Comparative Analysis of Clicks and Judgments for IR Evaluation

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

[1] Archives and Information Studies, University of Amsterdam, Amsterdam, The Netherlands
[2] ISLA, University of Amsterdam, Amsterdam, The Netherlands
[3] Department of Computer Science, University of Otago, Dunedin, New Zealand

ABSTRACT

Queries and click-through data taken from search engine transaction logs is an attractive alternative to traditional test collections, due to its volume and the direct 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? We compare a traditional test collection with manual judgments to transaction log based test collections—by using queries as topics and subsequent clicks as pseudo-relevance judgments for the clicked results.

Specifically, we investigate the following two research questions: Firstly, are there significant differences between clicks and relevance judgments. Earlier research suggests that although clicks and explicit judgments show reasonable agreement, clicks are different from static absolute relevance judgments. Secondly, are there significant differences between system ranking based on clicks and based on relevance judgments? This is an open question, but earlier research suggests that comparative evaluation in terms of system ranking is remarkably robust.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback, retrieval models*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms

Measurement, Performance, Experimentation

Keywords

Web information retrieval, Transaction log analysis, Wikipedia

WSCD'09, Feb 9, 2009 Barcelona, Spain.

1. INTRODUCTION

Until recently, the evaluation of retrieval effectiveness was synonymous with the use of information retrieval test collections—typically consisting of a frozen set of documents, search requests, and relevance judgments—such as constructed at the Text REtrieval Conference [TREC, 21]. Nowadays, there is an increasing interest in using queries and click-through data mined from search engine transaction logs, and web search click data is used as ground truth for evaluation [e.g., 13, 18]. This raises the following question: How does query log data differ from explicit human relevance judgments in information retrieval evaluation?

In this paper, we will investigate the differences between traditional information retrieval test collections—consisting of a frozen set of documents, search requests, and relevance judgments—and queries and click-through data mined from transaction logs. Specifically we look at the three different sources: humanly judged INEX 2008 ad hoc topics, a MSN search engine log, and a proxy log from a New Zealand high school. The INEX collection is based on a dump of the Wikipedia, and we will restrict our attention to Wikipedia targeting queries from the log files. We use a simple method of generating a document retrieval test collection from both a search engine's transaction log and a proxy log, based on submitted queries and user click-through data, and conduct a comparative analysis. Our analysis especially seeks to understand the differences between clicks and explicit judgments, and how these differences impact the comparative evaluation of retrieval effectiveness.

Our first research question is: are there significant differences between clicks and relevance judgments? Earlier research by Joachims et al. [14] suggests that although clicks and explicit judgments show reasonable agreement, clicks do not coincide with absolute relevance judgments but can be interpreted as relative relevance judgments. We investigate this question by comparing a search engine log, a proxy log, and a set of human judged ad hoc retrieval topics. The human judged topics will be a "complete" set of relevant documents (relative to the pooled documents). The proxy log will contain a complete user session, showing all viewed pages after an initial query. Finally, the search engine log contains only part of such a whole session, containing a query and one of more clicked results.

Our second research question is: are there significant differences between system rankings based on clicks and based on relevance judgments? This is an open question, but earlier research suggests that comparative evaluation in terms of system ranking is remarkably robust. As explained by

Voorhees [22], test collections are used "as a mechanism for comparing system performance." Thus, even though relevance judgments in IR test collections are typically not complete (i.e., not all documents are judged for each topic), the pooling method allows an unbiased comparison of retrieval effectiveness between systems that contributed to the pool. Thus, two test collections based on the same document collection and on the same type of topics, leading to the same system ranking can be considered equivalent in evaluating the retrieval performance of IR systems.

Hence, even if clicks and explicit judgments differ to some extent, they may still generate very comparable system rankings. This is very similar to the results of Zobel [23], who showed that despite limitations of the pooling methodology the resulting system ranking is reliable. In an experiment with a domain-specific test collection derived from a museum's transaction log, Arampatzis et al. [1] did show indeed similar system-rankings between the log-based evaluation and a manually constructed set of known-item search topics.

The rest of the paper is structured as follows. Section 2 discusses related work. Then, in Section 3 we describe the click-through data sets we will use and how we derive topics and relevance judgments from them. In Section 4 we give various statistics of the resulting queries plus clicked or judged documents. Section 5 compares the systems rankings produced by the test collections based on clicks with test collections based on explicit human judgments. In Section 6 we further analyse aspects specific to click-through data from both search engine logs and proxy logs and discuss how these clicks differ from each other. Finally, we end in Section 7 by summarizing our findings, and discussing their impact.

