11-442 / 11-642:
Search Engines

Evaluating Search Effectiveness

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

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

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 2015 tracks

- **Clinical decision support:** Link cases to relevant information
- **Contextual suggestions:** Suggest activities based on context
- **Dynamic domain:** Dynamic information needs
- **Live QA:** Answer questions from a live question stream
- **Microblog:** Search and filtering of Twitter data
- **Tasks:** Figure out the task a person is trying to accomplish
- **Temporal summarization:** Monitor an event over time
- **Total recall:** High recall with a human in the loop

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

Cranfield@TREC: International Siblings

Other sponsors have developed similar efforts to produce test collections focused on topics of interest to them

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

Cranfield@TREC: Summary

Characteristics of TREC test collections

- A relatively large number of queries
- Large pools of assessed documents
- Widespread use

Sort of an "open source" approach to creating datasets

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

Cranfield@TREC: Summary

TREC test collections are designed to be reusable

- The pool of judged documents is large enough and diverse enough to produce accurate measurements for techniques that did not contribute to the pool
- They must accommodate today's system(s) and future systems
- This 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

9

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

10

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?

Cranfield@Work: How Many Information Needs Are Needed?

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

- The rule of thumb has been
  - 25 provides a rough estimate
  - 50 is relatively reliable
  - 100 is reliable
  - 200 is very reliable

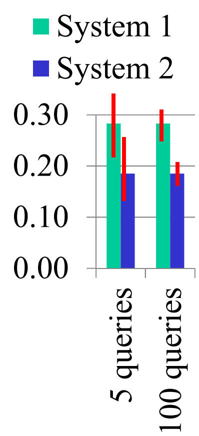
Are these heuristics valid? What is our goal?

- 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	0.283	0.056	[0.173, 0.393]
25	0.283	0.025	[0.234, 0.332]
150	0.283	0.010	[0.263, 0.303]



13

(Adapted from Sanderson & Zobel, 2005)

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

14

Cranfield@Work: Confidence Intervals

Example

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

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 important to researchers

Cranfield@Work: Relevance Assessments

Relevance is difficult to define precisely

- **A relevant document is one that <u>a person</u> judges as <u>useful</u> in the context of a <u>specific information need</u>**
  - Different people define "useful" differently
  - A person will define "useful" differently at different times
  - A person's judgment depends upon many factors
    - » E.g., what the person knew before reading the document
- **This is a <u>really important</u> concept**

Does it matter that people judge relevance differently?

17

Cranfield@Work: Reliability of Relevance Assessments

Common complaint: The relevance judgments are biased against my system

- Because the assessor made mistakes
- Because some relevant documents were not judged
  - And thus are considered non-relevant

Is this complaint justified?

18

Cranfield@Work: How Do Three TREC Assessors Compare?

R: Relevant
NR: Not relevant

Judgments

Documents judged relevant by all 3 assessors

Three TREC Assessors / TREC Topics (Info Needs)

A_1 / A_2 / A_3	NR NR NR	NR NR R	NR R NR	NR R R	R NR NR	R NR R	R R NR	R R R	Judged
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
208	171	21	1	7	6	23	2	23	254

(Voorhees and Over)

19

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

Switching assessments clearly affects objective evaluations

- Precision, Recall, MRR, R-Prec, …
- Objective evaluations describe the user experience
  - But they don't identify which technique is better

Does switching assessments affect comparative evaluations?

- i.e., system A vs. system B, or system A vs. system A'

20

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

Suppose you have 3 people judge 49 TREC topics

- You can create 3^{49} sets of relevance judgments
  - Use assessor i for topic j, generate all combinations of i and j

