Click Log Based Evaluation of Link Discovery

David Alexander
Dept. of Computer Science
University of Otago
Dunedin, New Zealand
dalexand@cs.otago.ac.nz

Andrew Trotman
Dept. of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Alistair Knott
Dept. of Computer Science
University of Otago
Dunedin, New Zealand
alik@cs.otago.ac.nz

Abstract *We introduce a set of new metrics for hyperlink quality. These metrics are based on users' interactions with hyperlinks as recorded in click logs. Using a year-long click log, we assess the INEX 2008 link discovery (Link-the-Wiki) runs and find that our metrics rank them differently from the existing metrics (INEX automatic and manual assessment), and that runs tend to perform well according to either our metrics or the existing ones, but not both. We conclude that user behaviour is influenced by more factors than are assessed in automatic and manual assessment, and that future link discovery strategies should take this into account. We also suggest ways in which our assessment method may someday replace automatic and manual assessment, and explain how this would benefit the quality of large-scale hypertext collections such as Wikipedia.*

Keywords Information Retrieval, Hypertext

1 Introduction

Link discovery is the automatic generation of new hyperlinks in existing hypertext. In recent years, it has seen substantial development within the information retrieval community. This is likely due to the similarity between link discovery and search: both tasks involve retrieving documents that are *relevant* to an information need, be it expressed in a *search query* or in the *anchor* (the clickable text) of a hyperlink. By making the assumption that topical relevance solely determines hyperlink quality, one can split the task of link discovery neatly into the two subtasks of *selecting appropriate anchors* and *finding relevant targets* for those anchors.

We challenge this assumption. While topical relevance is often useful in hyperlinks, it is not the only useful quality, nor do existing evaluation methodologies necessarily measure it in the most appropriate way. We propose that instead of making subjective judgements of relevance, we should instead measure the usefulness of hyperlinks directly by analysing recorded user behaviour. We do this using click logs, which are readily obtainable and can contain vast quantities of data. In doing so, we assume nothing about the nature of useful hyperlinks — only about the nature of users' interactions with such links.

In this article, we introduce a set of metrics for hyperlink quality that are calculated from the data in a click log. We compare these to the two most popular existing methods of assessment — *automatic assessment* and *manual assessment* — which are explained below.

Automatic assessment involves establishing a ground-truth for relevant hyperlinks based on the existing hyperlinks in the corpus. A ranked list of hyperlinks produced through link discovery is compared to the set of links in this ground-truth, and the link discovery strategy under assessment is scored according to the degree of similarity observed. In producing this list of links, the link discovery algorithm cannot refer to the ground-truth; instead, it refers to an *orphaned* version of the hypertext document to be linked: that is, one that has had its incoming and outgoing links removed. Other documents in the corpus are available to the algorithm in unorphaned form. Manual assessment also uses a ground-truth but establishes it from the subjective judgements of human assessors.

Our goal is to use click logs to develop an assessment method that can replace automatic and manual assessment while retaining the advantages of both. This could ultimately enable link discovery to enter mainstream use, where link discovery could be kept in check by rigorous and automatic user evaluations generated from a live click log, and thus highly experimental link discovery strategies could be used without fear of low precision. To determine the possible implications of such a replacement, we investigate the following research questions in this article:

1. Are click-based assessments capable of distinguishing topically relevant from topically non-relevant hyperlinks?

2. How does the use of click-based assessment affect the ranking of existing link discovery strategies from the INEX evaluation forum?

We compare the results of automatic and manual assessment with the results of click-based assessment, both on a per-hyperlink basis and across the entire test

Proceedings of the 16th Australasian Document Computing Symposium, Canberra, Australia, 2 December 2011.

corpus (a set of Wikipedia articles). We examine the ranking that each assessment method gives to a particular set of link discovery algorithms, and find that some perform well under our assessment method, some perform well under the existing assessment method, but few perform well under both. The consequence of this is that neither topical relevance nor the structure of existing Wikipedia hypertext appear to determine user behaviour.

2 Prior Work

In this section, we outline the qualities of the existing assessment methods, so that in the next section we may show how our method improves upon them.

2.1 Development of assessment methods

The INEX Link-the-Wiki track ran annually from 2007 to 2010. It assessed link discovery algorithms on the Wikipedia encyclopaedia, by providing *topics* (Wikipedia articles that had been "orphaned" by having their links removed) to be re-linked to the rest of the corpus. The algorithms produced ranked lists of links for each topic. Automatic assessment simply compares these links to the pre-orphan links.