2. RELATED WORK

We draw upon research in three areas—query or user intent, implicit feedback and automatic evaluation – which we will discuss in turn.

The best-known work on query or user intent is the Broder [4] taxonomy. That is, there are *navigational queries* (with the intent to reach a particular site), *informational queries* (with the intent to acquire some information present in some web pages), and *transactional queries* (with the intent to perform some web-mediated activity). Broder found that almost 50% were informational queries. Rose and Levinson [19] further refine the taxonomy by subdividing informational and transactional queries and found a somewhat higher fraction, above 60%, of informational queries. Automatic intent classification is studied by Jansen et al. [12], who found even higher fractions, above 80%, of informational queries. We will focus informational queries, and compare them to ad hoc search requests. Since we restrict our attention to Wikipedia results, there may be a navigational import in the informational query intent—a user may purposefully select the Wikipedia based on its credibility or reputation.

There has been substantial interest in using click-through data from transaction logs as a form of implicit feedback [9]. A range of implicit feedback techniques have been used for query expansion and user profiling in information retrieval tasks [16, 17]. Joachims et al. [14, p.160] conclude that "the implicit feedback generated from clicks shows reasonable agreement with the explicit judgments of the pages." Transaction logs, or more specifically, search logs, have been analysed [11] to study user search behaviour in Web search engines [6] and digital libraries [15], amongst others. In [6], user behaviour is studied using the transaction log of a website's search engine and is compared to that of general purpose search engines. They find that the number of query terms used for website search engines is comparable to queries submitted to general purpose search engines, but the search topics and terms are different. This signals that the subset of queries and clicks targeting Wikipedia may be different from the overall Web information seeking behavior.

Beitzel et al. [3] and Chowdhury and Soboroff [7] have used transaction logs to extract queries for automatic evaluation of on-line search services where queries were paired with pseudo-relevant documents, i.e., pages (from the Open Directory Project taxonomy) with titles of pages that exactly match the query. They use these topics to evaluate a number of web search engines that have indexed these pages (which may change or disappear over time). They find that these automatically generated topics lead to a system ranking that strongly correlates with a set of manual topics. Our approach, instead, leads to a test collection that, although requiring the search engine from which the logs are used to have a static document collection, is reusable by different web and non-web search engines. Moreover, as the pseudo-relevant documents are also based on user data, it combines implicit feedback analysis and automatic evaluation to create topics that are directly related to user search behaviour. Other alternatives of automatic evaluation include generating known-item queries from the document collection [2] and random picks from the pool of retrieved documents [20].

3. EXPERIMENTAL DATA

We want to analyse the "test collection" that would result from a transaction log, when (naively) using the queries issued as topics, and the clicked pages as relevance judgments. In this section we describe the data sets we use for our experiments. Specifically we describe how we mapped the log data to the INEX Ad hoc topics and Wikipedia collection.

3.1 Outline

We will compare three set of data with decreasing "completeness" of (pseudo) judgments:

INEX 2008 Ad hoc track Topics and manual relevance judgments on a collection of Wikipedia pages. The human judged topics will have a "complete" set of relevant documents (relative to the pooled documents).

Proxy log A proxy log from a New Zealand high school covering three months of traffic. The proxy log will contain a complete user session with all viewed pages after an initial query, including those resulting from browsing further pages.

MSN search log Queries and clicks from a major Internet search engine. The search engine log contains only part of such a whole session, containing a query and one of more clicked results.

In order to compare the log data to the INEX topics, we extract from the logs queries targeting Wikipedia, and the associated clicked Wikipedia pages.

Moreover, we have a large set of INEX 2008 Ad hoc track submissions—from participants who made runs with their

systems on the INEX 2008 Ad hoc topic set—that can be used to study the effect on system ranking. For this purpose, we can look at INEX topics occurring in the log files, and use only those queries and clicks to make a corresponding test collection. In fact, 150 queries from the proxy log were included in the INEX 2008 Ad hoc topic set.

3.2 Data

In our experiments we will compare the clicks and queries from both a search engine log and a proxy log to the human relevance judgments from the INEX 2008 Ad hoc test collection [10]. The INEX ad hoc track document collection is based on a snapshot of the English Wikipedia in early 2006 [8], and contains roughly 650,000 articles marked up in XML. We map all click-through data targeting Wikipedia pages to INEX's Wikipedia XML Corpus.

