Comparative Analysis of Clicks and Judgments for IR Evaluation

Jaap Kamps[1,3] Marijn Koolen[1] Andrew Trotman[2,3]

[1]University of Amsterdam, The Netherlands

[2]University of Otago, New Zealand

[3]INitiative for the Evaluation of XML Retrieval (INEX)

WSCD: Workshop on Web Search and Click Data
Barcelona, February 9, 2009

Overview

- Introduction
- Three sets of data: IR test collection and two log files
- Look at differences between clicks and relevance judgments
- Look differences between system rankings based on clicks and relevance judgments
- Discussion and conclusions

IR Evaluation

- Until recently: IR evaluation = Cranfield style test collection
- Recent alternative: queries and click data from search logs due to volume and relation to end-user querying
- The overall aim of this paper is to answer the question:
  - ⋆ How does click-through data differ from explicit human relevance judgments in information retrieval evaluation?

Idea of the Paper

- In a nut-shell:
  - ⋆ compare a traditional test collection with manual judgments
  - ⋆ to transaction log based test collections
- Q1: are there differences between clicks and relevance judgments?
  - ⋆ Earlier studies show reasonable agreement, but clicks are different from static absolute relevance judgments
- Q2: are there differences between system ranking based on clicks and based on relevance judgments?
  - ⋆ Open question, but system rankings are known to be remarkably robust

Three Sets of Data

- Decreasing completeness
  - ⋆ IR test collection: human judged topics with a “complete” set of relevant documents (relative to the pooled documents)
  - ⋆ Proxy log contains complete user sessions, showing all viewed pages after an initial query
  - ⋆ Search engine log contains only part of such a whole session, containing a query and one of more clicked results
- We’ll build three “test collections”
  - ⋆ using log queries as topics and subsequent clicks as pseudo-relevance judgments for the clicked results

(1) INEX 2008 Ad Hoc Track Test Collection

- A traditional test collection following *Cranfield*:
  - ⋆ Documents a snapshot of the English Wikipedia in early 2006, turned into XML mark-up
  - ⋆ Topics 135 ad hoc topics created by INEX participants
  - ⋆ Judgments explicit human judgments for 70 of those topics (pools of 600 articles)
- INEX judges highlight the exact relevant text
  - ⋆ Here we derive article-level qrels

(2) New Zealand Proxy Log

- A proxy log from a New Zealand high school covering three months of traffic.
  - ⋆ Complete user sessions, including browsing further pages
  - ⋆ Even with the user-ids!

- Extracted queries targeting Wikipedia, and the associated clicks
  - ⋆ 138 topics were added to the INEX 2008 topics set
  - ⋆ Selected on two criteria:
    1) the query leads to a click on a Wikipedia article, and
    2) the query was typed by more than one user.

(3) MSN Log

- Queries and clicks from a major Internet search engine.
  - ⋆ Captures only initial part of such a whole user session.
  - ⋆ Contains over 40,000 queries targeting Wikipedia
  - ⋆ Including 50 of the INEX topics (ad hoc or proxy log)
- MSN and proxy log clicks are mapped to INEX document ids
  - ⋆ MSN log roughly from the same period as the INEX collection
  - ⋆ Proxy log more recent

Wikipedia Clicks in the Logs

Description	MSN	Proxy
Total queries	8,831,281	36,138
Distinct queries	3,545,503	12,318
Total clicks	12,251,068	–
Distinct clicks	4,975,898	–
Clicks in Wikipedia	63,506	7,186
Total queries with Wiki clicks	59,538	3,211
Distinct queries with Wiki clicks	41,428	2,224

- Fair fraction of queries is targeting Wikipedia
  - ⋆ 1.2% of MSN queries, and 8.9% of the Proxy log queries
  - ⋆ MSN is huge, but we'll only use the 50 queries corresponding to the INEX topics
- On the set of INEX topics: How do these differ from judgments?

Distribution of Relevant Docs

	total #		per topic				
Topic set	topics	pages	min	max	median	mean	st.dev
Manual	70	4,850	2	375	49	69.31	68.73
Proxy	138	330	1	13	2	2.39	2.17
MSN	50	58	1	2	1	1.16	0.37

- Differences in # of relevant/clicked documents
  - Ad hoc topics have 70 relevant docs (max 375)
  - Proxy log has 2 (max 13)
  - MSN log has 1 (max 2)
- So there are striking differences in “completeness”

Impact on System Ranking?