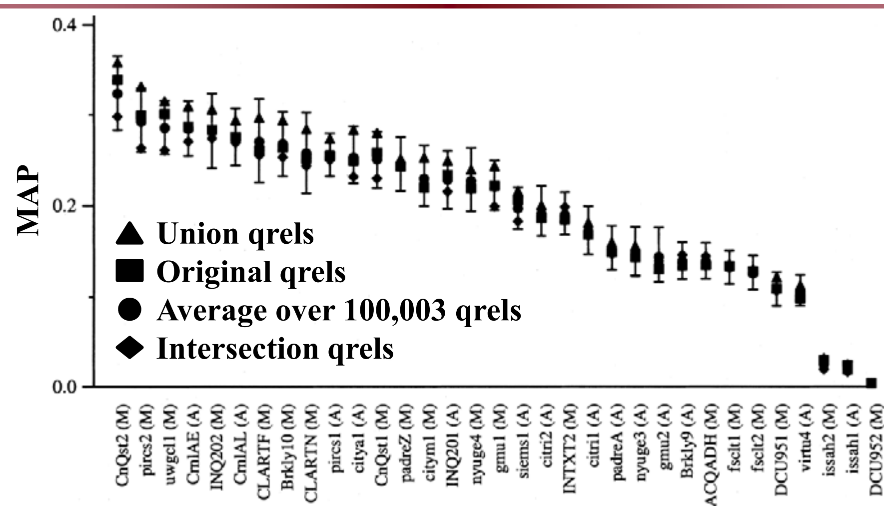
How often do the different sets of relevance assessments disagree about which system is best?

- Evaluate 33 systems using 100,005 sets of relevance judgments
  - A_1, A_2, A_3, Union (A_1, A_2, A_3), Intersection (A_1, A_2, A_3)
  - 100,000 randomly generated combinations
  - Use mean average precision (MAP) as the metric
- Rank the systems
- Compare the rankings produced by each set of assessments

21

(Voorhees, 1998)

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



22

(Voorhees, 1998)

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

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

System rankings are very 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

23

(Voorhees, 1998)

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

24

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

Characteristics of the Cranfield Methodology

Each user is as an expert (on their information need)

- This creates some implicit requirements or assumptions
  – Users are well-trained, don't get bored, don't make mistakes
- In reality, results for any individual query are unreliable
- The assessments are accurate enough to rank search engines

The test collection is static

- Relevance judgments collected today will be useful tomorrow
- Relevance of one document is independent of other documents

The test collection is reusable

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

The Cranfield methodology is difficult to apply to web search

- Tens of millions of queries per day
- Information needs for most queries are unknown
- Clicks are not relevance judgments
- The test collection is dynamic

Web search engines do use the Cranfield methodology

- They have use trained assessors, similar to NIST

But, they also use other metrics and methodologies…

Evaluation in a Dynamic Environment: Interleaved Testing

Interleaved testing is a common experimental methodology

- **Input:** Two rankings produced by different methods
- **Output:** One ranking composed of documents from each method
  - A fair ranking that does not favor either method
- **Gather data:** Which method produces better results?
  - Which method provides documents that get more clicks?

Ranking_1

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

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

Ranking_2

Interleaving Method

$\text{Ranking}_{\text{Interleaved}}$

$d_1\ d_2\ d_6\ d_5\ d_4\ d_9 \ldots$

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: 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

31

Evaluation in a Dynamic Environment: Balanced Interleaving

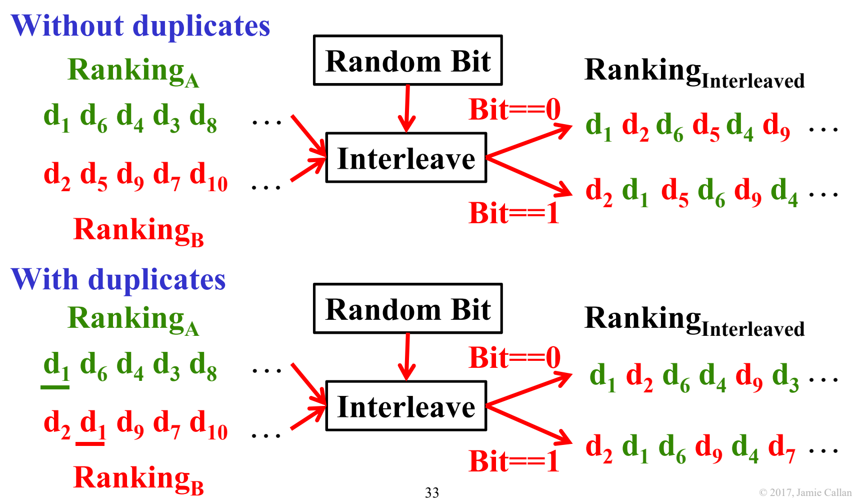
Input: Rankings $A = (a_1, a_2, \ldots)$ and $B = (b_1, b_2, \ldots)$
$I \leftarrow (); k_a \leftarrow 1; k_b \leftarrow 1;$
$AFirst \leftarrow RandomBit()$ .... *decide which ranking gets priority*
while $(k_a \leq |A|) \wedge (k_b \leq |B|)$ **do** .... *if not at end of A or B*
  if $(k_a < k_b) \vee ((k_a = k_b) \wedge (AFirst = 1))$ **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)