The search engine log we use is the Microsoft 2006 RPF data set, which contains 8.9 million queries and 12 million clicks taken from log data of the MSN search engine, from a period (May, 2006) close to the data of the snapshot.[1] As a consequence, we expect a good correspondence between the Wikipedia articles clicked by users and the articles in the INEX Wikipedia collection. The proxy log contains data from a New Zealand high school, covering a period from 29 august, 2007 to 29 November, 2007, and contains requested URLs, user IDs and timestamps of the Internet traffic in that period from 1,369 distinct users.

The test collections we derive from the log data will be compared to a set of topics using actual human judgments from the INEX 2008 Ad hoc track. This 2008 Ad hoc test collection comes with two sets of topics: one set of 135 topics created by INEX participants, and explicit human judgments for 70 of those topics, and one set of 150 topics from a proxy log. In fact, those 150 topics come from the same proxy log as we use in our experiments in this paper, and were 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.

We have a large set of official runs submitted to the INEX 2008 Ad hoc track, which include results for both the manually judged and the proxy log topics, allowing for detailed analysis of the system-rankings based on both these sets of topics. Using queries from the MSN and proxy logs that match topics in the INEX 2008 Ad hoc topic sets, we can use the corresponding clicks targeting Wikipedia to evaluate the runs submitted to the 2008 Ad hoc track.

3.3 Mapping

We have to derive queries plus corresponding sets of clicked pages from the log files. For the queries and clicks from the MSN log, this is reasonably straightforward. Each click in the data set corresponds to a specific query, identified by a query ID. We can then use each query as a topic and all corresponding clicks as relevance judgments, or group all identical queries and their corresponding clicks.

For the proxy log data, however, there is no direct correspondence between queries issued to search engines, and click-through data requesting pages related to that query. Since the data is obtained from the user side, we cannot see whether or not the URL requested after a query has been issued, is one of the results on the results list generated by the query. It could be that the user was working with multiple browser windows, or requested a page listed in the bookmarks.

Since we want to map the click data to the INEX Wikipedia collection, we only need to look at clicks within the English Wikipedia domain, and we have decided to use a naive approach of associating clicks with queries. We first split the proxy log data over user IDs and look for queries issued to search engines (the search engines we found by skimming the log data are Google, Live Search, MSN, Yahoo! Search and Wikipedia's own search facility), and treat any data with a time-stamp after query n and before query $n+1$ as related to query n. With this method, the time between two subsequent queries can be anything from a few seconds to multiple days – e.g. a user logs out after issuing query n and logs back in after a week and issues query $n+1$ – so, the interval between clicks associated with query n and the issuing of query n itself can be multiple days as well, in which the user can have switched tasks multiple times.

[1]The RPF data set actually comes with two sets of data. One set of 15M queries with corresponding sessions, and one set of 8.9M queries with corresponding clicks. We only work with the latter set in our experiments.

Table 1: Query and click statistics of the MSN and proxy log files

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

4. CLICKS AND JUDGMENTS

In this section we start answering our first research question: are there significant differences between clicks and relevance judgments? Specifically we will give various statistics of the resulting queries plus clicked or judged documents.

4.1 Queries and Clicks

Table 1 presents statistics on the queries and clicks in the two log files. We will discuss the MSN log first. There are 8.83 million queries (identified by query ID) and 12.25 million clicks (1.39 clicks per query). If we look at the set of queries and the set of clicks, that is, group identical queries and identical clicks, we find 3.55 million distinct queries and 4.98 million distinct URLs.[2] This means an average query frequency 2.49 and and each URL in the set is clicked 2.46 times.

The proxy log contains 36,138 queries in total, and 12,318 distinct queries (average query frequency is 2.93) from 687 users. The proxy log data set contains 2.84 million URL requests. Since a single click can lead to multiple URL requests (for instance, a URL for the requested page and a URL for an image on that page), we cannot report the total number of clicks in the proxy log data.

How different are the queries found in the MSN log and the proxy log? The two query sets show substantial over-

[2]To group identical queries, we ignore case. As URLs are case sensitive, we leave them as is for grouping.

Table 2: Overlap between INEX 2008 topic titles and queries in the MSN and proxy log files. The numbers in parentheses represent the overlap with the two subsets of the 2008 topics (topics 544-678/topics 679–828)

Description	MSN	Proxy
INEX query	121 (4/117)	150 (0/150)
INEX query + Wiki click	50 (4/46)	138 (0/138)

Table 3: Distribution of relevant documents over topics

	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

lap. There is a total of 2,786 distinct queries that occur in both data sets. That is 22.62% of the distinct queries in the proxy log and 0.79% of the distinct queries in the MSN log (due to the different size). However, these 2,786 distinct queries match with 15,585 individual queries in the proxy log (43.13% of the total number of queries) and 790,758 queries in the MSN log (22.30%). This means that the queries occurring in both logs have an average frequency of 5.59 in the proxy log and and an average frequency of 283.83 in the MSN log, which is well above the average frequency of the total query sets, showing that the overlapping queries are the more frequent queries.

