own corpus …how many information needs do you need?

- Typical heuristics
  - 25 provides a rough estimate
  - 50 is relatively reliable
  - 100 is reliable
  - 200 is very reliable

Are these heuristics valid? What is our goal?

- To calculate MAP reliably?
- To distinguish among systems reliably?

Cranfield@Work: How Many Information Needs Are Needed?

Evaluation based on a few information needs is unreliable

Info Needs	MAP	Standard Deviation	95% Confidence Interval
5	0.172	0.039	[0.095, 0.250]
10	0.276	0.020	[0.238, 0.315]
25	0.265	0.029	[0.208, 0.321]
50	0.260	0.020	[0.221, 0.299]
100	0.290	0.014	[0.263, 0.317]
148	0.287	0.014	[0.259, 0.315]

Indri
Gov2
BOW queries

The population changes

Usually 50 information needs is considered "good enough"

- 100-200 information needs is considered very reliable

Cranfield@Work: Confidence Intervals

Example

- MAP = 0.283
- N = 25 (information needs)
- Standard deviation = 0.025
- $CI_{95\%}$ = [Mean – $ZValue_{95\%}$ × StdDev, Mean + $ZValue_{95\%}$ × StdDev]
  = [0.283 – 1.96 × 0.025, 0.283 + 1.96 × 0.025]
  = [0.234, 0.332]
  - 95% of samples will have MAP ∈ [0.234, 0.332]
  - It does not mean that the true MAP ∈ [0.234, 0.332]

Cranfield@Work: How Many Information Needs Are Needed?

Usually 50 information needs is considered "good enough"

- Good enough to identify the best system relatively reliably
- Maybe not good enough to provide a reliable estimate of MAP
- 100-200 information needs is considered very reliable

Industry often uses hundreds of information needs (queries)

Why this difference?

- Researchers have fewer resources
- Small differences can be important to industry, but are less useful to researchers

Cranfield@Work:
Reliability of Relevance Assessments

A relevant document is one that a person judges as useful in the context of a specific information need

- Does it matter that people judge relevance differently?

Common complaint: The relevance judgments don't measure my system fairly

- Because the assessor made mistakes on some documents
- Because some highly-ranked documents were not judged
  – And thus are considered non-relevant by trec_eval

Is this complaint justified?

15

15

Cranfield@Work:
How Do Three TREC Assessors Compare?

Documents judged not relevant by all 3 assessors | **Documents judged relevant by all 3 assessors**

Relevance Assessments

Three TREC Assessors									
A_1	NR	NR	NR	NR	R	R	R	R	
A_2	NR	NR	R	R	NR	NR	R	R	
A_3	NR	R	NR	R	NR	R	NR	R	Judged
TREC Topics (Info Needs)									
202	32	168	0	0	21	127	1	51	400
203	194	1	4	1	20	3	4	6	233
204	138	55	1	6	119	39	8	34	400
205	200	0	0	0	119	20	59	2	400
206	200	0	0	0	17	6	16	8	247
207	171	7	5	17	6	16	3	49	274

Major disagreement

(Voorhees and Over)

16

16

Cranfield@Work: Does it Matter Which Assessments You Use?

Switching assessments affects objective evaluations

- Precision, Recall, MRR, R-Prec, …
- Objective evaluations describe the user experience

Does switching assessments affect comparative evaluations?

- system A vs. system B
- system A vs. system A′

A study used TREC data to answer this question

17

17

Cranfield@Work: Does it Matter Which Assessments You Use?

Use TREC assessments to generate 100,000 artificial assessors

- A_1: Relevant if A_1 calls it relevant
- …
- A_4: Relevant if A_1 and A_2 and A_3 call it relevant ("union")
- A_5: Relevant if A_1 or A_2 or A_3 call it relevant ("intersection")
- A_6: Use A_1 for q_1, A_2 for all other queries
- …

Use each artificial assessor to rank systems (n rankings)