32

Evaluation in a Dynamic Environment: Balanced Interleaving



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

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)

Evaluation in a Dynamic Environment: Balanced Interleaving

Example

- Clicked: ✓
- $c_{max}=3$
- k=2
  - Depth needed in R_1 or R_2 to find all clicked docs
- $\# clicks_{R_1}=1$
- $\# clicks_{R_2}=2$

R_2 wins this trial

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)

Evaluation in a Dynamic Environment: Balanced Interleaving

Example

- Clicked: ✓
- c_{max}=3
- k=2
  - Depth needed in R_1 or R_2 to find all clicked docs
- # $clicks_{R_1}$=1
- # $clicks_{R_2}$=2

R_2 wins this trial

	Input Ranking		Interleaved Ranking
Rank	R_1	R_2	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)

37

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
cat videos	Ye	R_1	R_2
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

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

Evaluation in a Dynamic Environment: Balanced Interleaving

Balanced Interleaving can behave unexpectedly

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

Why?

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

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

(Chapelle et al, 2012)

39

Evaluation in a Dynamic Environment: Team-Draft Interleaving

```
Input: Rankings A = (a1, a2, ...) and B = (b1, b2, ...)
Init: I ← (); TeamA ← ∅; TeamB ← ∅;
while (∃i : A[i] ∉ I) ∧ (∃j : B[j] ∉ I) do ................. if not at end of A or B
  if (|TeamA| < |TeamB|) ∨
     ((|TeamA| = |TeamB|) ∧ (RandBit() = 1)) then
     k ← min_i{i : A[i] ∉ I} ............................ top result in A not yet in I
     I ← I + A[k]; ............................................... append it to I
     TeamA ← TeamA ∪ {A[k]} ............................... clicks credited to A
  else
     k ← min_i{i : B[i] ∉ I} ............................ top result in B not yet in I
     I ← I + B[k] ................................................ append it to I
     TeamB ← TeamB ∪ {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)

40

Evaluation in a Dynamic Environment: Team Draft Interleaving

Without duplicates

Ranking$_A$: $d_1\ d_6\ d_4\ d_3\ d_8\ \ldots$

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

Random Bit → Interleave

Ranking$_{Interleaved}$

Bit==0: $d_1\ d_2\ d_6\ d_5\ d_4\ d_9\ \ldots$

Bit==1: $d_2\ d_1\ d_5\ d_6\ d_9\ d_4\ \ldots$

With duplicates

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

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

Random Bit → Interleave

Ranking$_{Interleaved}$

Bit==0: $d_1\ d_2\ d_6\ d_9\ d_4\ d_7\ \ldots$

Bit==1: $d_2\ d_1\ d_9\ d_6\ d_7\ d_4\ \ldots$

41

Evaluation in a Dynamic Environment: Team-Draft Interleaving

Consider an interleaved ranking I with clicks C

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

Clicks attributed to each method are

$\#\ clicks_a = |c_j : i_{c_j} \in Team_a|$ **clicks on docs from a**

$\#\ clicks_b = |c_j : i_{c_j} \in Team_b|$ **clicks on docs from b**

The method that gets the most clicks wins the trial

Aggregate results for all trials to find the best ranker

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

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)

42

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

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

- But, if users click on only the first relevant document, R2 wins 51% of the trials