4.2 Queries and Wikipedia Clicks

In the MSN log, there are 63,506 clicks on Wikipedia articles, corresponding to 59,538 query IDs and 41,428 distinct queries and in the proxy log, there are 7,186 clicks on Wikipedia articles, corresponding to 3,211 queries (2,224 distinct queries) from 410 users.

We already know that there are at least 150 queries in the proxy log matching topics in the INEX 2008 Ad hoc test collection. In the next section we look in more details at the overlap between the log queries targeting Wikipedia, and the INEX Ad hoc topics.

We now look at the overlap between the queries and clicks in the MSN and proxy log data on one side and the INEX Wikipedia collection and the INEX 2008 Ad Hoc topic set on the other side.

Of the 63,506 MSN log clicks on Wikipedia articles, there are 50,361 matches (79%) with articles in the INEX Wikipedia collection. This high overlap is in line with our earlier remark that the log data is from a period close to the moment of the snapshot on which the INEX Wikipedia collection is based. For the proxy log, 3,746 of the 7,186 (52%) clicks correspond to articles in the INEX Wikipedia collection.

4.3 INEX 2008 Ad hoc Topics

Next, we map the queries from the MSN and proxy logs to the 2008 Ad hoc topics and the clicks on Wikipedia pages to the Wikipedia articles in the INEX 2008 Wikipedia collection. Table 2 shows the overlap between the INEX 2008 Ad Hoc topic set and the queries in the MSN and proxy log data. The first line shows the overlap between log queries and topic titles. If a query has corresponding clicks on Wikipedia pages that do not appear in the INEX Wikipedia snapshot, we cannot compare that query to the INEX 2008 Ad Hoc topic set. Therefore, the second line in Table 2 shows the overlap between log queries and topic title, where the log queries have at least one click corresponding to a document in the INEX Wikipedia collection. Of the 285 INEX 2008 topics (135 manual and 150 log topics), there are 121 topic titles (4 in the manual topics, 117 in the log topics) that appear as queries in the MSN log. Of these 121 queries, there are 50 queries with at least one click on a Wikipedia article that also appears in the INEX Wikipedia collection. Since the proxy log is the source of the INEX 2008 log topics, all 150 INEX 2008 log topics appear in that proxy log. However, of those 150 queries, there are 138 with a click on a Wikipedia page that is also available in the INEX Wikipedia collection. The MSN log queries show much more overlap with the INEX 2008 log topics than with the manually created topics. This is to be expected, as both sets of queries were issued as Web search queries, whereas the manually created INEX topics are explicitly created – and possibly revised by the topic creator after exploring the Wikipedia collection – as Ad Hoc topics.

Table 3 shows the distribution of relevant documents over the different topic sets. The most immediately visible difference between the human judgments and the clicks is the number of relevant documents per topic. The 70 manually assessed INEX Ad hoc topics, contain explicit relevance judgments for pools of at least 500 documents, with an average of 69 relevant documents per topic. In contrast, the 50 MSN log topics have an average of 1.16 clicked documents (which we regard as relevant, since we treat clicks as positive relevance judgments), and the proxy log topics have 2.39 relevant documents per topic. Note that although the MSN log is based on one system, there is a multitude of users and we see clicked pages more than once in the log, on average 7.24 times for the same query. To a lesser extent, the same holds for the proxy log where we see an average of 1.94 for each query/clicked document pair. The latter one being in part a result of the selection of queries issued by at least 2 different users.

Although the average number of relevant documents per topic for the proxy log topics is not much higher than for the MSN log topics, the distribution is very different. The MSN log topic set has 8 topics with two relevant documents, all other topics have one relevant document. The proxy log topic set has four topics with 10 or more relevant documents and 14 topics with 5 or more relevant documents.

In this section we compared manually created and assessed Ad hoc topics with topics derived from the log files. The most striking difference is the number of relevant documents per topic. The topics derived from the MSN log mostly have one relevant document. Some of the proxy log topics have quite a few more relevant documents, with a maximum of 13 relevant documents per topic. However, in comparison with the ad hoc topics, with on average 69 relevant documents, it is clear that the proxy log clicks are far less exhaustive than the pools used in the INEX Ad hoc Track. This can have important consequences for evaluation, which we will explore in the next section.