- We have seen that there are considerable differences
  - ⋆ But how does this impact comparative IR evaluation?
  - ⋆ What is the impact on the ranking of systems?
- This is the main goal of our experiment:
  - ⋆ We have 3 sets of qrels (Ad hoc, Proxy, MSN)
  - ⋆ but also 163 INEX submissions for these topics!
- Will the rankings of these runs agree?

System Ranking (Top 10)

Ad hoc	map	Proxy log	map	MSN log	map
1	0.3753	45	0.4625	42	0.6999
2	0.3686	39	0.4601	41	0.6982
3	0.3601	40	0.4601	43	0.6977
4	0.3489	41	0.4471	30	0.6963
5	0.3412	42	0.4467	25	0.6963
6	0.3390	43	0.4464	75	0.6904
7	0.3383	6	0.4368	39	0.6866
8	0.3371	7	0.4368	40	0.6866
9	0.3344	9	0.4368	36	0.6848
10	0.3333	26	0.4368	31	0.6848

- Run label is Ad hoc rank
  - ⋆ Ad hoc and Proxy have 3 runs in common
  - ⋆ Ad hoc and MSN have no runs in common
  - ⋆ Proxy and MSN have 5 runs in common

System Rank Correlation (163 runs)

	map			1/rank		
Collection	Ad hoc	Proxy	MSN	Ad hoc	Proxy	MSN
Ad hoc	1.000	0.360	0.296	1.000	0.442	0.379
Proxy		1.000	0.784		1.000	0.788
MSN			1.000			1.000

- Overall there is “some” agreement
  - ⋆ Ad hoc agrees 30% (MSN) to 36% (Proxy)
  - ⋆ Reciprocal rank somewhat better
- The rankings differ, but which one is “better”?

Significant Differences

Ad hoc	1	2	3	4	5	6	7	8	9	10
1		-	-	-	⋆	⋆	⋆	⋆	⋆	⋆
2			-	-	-	-	-	⋆	⋆	⋆
3				-	-	-	-	-	-	-
4					⋆	-	-	-	-	-
5						-	-	-	-	-
6							⋆	-	⋆	-
7								-	⋆	-
8									-	-
9										-
10										

Proxy	1	2	3	4	5	6	7	8	9	10
45		-	-	-	-	-	-	-	-	-
39			-	-	-	⋆	-	-	-	-
40				-	-	⋆	-	-	-	-
41					-	-	-	-	-	-
42						-	-	-	-	-
43							-	-	-	-
6								⋆	⋆	⋆
7									⋆	⋆
9										⋆
26										

MSN	1	2	3	4	5	6	7	8	9	10
42		-	-	-	-	⋆	-	-	-	-
41			-	-	-	⋆	-	-	-	-
43				-	-	⋆	-	-	-	-
30					-	⋆	-	-	-	-
25						⋆	-	-	-	-
75							-	-	-	-
39								-	-	-
40									-	-
36										-
31										

- There is some support for the ad hoc ranking
  - ⋆ Proxy log: high-ranked ad hoc runs (6, 7, 9) really better
  - ⋆ MSN log: low-ranked ad hoc run (75) really worse

What's the Bias?

- Clicks are less "complete" than human judgments
  - ⋆ Ad hoc 70 per topic, versus 1-2 clicks per query
- An unbiased sample would result in comparable system-rankings
  - ⋆ We see clear upsets
  - ⋆ What's causing the bias?
- We ignore user-biases, and look at the relation between query and clicked/relevant document

Title Bias

	Test collections			Complete log	
	Ad Hoc	Proxy	MSN	Proxy	MSN
titlestat_rel	0.061	0.508	0.953	0.524	0.689

- Wikipedia title (in URL) prevails in log clicks
  - ⋆ Only 6% of ad hoc's relevant pages
  - ⋆ 51% of the proxy's clicked pages
  - ⋆ 96% of the MSN's clicked pages
- There is striking title bias
  - ⋆ Casts doubt on measuring recall aspects

Discussion and Conclusions

- Traditional IR evaluation is based on IR test collections
  - ⋆ Industry moves to "operational" testing using queries and clicks
  - ⋆ Attractive: costs, quantity, and relation to end-user querying

- Logs are less "complete"
  - ⋆ Search engine log 1-2 clicked Wikipedia pages
  - ⋆ Proxy log slightly more, but still a fraction of explicit judgments
  - ⋆ There is a strong title bias
  - ⋆ Difficult to measure any recall effect

- Use with care: log data are no silver bullet
  - ⋆ Incredibly rich, but potentially biased and shallow
  - ⋆ Still, I'd love to use them if they were available for research!

Thank You

- Questions?