It was through Link-the-Wiki that manual assessment was first introduced. Huang et al. [5] performed an experiment at INEX 2008 in which they added the pre-orphan links (i.e. the automatic assessment groundtruth) to the pool of links to be assessed, and then asked participants to manually assess the resulting pool. They found that many of the pre-orphan links were assessed as non-relevant by the manual assessors, suggesting that automatic assessment based on these links was not accurate.

These evaluation methods can be seen as the beginning of a progression: first, a criterion for hyperlink quality (namely relevance) is specified, and in automatic assessment a particular source (Wikipedia) is assumed to exhibit this quality due to *implicit* norms among its authors; then, with manual assessment, assessors attempt to judge this criterion *explicitly*. Our click log based method is intended as a third step in this progression, whereby the measurement of hyperlink quality is not based on *any* assumptions or judgements, but is instead carried out through direct observation of users.

2.2 Quality of assessment data

The principle of automatic assessment is that it harnesses a vast quantity of data of assumed high quality. When automatic assessment is used, we can expect that new documents added to a corpus will be hyperlinked with quality approximately as good as that of the existing documents. However, the research of Huang et al. [5], showing that Wikipedia contains many links assessed as non-relevant by assessors, calls this quality into question.

By contrast, manual assessment allows assessors to determine hyperlink relevance however they see fit. This would allow them to take into account arbitrarily complex use cases of the hyperlinks — if it were not extremely time-consuming for them to do so. Instead, manual assessment relies on the assessor's intuitions and is prone to inconsistency.

This inconsistency has been observed in *ad hoc* information retrieval in several studies, including that of Sanderson et al. [9] who note that one's criteria for relevance develop gradually as more documents are assessed — and we believe that similar results can be expected in link discovery.

The data used in manual assessment is also often incomplete, but this problem can be mitigated in several ways, for example by using metrics such as *bpref* (Buckley & Voorhees [2]), which ignores unassessed results.

2.3 Relevance

Relevance, as the term is used in the evaluation of search engines, means that search results are of use to a person with a stated information need. In link discovery, it means that the *topic* of the target document is appropriate for the anchor text as it appears in the source document.[1]

This entails a very simple relationship between the source and target documents: it is only the strength of the connection between *those two documents* that determines the strength of the link. Indeed, link discovery algorithms based on this formulation of relevance need only to perform two tasks: (1) finding phrases in the new document that refer to the topics of existing documents; and (2) deciding which such phrases represent *topically relevant* hyperlinks. The second of these tasks is often implemented with a simple document similarity metric, or even left out altogether (as in the highly successful *Structural Threshold* algorithm of Itakura & Clarke [6]).

These tasks are analogous to the tasks of selecting and ranking results for a search engine. However, search engines are not used in the same way as freeform hypertext, nor should they be evaluated in the same way. Search engines use hyperlinks to point to specific results, whereas freeform hypertext uses them to provide context. This context can take many forms, some of which fit directly into the categories of *non-relevant* hyperlinks identified by Huang et al. [5] (such as links to Wikipedia articles listing general events in specific periods of history).

3 Our approach

A hypertext browsing session does not necessarily have the same structure as a search engine session. Though

[1] This distinction exists because users of hypertext do not specify *information needs* explicitly (as search engine users do with queries). This means that users' information needs must be assumed to relate to the documents they view.

users often arrive at a hypertext with a certain information need, there may be no single document that satisfies it completely; instead, multiple documents, and even the structure of the links between them, can provide valuable information to the user. This raises the question of what unit of information retrieval activity (e.g. the *session*, the *document*, the *passage*, and so on) should be measured.

Our goal is to evaluate individual hyperlinks, rather than entire hypertexts, because the granularity of per-link assessments is ideal for directing improvements to the hypertext. We therefore favour a bottom-up approach to assessment: instead of assigning click based scores to entire sessions (and using these to calculate scores for links), we directly assign a score to each link based on behaviour recorded from users interacting with that link.

3.1 Click log based metrics

One of the advantages of our approach is that it directly measures the effectiveness of hyperlinks in real-life user interactions. This involves first recording data from users and then interpreting that data. There are several ways of gathering data on user behaviour, but we believe that *click log analysis* is the best way because it allows for a continuous stream of data, unencumbered by the bias of explicit judgements, to be gathered without human intervention or the disruption of users.

The raw data of a click log consists of a sequence of page-views (with timestamps) for each user. No detailed information is recorded about users' interaction with their browsers, such as the specific anchors that are clicked to generate each page-view.