Table 4: Top 10 runs: Ad hoc judgments (left), Proxy log clicks (middle), MSN log clicks (right)

Run	P5	P10	1/rank	map	Run	P5	P10	1/rank	map	Run	P5	P10	1/rank	map
1	0.6200	0.5257	0.8711	0.3753	45	0.1594	0.0877	0.5904	0.4625	42	0.2000	0.1000	0.7133	0.6999
2	0.6257	0.5300	0.8509	0.3686	39	0.1623	0.0870	0.5776	0.4601	41	0.2000	0.1000	0.7133	0.6982
3	0.6371	0.5843	0.8322	0.3601	40	0.1623	0.0870	0.5776	0.4601	43	0.1960	0.1000	0.7128	0.6977
4	0.5914	0.5386	0.8635	0.3489	41	0.1594	0.0855	0.5674	0.4471	30	0.1840	0.1000	0.7126	0.6963
5	0.6000	0.5371	0.8724	0.3412	42	0.1594	0.0862	0.5673	0.4467	25	0.1840	0.1000	0.7126	0.6963
6	0.5686	0.5214	0.7868	0.3390	43	0.1580	0.0855	0.5673	0.4464	75	0.1960	0.1020	0.7189	0.6904
7	0.5686	0.5214	0.7868	0.3383	6	0.1507	0.0833	0.5656	0.4368	39	0.2000	0.1000	0.7018	0.6866
8	0.5800	0.4943	0.8161	0.3371	7	0.1507	0.0833	0.5656	0.4368	40	0.2000	0.1000	0.7018	0.6866
9	0.5686	0.5214	0.7868	0.3344	9	0.1507	0.0833	0.5656	0.4368	36	0.1760	0.1000	0.7025	0.6848
10	0.5543	0.5100	0.7894	0.3333	26	0.1507	0.0833	0.5656	0.4368	31	0.1760	0.1000	0.7025	0.6848

5. COMPARATIVE EVALUATION

In this section we will start to address our second research question: are there significant differences between system rankings based on clicks and system rankings based on relevance judgments? We do this by comparing the system ranking of the official runs submitted to the INEX 2008 Ad hoc track, 163 different runs in total, over the three set of relevant/clicked pages: 70 manually judged ad hoc topics, 50 queries and clicks derived from the MSN log, and 138 queries and clicks derived from a proxy log.

5.1 System Ranking

Having derived test collections from the MSN and proxy logs, we now investigate their ability to rank systems. Recall that the value of test collections is that they allow us to compare the *relative* effectiveness of retrieval systems. If we use two test collections to evaluate the same group of systems, and both test collections rank those systems in the same order, we say that their ability to rank systems is equal.

The runs submitted to the INEX 2008 Ad hoc Track contain results for both the 135 manually created topics and the 150 proxy log topics. With the human judgments for 70 of those 135 topics, and the two sets of relevance judgments derived from the MSN and proxy log, we can evaluate all the runs with these three test collections and compare the system rankings. The top 10 runs for each of the three test collections is shown in Table 4. We label all runs with their system rank based on the manually judged Ad hoc topics (using map), hence on the left-hand side we see run labels 1–10. We see a considerably different ranking for the proxy log (middle of Table 4): only three runs of the best ad hoc systems occur in the top 10, and the rest comes from deep down the ad hoc ranking. The best scoring run is ranked 45 on the ad hoc topics. Looking at the MSN log (right-hand side of Table 4), we see an even more different ranking: the best ad hoc topics system rank is 25, and a run ranked as low as rank 75 is in the top 10. The ranking is far more similar to the proxy log ranking having 5 of the 10 runs in common.

5.2 System Rank Correlations

Table 5(a) shows the system rank correlations (using Kendall's tau) over all 163 runs between the three topic sets. Using map, the ad hoc topics and the proxy log agree on 36% of the pairwise comparisons of systems, and the ad hoc and MSN log agree on 30% of the pairwise comparisons. Given our observations above this is still reasonable, and suggests that that clicked articles are at least a weak indication of relevance. We also show the mean reciprocal rank—the measure of choice for known-item search—which leads to some but limited increase in the correlations of 44% between ad hoc and proxy log, and 38% between ad hoc and MSN log.