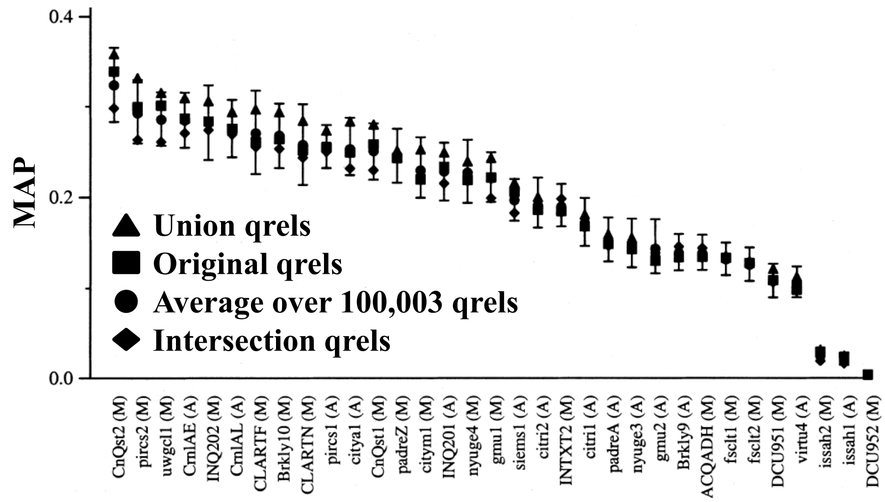
- Compare the rankings produced by each set of assessments
- How well do they agree about which systems are best/worst?

(Voorhees, 1998)

18

18

Cranfield@Work:
Does it Matter Which Assessments You Use?



(Voorhees, 1998)

19

19

Cranfield@Work:
Does it Matter Which Assessments You Use?

Significant overlap in the bars, so this looks bad … is it?

System rankings are <u>very</u> similar with different assessments

- On average, swap 3% of entries to convert between rankings
- Most swaps are between systems that have ΔMAP < 1%
- Probability of a swap is very low if ΔMAP is ≥ 0.05

Systems tend to move together

- A set of assessments affects most systems in the same way
  - "Easy" assessors, "hard" assessors

(Voorhees, 1998)

20

20

Cranfield@Work: Creating Test Collections

So, you're evaluating search engines for some organization…
…how do you build them a test collection?

1. **Collect a large set of representative documents**
   – Easy
2. **Collect a set of representative information needs**
   – At least 25, preferably 50-100
3. **Translate each information need into a set of queries**
   – At least several queries per information need

21

21

Cranfield@Work: Building Your Own Test Collection

4. **Run each query against each search engine**
   – Save the top N documents
   – Preferably at least 50 documents per query
5. **Pool all results for an information need**
   – Different queries, different engines
   – Sort them into random order
6. **Have a person judge each document**
   – One person judges all documents for one information need
      » **Important!:** The work can't be split among people
   – Ideally, the judge created the information need

22

22

Overview of the Evaluation Unit

Introduction to evaluation

The Cranfield methodology

- Overview and introduction
- Test collections
- Metrics

Creating test collections

- Cranfield @ TREC and other evaluation forums
- Cranfield @ work

Evaluation in a dynamic environment

Evaluation in a Dynamic Environment

Web search engines do use the Cranfield methodology

- They have use trained assessors, similar to NIST

But, they also use other metrics and methodologies…

- It would be too expensive to apply the Cranfield methodology to a large percentage of query volume
- Information needs for most queries are unknown
- The document collection is dynamic
- Clicks are not relevance judgments

Evaluation in a Dynamic Environment: Interleaved Testing Procedure

One trial

- User submits query
- Select two rankers ("A" and "B")
- Interleave the rankings produced by "A" and "B"
- Track the user's clicks on the interleaved document ranking
- When the user stops clicking
  - Assign credit to "A" and "B" based on clicks
  - Declare "A" or "B" the winner of this trial

Repeat until enough trials are collected

- Each trial is a different query and a different user

Evaluation in a Dynamic Environment: Interleaved Testing

Requirements for an interleaving procedure

- The user should not notice it
- It should be robust to user biases
- It shouldn't alter the search experience
- It should lead to user behavior that reflects user preferences

We consider two interleaving methods

- Balanced interleaving
- Team-draft interleaving

There are other methods, but this gives you the general idea

Evaluation in a Dynamic Environment: Balanced Interleaving

Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
$I \leftarrow (); k_a \leftarrow 1; k_b \leftarrow 1;$
$AFirst \leftarrow RandomBit()$ .... *decide which ranking gets priority*
while $(k_a \leq |A|) \wedge (k_b \leq |B|)$ **do** .... *if not at end of A or B*
  if $(k_a < k_b) \vee ((k_a = k_b) \wedge (AFirst = 1))$ **then**
    if $A[k_a] \notin I$ **then** $I \leftarrow I + A[k_a]$ .... *append next A result*
    $k_a \leftarrow k_a + 1$
  else
    if $B[k_b] \notin I$ **then** $I \leftarrow I + B[k_b]$ .... *append next B result*
    $k_b \leftarrow k_b + 1$
  end if
end while
Output: Interleaved ranking I