  - This is an artifact of how duplicates are handled
  - Only the method that suggested the document higher gets credit

43

(Chapelle et al, 2012)

Evaluation in a Dynamic Environment: Search Engines Tested

ArXiv.org

- 700K academic articles, scientific users, about 70K searches
- Ranking strategies created by degrading a baseline

Bing

- Team-Draft interleaving was performed on a % of US traffic
- Five pairs of proprietary ranking functions, 220K searches
  - 3 functions with Δ MAP and NDCG > 0.5% absolute
  - 2 functions with Δ MAP and NDCG < 0.2% absolute
- 12,000 queries were also manually assessed on a 5-point scale

(Chapelle et al, 2012)

44

Evaluation in a Dynamic Environment: Search Engines Tested

Yahoo

- Balanced interleaving was performed on a % of US traffic
- All pairs of four proprietary ranking functions, about 20M searches
  - The current production method and 3 candidates for next release
  - Two rankers were very similar (variants on a theme)
  - The maximum differences in MAP and NDCG are < 0.65% relative
- 2,000 queries were also manually assessed

(Chapelle et al, 2012)

45

Evaluation in a Dynamic Environment: Data Collected

	Experimental Condition: Type	Experimental Condition: Function(s)	Number of Searches	Number of Days	First Day
Bing	Team-Draft	$\mathcal{B}_B \succ \mathcal{A}_B$	220,000	4	July 21, 2009
Bing	Team-Draft	$\mathcal{C}_B \succ \mathcal{A}_B$	190,000	4	Aug 4, 2009
Bing	Team-Draft	$\mathcal{C}_B \succ \mathcal{B}_B$	220,000	4	Aug 11, 2009
Bing	Team-Draft	$\mathcal{D}_B \succ \mathcal{C}_B$	220,000	4	July 7, 2009
Bing	Team-Draft	$\mathcal{F}_B \succ \mathcal{E}_B$	220,000	4	Sept 1, 2009
Yahoo!	Non-Comp	$\mathcal{A}_Y$	73.9 M	33	Mar 17, 2010
Yahoo!	Non-Comp	$\mathcal{B}_Y$	10.4 M	33	Mar 17, 2010
Yahoo!	Non-Comp	$\mathcal{C}_Y$	41.8 M	33	Mar 17, 2010
Yahoo!	Non-Comp	$\mathcal{D}_Y$	72.4 M	33	Mar 17, 2010
Yahoo!	Balanced	$\mathcal{D}_Y \succ \mathcal{C}_Y$	13.9 M	42	May 12, 2010
Yahoo!	Balanced	$\mathcal{D}_Y \succ \mathcal{B}_Y$	1.5 M	5	Apr 14, 2010
Yahoo!	Balanced	$\mathcal{D}_Y \succ \mathcal{A}_Y$	677,000	2	Apr 7, 2010
Yahoo!	Balanced	$\mathcal{C}_Y \succ \mathcal{B}_Y$	1.5 M	5	Apr 14, 2010
Yahoo!	Balanced	$\mathcal{C}_Y \succ \mathcal{A}_Y$	680,000	2	Apr 7, 2010
Yahoo!	Balanced	$\mathcal{B}_Y \succ \mathcal{A}_Y$	1.6 M	5	Apr 9, 2010

(Chapelle et al, 2012)

46

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

ArXiv.org

- Varying amounts of manual degradation of current ranker
- 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)

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)

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
- **Clicks per query,** **Clicks@1**
- **pSAT-clicks:** % of documents with dwell time > 30 seconds
- **pSkip:** % of documents that are skipped
- **Max Reciprocal Rank,** **Mean Reciprocal Rank**
- **Time to First Click,** **Time to Last Click**

(Chapelle et al, 2012)

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

Interleaving does not predict changes 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:**
  - A few hundred thousand queries for rankers of different quality
  - A few million queries for rankers of similar quality

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

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	Optional	N
Dynamic test collections	N	Y
Quickly test new methods	Optional	Y

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.