The data for our metrics, which are detailed below, is collected as follows. First, the requests recorded in a click log (the log we use in this work is described in Section 4.1) are grouped by user, and each user's entire history of requests is grouped into *sessions* such that no two consecutive requests within a session have a gap of more than 30 minutes between them. Then, page-views are associated with the links that appear to have caused them, if any.[2]

We introduce four metrics for calculating scores from this data. Each of our metrics builds on the previous, encapsulating more information about user behaviour.

The various metrics are all based on the idea of users implicitly "voting" for the quality of links. We assume that users have a limited number of votes to cast (because their time is limited) and therefore a vote for one link is cast at the expense of votes for lower-quality links. We also assume that even when users' votes are individually binary, as more votes are collected they will tend towards a proportionate representation of each link's quality.[3] We need not understand what *causes* each link to be worthy of inclusion: it may sometimes be strict relevance; other times, the context that it provides to the user; and perhaps occasionally, sheer curiosity on the part of users.

Our simplest metric is based on the assumption that high-quality links will be popular:

1. **Click-Voting (CV)** counts every click on a link as a vote in support of that link. It is assumed that users will be able to estimate the quality of links before clicking on them, and will therefore be more likely to click on high-quality links. $C(l)$ denotes the set of clicks recorded for the link l.

$$U_{\mathrm{CV}}(l) = |C(l)|$$

However, CV does not account for the problem whereby links that occur on popular pages receive unduly high numbers of votes. This causes the scores of links on the most frequently visited pages to dwarf all others. We address this problem by introducing another factor:

2. **Proportional Click-Voting (pCV)** measures the number of votes for each link as a proportion of the number of "voters" — i.e. visitors to the page containing the link. $V(l)$ denotes the set of page-views recorded for the source document of link l, whether or not they result in a click for that link.

$$U_{\mathrm{pCV}}(l) = |C(l)| \cdot \frac{1}{|V(l)|}$$

pCV is the probability that a visitor to the source page of a link will click that link, assuming each visitor clicks exactly one link. However, this probability is only an accurate measure of quality if the page visitor examines every link to decide whether to click it, and therefore the decision regarding a link's quality is made with full knowledge of the link's existence.

This is rendered unlikely by the phenomenon of *position bias*, in which links occurring early in a document are highly likely to be clicked regardless of the quality of links occurring later. This is assumed to be because users do not always view the later links. It can be expressed as a diminishing probability of links being clicked the further down the page they appear. We introduce a further factor to address this problem:

3. **Biased Proportional Click-Voting (bpCV)** corrects for position bias by estimating the probability that the link would be clicked if position bias did not exist. The model we use to calculate the correction factor, denoted by $B(l)$, is described in Section 3.2.

$$U_{\mathrm{bpCV}}(l) = |C(l)| \cdot \frac{1}{|V(l)|} \cdot B(l)$$

[2] This is determined for each page-view by finding the most recent previously viewed page in the same session that contains a link to the requested page.

[3] Specifically, the votes should approximate the *average* perceived quality over all users, since different users may find links to be more or less useful depending on their needs.

The normalisations in pCV and bpCV account for *false negatives* — i.e. links that are unused for reasons other than low quality. However, users cannot necessarily determine the quality of a link until after clicking it, which means that the use of a link does not necessarily imply that its quality is high. We address this by observing the behaviour of the user *after* clicking the link:

4. **Normalised Reading Time (nRT)** counts users' votes proportionally to their individually measured information gain upon clicking a link. This is represented by the amount of time that the user spends on the resulting page, measured as the number of seconds between the page request in question and the one that follows it.[4] This metric is normalised in the same way as above: i.e. by the number of source page visitors and the position bias. $R(c)$ denotes the reading time recorded for the page-view resulting immediately from the link-click c.

$$U_{\text{nRT}}(l) = \sum_{c \in C(l)} R(c) \cdot \frac{1}{|V(l)|} \cdot B(l)$$

We prefer Normalised Reading Time as a click-based metric, because it is an indicator of not only the number of users who found a link useful, but of *how much* useful information was gained. We believe that the time spent reading a page can be expected to correlate with the user's information gain since the assimilation of information is a time-consuming process which users are unlikely to perform on information that is not useful to them.

3.2 Position bias

While the other constituents of the click-based metrics can be calculated straightforwardly, the position bias is a quantity that must be estimated using a model of user behaviour.