- **Decide (once) which method goes first**
- **When a duplicate document is found, increment the counter**
  - But, the document is not added to the interleaved ranking

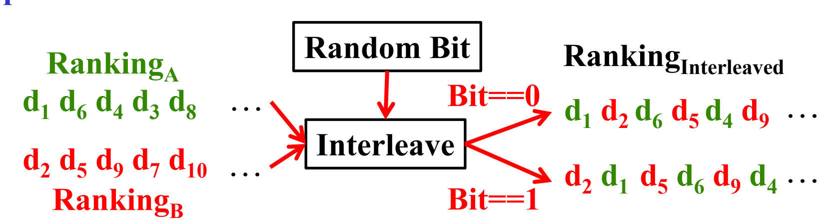
(Chapelle et al, 2012)

27

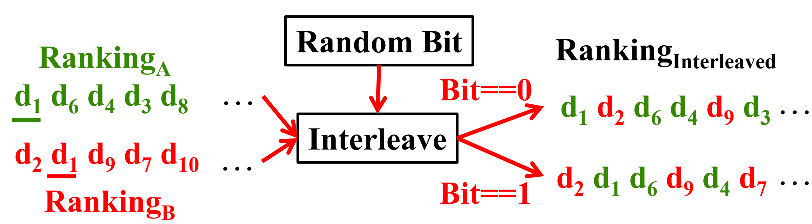
27

Evaluation in a Dynamic Environment: Balanced Interleaving

Without duplicates



With duplicates



28

28

Evaluation in a Dynamic Environment: Balanced Interleaving

Assume that people read from top to bottom

- They click on documents that look interesting
- They stop when they are satisfied or frustrated

At each rank, each method contributes about 50% of the documents

- **Fair:** Each method has an equal opportunity to present documents
- A random clicker would click equally on documents from each method

29

Evaluation in a Dynamic Environment: Balanced Interleaving

Given an interleaved ranking I with clicks C

- c_{max}: Rank of the last click (the last document viewed)

Use rankings to depth $k = \min\{ j: (i_{c_{max}}=a_j) \vee (i_{c_{max}}=b_j) \}$

- $\{a_1, .., a_k\} \cup \{b_1, .., b_k\}$ covers all docs in $\{i_1, .., i_{c_{max}}\}$
- $\# clicks_a = |c_j : i_{c_j} \in \{a_1, \ldots, a_k\}|$ **clicks on a's top k**
- $\# clicks_b = |c_j : i_{c_j} \in \{b_1, \ldots, b_k\}|$ **clicks on b's top k**

The method that gets the most clicks wins the trial

Aggregate results for all trials to find the best ranker

$$\Delta(A,B) = \frac{wins(A) + 0.5 \times ties(A,B)}{wins(A) + wins(B) + ties(A,B)}$$

I	C
i_1	
i_2	c_1
i_3	
i_4	
i_5	c_2
i_6	
i_7	
i_8	c_{max}
i_9	
:	

(Chapelle et al, 2012)

30

Evaluation in a Dynamic Environment: Balanced Interleaving

Example

- Clicked: ✓
- c_{max}=3
  - Rank of last clicked doc
- k=2
  - Min depth needed to find last clicked doc in R_1 or R_2
- # $clicks_{R_1}$=1
- # $clicks_{R_2}$=2

R_2 wins this trial

k=2

Rank	Input Ranking R_1	Input Ranking R_2	Interleaved Ranking R_1 first
1	a	b	a
2	b	e	b ✓
3	c	a	e ✓
4	d	f	c
5	g	g	d
6	h	h	f
:	:	:	:

(Chapelle et al, 2012)

31

31

Evaluation in a Dynamic Environment: Balanced Interleaving

Example

- Clicked: ✓
- c_{max}=3
  - Rank of last clicked doc
- k=2
  - Min depth needed to find last clicked doc in R_1 or R_2
- # $clicks_{R_1}$=1
- # $clicks_{R_2}$=2