Table 5: System rank correlation coefficients for the three test collections

(a) All 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

(b) Top 10 runs

	map			1/rank		
Collection	Ad hoc	Proxy	MSN	Ad hoc	Proxy	MSN
Ad hoc	1.000	-0.244	-0.200	1.000	0.600	0.333
Proxy	-0.289	1.000	-0.644	-0.022	1.000	-0.644
MSN	0.378	0.378	1.000	-0.022	-0.022	1.000

How well do the different test collections agree on the best runs? In Table 5(b) we show the system rank correlation over the top 10 runs. Here we look at the top 10 system according to the set in the row against their ranking in terms of the set in the column, which is not symmetric. For map, we see that the top 10 runs according to the ad hoc topic set is correlating negatively with the log sets, and the top 10 runs of the proxy log set is correlating negatively with the other two. Interestingly, the top 10 according to the MSN log correlate 38% with the other two—there is reasonable agreement on the ranking of the top 10 runs of the MSN log. When looking at reciprocal rank, we see that the top 10 runs based on the ad hoc test collection lead to reasonable agreement: 60% with the proxy topic set, and 33% with the MSN topic set.

5.3 Significant Differences

Although we have established above that the rankings are considerably different with respect to the best systems, it is not immediately clear which of the rankings is better. One criterion would be if the topic set agrees with itself, that is, whether the system rank is not too dependent on the particular choice of topics in the set. We tested whether higher ranked systems were significantly better than lower ranked system, using a t-test (one-tailed) at 95%. Table 6 shows the results, with significant differences indicated with $\star$.

Table 6: Statistical significance (t-test, one-tailed, 95%): Ad hoc judgments (left), Proxy log clicks (middle), MSN log clicks (right)

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

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

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

For the ad hoc topics, we see that the highest scoring system is significantly better than the systems ranked 5 to 10, etc. We see that 13 of the 45 system comparisons are significant, also a result of several close variants of the same run being in the top 10. For the proxy log topics, we see 8 significant differences. However, the top ranking runs tend not to be significantly better than the rest, but the three runs also occurring in the top 10 of the ad hoc top set (labeled 6, 7, 9) are significantly better than lower ranked systems. This can be interpreted as a sign that ad hoc ranking (or where the ad hoc ranking agrees with the proxy log ranking) reflects the inherently better systems. For the MSN log topics, we see 5 significant differences. All these are a comparison with the sixth ranked run, which was ranked 75 over the ad hoc topics. Again, this can be interpreted as a sign that the ad hoc ranking reflect the inherent system quality better.

In this section, we saw that the impact on the comparative evaluation of systems is considerable. There is reasonable agreement over all 163 runs between the three topic sets. Using map, the ad hoc topics and the proxy log agree on 36% of the pairwise comparisons of systems, and the ad hoc and MSN log agree on 30% of the pairwise comparisons. However, the system-ranking for the best 10 runs per set differ radically. There is some evidence that the ad hoc ranking corresponds better to inherent system quality: systems ranked high/low on the ad hoc set tend to be also significantly better/worse on the other log-based sets.

6. FURTHER ANALYSIS

In this section we further analyse the queries and clicks derived from the proxy and search engine log, trying to understand how these may be biased toward particular documents.

6.1 Completeness

In the previous section we saw that the system ranking based on log data differs considerably from the system ranking based on a traditional test collection. The most striking difference between the sets is, as detailed in Section 4, the number of relevant or clicked pages per query. The human judged topics have "complete" sets of relevant documents (relative to the pooled documents). The proxy log contains complete user sessions, showing all viewed pages after an initial query. Finally, the search engine log contains only part of such a whole session, containing a query and one of more clicked results. But in what sense is the log data "incomplete"? If the logs would contain an unbiased sample of the complete set of relevant pages, we could expect a relatively similar system ranking. Since the system-rankings are relatively different, especially for the top ranking systems, we may expect a bias in the set of clicked pages. We will ignore user biases, such as detailed in Joachims et al. [14], for now and focus on features of the resulting queries and sets of clicked pages. In particular, we look at the nature of the data captured in the log files.

The search engine log contains only clicks on the result list, but no subsequent clicks from the requested page to further pages. Although the search engine might retrieve several Wikipedia pages in response to a query, some of them having only a few or none of the query terms in the title, these pages will typically be ranked lower than a page with a title exactly matching the query. Wikipedia being an encyclopedia, a page with a title exactly matching the query is the natural entry point, and search engines will thus rank that page higher than other Wikipedia pages. Added to that, a search engine will typically show results from different sites, and hence suppress further results from Wikipedia. So, even if further Wikipedia results occur in the ranking, they will be ranked low after results from various other Web sources and will receive no clicks.

Note that this is no problem for the proxy log, where we capture the complete session of the user from the query issued to the search engine, to any clicks following that query, whether they are clicks on the results list or not.