R_2 wins this trial

k=2

Rank	Input Ranking R_1	Input Ranking R_2	Interleaved Ranking R_2 first
1	a	b	b ✓
2	b	e	a
3	c	a	e ✓
4	d	f	c
5	g	g	f
6	h	h	d
:	:	:	:

(Chapelle et al, 2012)

32

32

Evaluation in a Dynamic Environment: Balanced Interleaving

Example

- Clicked: ✓
- c_{max}=5
  - Rank of last clicked doc
- k=4
  - Min depth needed to find last clicked doc in R_1 or R_2
- # clicks_{R_1}=2
- # clicks_{R_2}=2

This trial is a tie

Rank	Input Ranking R_1	Input Ranking R_2	Interleaved Ranking R_1 first
1	a	c	a ✓
2	b	a	c
3	c	i	b
4	d	b	i ✓
5	e	g	d ✓
6	f	e	e
:	:	:	:

k=4

(Chapelle et al, 2012)

33

33

Evaluation in a Dynamic Environment: Interleaving

Interleaving is repeated for many trials

Query	User	First Ranker	Winner
buy ipad	Hongyu	R_2	R_1
deep learning tutorial	Vallari	R_1	R_1
pittsburgh weather	Arpita	R_2	Tie
shoes	Varshini	R_2	R_1
gifts for mom	Qing	R_1	R_2
: : :		:	:

Tally results from all trials to declare a winner

$$\Delta(R_1, R_2) = \frac{wins(R_1) + 0.5 \times ties(R_1, R_2)}{wins(R_1) + wins(R_2) + ties(R_1, R_2)}$$

R_1 wins if $\Delta(R_1, R_2) > 0.5$

34

34

Evaluation in a Dynamic Environment: Balanced Interleaving

Balanced Interleaving can behave unexpectedly

- Suppose a user clicks on just <u>one</u> result randomly
- ¾ of the outcomes favor R_2

Why?

- ¾ of the documents are ranked higher by R_2 than R_1
- Cutoff k considers too little information

Rank	Input Ranking R_1	Input Ranking R_2	Balanced R_1 first	Balanced R_2 first
1	a	b	a	b
2	b	c	b	a
3	c	d	c	c
4	d	a	d	d

R_1 first

✓	C_{max}	k	Win
a	1	1	R_1
b	2	1	R_2
c	3	2	R_2
d	4	3	R_2

(Chapelle et al, 2012)

35

35

Evaluation in a Dynamic Environment: Team-Draft Interleaving

Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
Init: $I \leftarrow ()$; $TeamA \leftarrow \emptyset$; $TeamB \leftarrow \emptyset$;
while $(\exists i : A[i] \notin I) \wedge (\exists j : B[j] \notin I)$ **do** .... *if not at end of A or B*
  if $(|TeamA| < |TeamB|) \vee$
    $((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ **then**
    $k \leftarrow \min_i\{i : A[i] \notin I\}$ .... *top result in A not yet in I*
    $I \leftarrow I + A[k]$; .... *append it to I*
    $TeamA \leftarrow TeamA \cup \{A[k]\}$ .... *clicks credited to A*
  else
    $k \leftarrow \min_i\{i : B[i] \notin I\}$ .... *top result in B not yet in I*
    $I \leftarrow I + B[k]$ .... *append it to I*
    $TeamB \leftarrow TeamB \cup \{B[k]\}$ .... *clicks credited to B*
  end if
end while
Output: Interleaved ranking I, $TeamA$, $TeamB$

- **On each round, randomize which method goes first**
- **When a duplicate document is encountered, skip to the next**

(Chapelle et al, 2012)

36

36

Evaluation in a Dynamic Environment: Team Draft Interleaving

Without duplicates

Ranking_A: $d_1\ d_6\ d_4\ d_3\ d_8\ \ldots$

$d_2\ d_5\ d_9\ d_7\ d_{10}\ \ldots$:Ranking_B

Random Bits → Interleave → $\text{Ranking}_{\text{Interleaved}}$: $d_1\ d_2\ d_5\ d_6\ d_9\ d_4\ \cdots$

With duplicates

Ranking_A: $\underline{d_1}\ d_6\ d_4\ d_3\ d_8\ \ldots$

$d_2\ \underline{d_1}\ d_9\ d_7\ d_{10}\ \ldots$:Ranking_B

Random Bits → Interleave → $\text{Ranking}_{\text{Interleaved}}$: $d_2\ d_1\ d_9\ d_6\ d_4\ d_7 \cdots$

37

37

Evaluation in a Dynamic Environment: Team-Draft Interleaving

Consider an interleaved ranking I with clicks C

Clicks attributed to each method are

$\#\ \text{clicks}_a = |c_j : i_{c_j} \in \text{Team}_a|$ — **clicks on docs selected by a**

$\#\ \text{clicks}_b = |c_j : i_{c_j} \in \text{Team}_b|$ — **clicks on docs selected by b**

The method that gets the most clicks wins the trial

Aggregate results for all trials to find the best ranker

$$\Delta(A,B) = \frac{wins(A) + 0.5 \times ties(A,B)}{wins(A) + wins(B) + ties(A,B)}$$

R_1 wins if $\Delta(R_1, R_2) > 0.5$

I	C
i_1	
i_2	c_1
i_3	
i_4	
i_5	c_2
i_6	
i_7	
i_8	c_3
i_9	
:	

(Chapelle et al, 2012)

38

38

Evaluation in a Dynamic Environment: Team-Draft Interleaving

Team-Draft can behave unexpectedly

- Suppose a query has 3 intents
  - 49% of the users: a is relevant
  - 49% of the users: b is relevant
  - 2% of the users: c is relevant

Rank	Input Ranking R_1	Input Ranking R_2	Interleaved Ranking
1	a	b	b
2	b	c	a
3	:	:	c

✓ Win	
a	R_1
b	R_2
c	R_2

R1 satisfies 98% of search intents with the top 2 results

- But, if people click on just one result randomly, R_2 wins 51% of trials
  - This is an artifact of how duplicates are handled
  - Only the method that suggested the document higher gets credit
    - » R_1 gets credit for a, and R_2 gets credit for b and c

(Chapelle et al, 2012)

39

39

Evaluation in a Dynamic Environment: Team-Draft Interleaving

Team-Draft can behave unexpectedly

- Suppose a query has 3 intents
  - 49% of the users: a is relevant
  - 49% of the users: b is relevant
  - 2% of the users: c is relevant

Rank	Input Ranking R_1	Input Ranking R_2	TeamDraft R_1 First	TeamDraft R_2 First
1	a	b	a	b
2	b	c	b	a
3	:	:	c	c

✓ Win	
a	R_1
b	R_2
c	R_2

R1 satisfies 98% of search intents with the top 2 results

- But, if people click on just one result randomly, R_2 wins 51% of trials
  - This is an artifact of how duplicates are handled
  - Only the method that suggested the document higher gets credit
    - » R_1 gets credit for a, and R_2 gets credit for b and c

(Chapelle et al, 2012)

40

40

Evaluation in a Dynamic Environment: Does Interleaving Agree With Assessors?

A large evaluation was done with ArXiv.org, Bing, and Yahoo

- **ArXiv:** 700K academic articles, scientific users, 70K searches
  - Ranking functions created by degrading a baseline
- **Bing:** Team-Draft interleaving on a % of US traffic
  - Five pairs of proprietary ranking functions, 220K searches
  - 12,000 queries were also manually assessed (5-point scale)
- **Yahoo:** Balanced interleaving on a % of US traffic
  - All pairs of four proprietary ranking functions, 20M searches
  - 2,000 queries were also manually assessed

(Chapelle et al, 2012)

41

41

Evaluation in a Dynamic Environment: Does Interleaving Agree With Assessors?

ArXiv.org

- Interleaving identifies the better ranker (usually w/ significance)

Bing & Yahoo

- When assessors find a significant difference, interleaving agrees
- Interleaving may find a difference significant that assessors don't

Often interleaving can provide statistically significant results where manual assessments cannot

- A "small" number of manually-assessed queries

(Chapelle et al, 2012)

42

42

Evaluation in a Dynamic Environment: Does Interleaving Agree With Assessors?

Interleaving identifies the best ranker
… does it also indicate the magnitude of the difference?