6.2 Title Bias

In the MSN data, we found 59,538 queries leading to 63,506 clicks in Wikipedia, giving an average of 1.07 clicks per query. This suggests a direct correspondence between those queries and clicks. Looking at the list of queries and the titles of the Wikipedia pages the were clicked on as results of these queries, we found that, in most cases, the title of the clicked Wikipedia page exactly matches the query. Using the proxy log, we found 3,211 queries leading to 7,186 clicks on Wikipedia articles, an average of 2.24 clicks per query. In most cases, a query in the proxy log has one Wikipedia article with a title matching the query. The other clicks often correspond to Wikipedia articles that are related to the query, but with very different titles.

This bias in the MSN log towards Wikipedia pages that have most of the query terms in the title is similar to the bias reported by Buckley et al. [5] in the TREC 2005 HARD and robust tracks. They define the measure *titlestat_rel* as the

Table 7: Titlestat_rel over Ad Hoc, MSN log and Proxy log topics

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

fraction of a set of documents that a topic title term occurs in, and show that the document pools that were judged for the topic sets used in those tracks contained a bias towards documents that contain terms from the topic title. We use *titlestat_rel* in a slightly different way; instead of looking for the occurrence of query terms in the whole *content* of each relevant document, we only look at the *title* of each relevant Wikipedia page. That is, we look at the bias in the search results towards Wikipedia pages that have most query terms in the title. We compute *titlestat_rel* as follows:

$$titlestat\_rel_T = \frac{1}{|T|} \sum_{t \in T} \frac{|C_t|}{|C|} \quad (1)$$

Where t is a title term in topic T, C is the set of documents relevant to topic T, and C_t is the set of documents in C that contain t. The titlestat value for the whole topic set is the average of the per-topic titlestat scores. A maximum of 1.0 means all titles of the relevant Wikipedia pages contain all the query terms. The minimum value is 0.0, meaning none of the Wikipedia titles contain any query terms.[3]

Table 7 shows the *titlestat_rel* of the three test collections. The term overlap between the Ad Hoc topic titles and the titles of the relevant Wikipedia articles is very small, only 0.061. With an average of 69 relevant documents per topic in the manually judged INEX 2008 topics, this low number is not surprising. For the Qrels derived from the MSN log, the *titlestat_rel* is 0.953, meaning an almost perfect match between the user queries and the titles of the Wikipedia pages they click on. The titlestat_rel value for the proxy log Qrels, 0.508, is much higher than the value for the manually judged topics, but much lower than the value for the MSN log topics. Table 7 also shows the *titlestat_rel* of the all query and click pairs in the log files. Here we see roughly the same fraction, 0.524, for the Proxy log, and a lower fraction of 0.689 for the MSN log—still considerably higher than the other two sets.

In this section, we further analysed the queries and clicks derived from the transaction logs. The main observation was that number of relevant pages whose title contained the complete query is low for the ad hoc judgments (6%), moderately high for the proxy log clicks (51%), and extreme for the MSN log click (95%). This casts considerable doubt on the ability of the search engine log-based test collections to measure recall-related aspects of retrieval systems.

[3]Another difference with the implementation of Buckley et al. is that we do not normalise the fraction by taking the minimum of $|C|$ and the collection frequency of t. In the original measure, this normalisation is necessary for rare terms. If a topic title term has a collection frequency smaller than $|C|$, then $|C|$ is replaced by df_t. In our case, we only look at the document title, while the systems that contributed to the pools—the participating systems in the INEX Ad Hoc Track and Internet search engines in the log data—had access to all the document content. Therefore, the issue of low frequency terms is less important here.

7. DISCUSSION AND CONCLUSIONS

In this paper, we investigated the differences between traditional information retrieval test collections—typically consisting of a frozen set of documents, search requests, and relevance judgments—and queries and click-through data mined from transaction logs. Our main research question was: How does query log data differ from explicit human relevance judgments in information retrieval evaluation? Specifically we looked at the three different sources: humanly judged INEX 2008 ad hoc topics, a MSN search engine log, and a proxy log from a New Zealand high school. We used a simple method of generating a document retrieval test collection from both a search engine's transaction log and a proxy log, based on submitted queries and user click-through data, and conducted a comparative analysis. Our analysis especially seeks to understand the differences between clicks and explicit judgments, and how these differences impact the comparative evaluation of retrieval effectiveness.

Our first research question was: are there significant differences between clicks and relevance judgments? The most striking difference between the manually created and assessed Ad hoc topics and the topics derived from the log files, is the number of relevant documents per topic. The topics derived from the MSN log mostly have one relevant document. Some of the proxy log topics have quite a few more relevant documents, with a maximum of 13 relevant documents per topic. However, in comparison with the ad hoc topics, with on average 69 relevant documents, it is clear that the proxy log clicks are far less exhaustive than the pools used in the INEX Ad hoc Track.