- **Bing**
  - 0.88 correlation w/ NDCG@5 (Team-Draft)
  - 0.69 correlation w/ MAP (Team-Draft)
- **Yahoo**
  - 0.70 correlation w/ DCG@5 (Balanced)

Note that the number of queries affects the error bars

- 12,000 queries for Bing
- 2,000 queries for Yahoo

(Chapelle et al, 2012)

43

43

Evaluation in a Dynamic Environment: Metrics

Dynamic environments often use metrics based on user behavior

- **Abandonment rate:** % of queries that receive no clicks
- **Reformulation rate:** % of queries that are reformulated
- **Queries per session:** Session == Information need
- **pSAT-clicks:** % of documents with dwell time > 30 seconds
- **pSkip:** % of documents that are skipped
- **Clicks per query,** **Clicks@1**
- **Max Reciprocal Rank,** **Mean Reciprocal Rank**
- **Time to First Click,** **Time to Last Click**

(Chapelle et al, 2012)

44

44

Evaluation in a Dynamic Environment: Does Interleaving Agree With Behavior?

Interleaving does not predict <u>changes</u> in user behavior well

- E.g., Queries per Session, Abandonment Rate, …
- It predicts Clicks@1, but only with very large numbers of queries
  - The Yahoo experiment

(Chapelle et al, 2012)

Evaluation in a Dynamic Environment: How Many Queries Are Needed?

To achieve 95% confidence

- **ArXiv.org:** About 200K queries
- **Yahoo:**
  - Rankers of different quality: A few hundred thousand queries
  - Rankers of similar quality: A few million queries

Interleaving reaches significance faster than Clicks@1

- 1 hour for interleaving vs. 1 day for Clicks@1

(Chapelle et al, 2012)

Evaluation in a Dynamic Environment

More sophisticated methods of counting clicks improve the sensitivity and convergence rates for Team-Draft Interleaving

- Not covered due to lack of time
- This is an active research topic

Overview of the Evaluation Unit

Introduction to evaluation

The Cranfield methodology

- Overview and introduction
- Test collections
- Metrics

Creating test collections

- Cranfield @ TREC and other evaluation forums
- Cranfield @ work

Evaluation in a dynamic environment

Overview of the Evaluation Unit: Cranfield vs. Interleaved Evaluation

We focused more on Cranfield than interleaving … why?

- Cranfield is more established
  - It has been used for years and is well-understood
- Cranfield supports a wide variety of metrics
  - It provides better information about ranking behavior
- Cranfield can be used in most situations
  - Interleaving requires query traffic that you may not have

However, interleaving is a powerful tool, when you can use it

- Inexpensive, adaptive, sensitive to small differences

49

49

Overview of the Evaluation Unit: Cranfield vs. Interleaved Evaluation

Use the method that has the properties you need

Property	Cranfield	Interleave
Relevance == satisfying an information need	Y	Y
The assessor has the information need	Usually	Y
Requires human assessors	Y	N
Requires a large amount of query traffic	N	Y
Supports a variety of metrics	Y	Y
Sensitive to small differences among methods	N	Y
Reusable test collections	If desired	N
Dynamic test collections	N	Y
Quickly test new methods	If desired	Y

50

50

Next Semester…

This course will be offered next semester

I will need Teaching Assistants

- 8-12 hours/week × 7 weeks (“grading weeks”)
- 3-5 hours/week × 9 weeks (office hours, piazza)

Please send me email if you are interested in being a TA

- I will start TA interviews after grades are posted
  (probably Monday, May 10 or Tuesday, May 11)

51

51

For More Information

- C. Buckley and E. M. Voorhees. “Evaluating evaluation measure stability.” Proceedings of SIGIR 2000. pp. 33-40. 2000.
- C. Buckley and E. M. Voorhees. “Retrieval evaluation with incomplete information.” Proceedings of SIGIR 2004. pp. 25-32. 2004.
- O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. Transactions on Information Systems, 30(1). 2012.
- E. M. Voorhees. “Variations in relevance judgements and the measurement of retrieval effectiveness.” Proceedings of SIGIR ’98. pp. 315- 323. 1998.
- E. M. Voorhees. “Evaluation by highly relevant documents.” Proceedings of SIGIR 2001. pp. 74-82. 2001.
- E. M. Voorhees and C. Buckley. “The effect of topic set size on retrieval experiment error.” Proceedings of SIGIR 2002. pp. 316-323. 2002.
- M. Sanderson and J. Zobel. “Information retrieval system evaluation: Effort, sensitivity, and reliability.” Proceedings of SIGIR 2005. pp. 162-169.

52

52