Our second research question is: are there significant differences between system rankings based on clicks and based on relevance judgments? The impact on the comparative evaluation of systems is considerable. There is reasonable agreement over all 163 runs between the three topic sets. For map, the ad hoc topics and the proxy log agree on 36% of the pairwise comparisons of systems, and the ad hoc and MSN log agree on 30% of the pairwise comparisons. For mean reciprocal rank, the agreement is somewhat higher with 44% between ad hoc and proxy log, and 38% between ad hoc and MSN log. However, the system-ranking for the best 10 runs per set differ radically. There is some evidence that the ad hoc ranking corresponds better to inherent system quality: systems ranked high/low on the ad hoc set tend to be also significantly better/worse on the other log-based sets. We further analysed the queries and clicks derived from the transaction logs, trying to uncover particular biases in the clicked pages. The main observation was that number of relevant pages whose title contained the complete query is low for the ad hoc judgments (6%), moderately high for the proxy log clicks (51%), and extreme for the MSN log click (95%). This casts considerable doubt on the ability of the search engine log-based test collections to measure recall-related aspects of retrieval systems.

One of the greatest attractions of log data is that it comes in large volumes. Our analysis necessitated us to focus on small samples of queries derived from huge log files. Although these samples reasonably agree with the overall statistics, we make no particular claims on the representativeness of these samples for the whole logs. Our investigation focused on informational queries in terms of [4]: queries with the intent to acquire some information present in some web pages. And an encyclopedia like Wikipedia is a prototypi-

cal example of a resource to satisfy informational information needs. Still, the near one-to-one relationship between queries and clicked Wikipedia pages in the search engine log suggests behavior very similar to navigational queries. But rather than navigating to a particular website (like the Wikipedia's home-page) the user was directed to the dedicated entry in Wikipedia—an interesting mix of informational and navigational intent.

Acknowledgments

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501). Marijn Koolen was supported by NWO (# 640.001.501).

REFERENCES

[1] A. Arampatzis, J. Kamps, M. Koolen, and N. Nussbaum. Deriving a domain specific test collection from a query log. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 73–80. Association for Computational Linguistics, 2007.

[2] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *Proceedings SIGIR 2007*, pages 455–462. ACM, 2007.

[3] S. M. Beitzel, E. C. Jensen, A. Chowdhury, and D. A. Grossman. Using titles and category names from editor-driven taxonomies for automatic evaluation. In *CIKM*, pages 17–23. ACM, 2003.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2): 3–10, 2002.

[5] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10:491–508, 2007.

[6] M. Chau, X. Fang, and O. R. L. Sheng. Analysis of the query logs of a web site search engine. *Journal of the American Society for Information Science and Technology*, 56(13): 1363–1376, 2005.

[7] A. Chowdhury and I. Soboroff. Automatic evaluation of world wide web search services. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 421–422, New York, NY, USA, 2002. ACM Press.

[8] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40:64–69, 2006.

[9] S. Dumais, T. Joachims, K. Bharat, and A. Weigend. SIGIR 2003 workshop report: implicit measures of user interests and preferences. *SIGIR Forum*, 37:50–54, 2003.

[10] INEX. INitiative for the Evaluation of XML retrieval, 2009. `http://www.inex.otago.ac.nz/`.

[11] B. J. Jansen. Search log analysis: What is it; what's been done; how to do it. *Library and Information Science Research*, 28(3):407–432, 2006.

[12] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing and Management*, 44:1251–1266, 2008.

[13] T. Joachims. Evaluating retrieval performance using clickthrough data. In *Proceedings of the SIGIR 2002 Workshop on Mathematical/Formal Methods in Information Retrieval*. ACM Press, 2002.

[14] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. ACM Press, New York NY, USA, 2005.

[15] S. Jones, S. J. Cunningham, R. J. McNab, and S. J. Boddie. A transaction log analysis of a digital library. *Int. j. on Digital Libraries*, 3(2):152–169, 2000. URL `citeseer.ist.psu.edu/jones00transaction.html`.

[16] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37:18–28, 2003.

[17] D. W. Oard and J. Kim. Modeling information content using observable behavior. In *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*, pages 38–45, 2001.

[18] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*, pages 43–52, New York, NY, USA, 2008. ACM.

[19] D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM Press, New York NY, USA, 2004.

[20] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–73, New York, NY, USA, 2001. ACM.

[21] TREC. Text REtrieval Conference, 2009. `http://trec.nist.gov/`.

[22] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2002.

[23] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314. ACM Press, New York NY, USA, 1998.