Position bias can be modelled in numerous ways: we use the "cascade model" introduced by Craswell et al. [3], which states that the probability of a given link being clicked is dependent on both the user's perception of the relevance of the link and the probability that *none of the previous links were clicked* before reaching it. It assumes that the user views the links in a document from top to bottom, and that once a link is clicked, the user does not return to click further links.

If c_i is the event of clicking link l_i, and v_i is the event of viewing link l_i, the equation given by Craswell et al. for the cascade model shows how to calculate $P(c_i)$ given $P(c_i|v_i)$ (assuming the reader is already on the page containing l_i). We have $P(c_i)$ empirically — it is equivalent to the pCV metric — so we rearrange the equation to give $P(c_i|v_i)$, i.e. the probability of the link being clicked if position bias did not exist. To calculate this, we multiply the value of pCV with the following correction factor, derived from the equation in Craswell et al.:[5]

$$B(l_i) = \prod_{j=0}^{i-1} (1 - P(c_j) \cdot B(l_j))^{-1}$$

4 Experiments

In this section we present the results of two experiments.

The first experiment is a "sanity check" that establishes that the click-based metrics are sensitive to topical relevance (as well as to other qualities), and therefore are in general agreement with systems that measure topical relevance.

The second experiment determines the practical implications of using a click-based assessment method rather than manual assessment by performing INEX-style assessments on the runs submitted to INEX 2008. One set of assessments is produced for each of our four metrics, with each metric providing a score for each link. According to INEX tradition, Mean Average Precision scores are calculated for each run from the per-link scores. We also use the per-link scores to calculate Normalised Discounted Cumulative Gain scores. We then compare the rank-order of runs under these click-based assessments to the rank-order given by INEX manual assessment.

Before presenting the results of these experiments, we describe the data sources and our handling of them.

4.1 Data source

To gather data for click-based scores, we use the 2008 click log of the University of Otago student web proxy, which records all student web access through campus computers. Our anonymising process involves removing non-Wikipedia requests and personal information such as usernames, but allows us to separate the requests of the 17,635 unique users recorded in the log.

Although the recorded usage (and therefore our evaluation) pertains to a constantly changing Wikipedia, the INEX assessments against which we compare our evaluations are based on a static snapshot of Wikipedia. This snapshot was taken in 2005,[6] and included the 659,388 articles that Wikipedia contained at the time.[7] However, the 2,595,572 Wikipedia requests in the click log cover only 451,979 unique articles.

[4] This cannot be calculated for the last page-view in each session, since no page-view follows it. Instead, the reading time for the last page-view is estimated (pessimistically) as the average reading time across the session.

[5] If multiple anchors with the same target occur in a document, we use the first anchor to calculate this factor.

[6] For further information on the Wikipedia collection and how it is used in INEX assessments, consult the following papers from 2006 — the year in which the collection was introduced — Denoyer & Gallinari [4] and Malik et al. [8].

[7] The 2005 Wikipedia snapshot was used because it was the official document collection for INEX 2008. We chose INEX 2008 because it was run at approximately the same time as our click log was recorded.

Name	Significance	Significance (adj.)
CV	p = 0.000	p = 0.001
pCV	p = 0.004	p = 0.009
bpCV	p = 0.000	p = 0.000
nRT	p = 0.000	p = 0.000

Table 1: Results of *t*-tests comparing click-based scores for manually-judged relevant and non-relevant links. (Experiment 1)

Because of the limited scope of the click log, many hyperlinks in Wikipedia have no recorded clicks in the log. There are 24,094,179 hyperlinks in all of the requested Wikipedia articles, but only 473,510 of these have any recorded clicks. Of these, only 311 also have manual assessments. For each distinct hyperlink that is clicked, a set of click-based scores can be calculated from our four metrics.

In short, there are 473,510 distinct sets of click-based scores — one for each hyperlink — that contribute to the calculation of the performance metrics in our second experiment, but only 311 sets of click-based scores are used in the first experiment (because it only applies to the intersection of manual and click-based assessments). However, this number of assessable links is sufficient for our current purpose: to examine our new evaluation method and compare its results to those of the existing methods. In the Further Work section, we suggest sources of further click log data and ways in which the existing data could be extended to cover more links.

4.2 Experiment 1: compatibility

The first of our two research questions concerns the extent to which click-based metrics can distinguish between topically relevant and non-relevant links. We acknowledge that manual assessment, despite its problems, is the most accurate way of determining the relevance (although not necessarily the overall quality) of hyperlinks. Therefore, we use INEX manual assessment as the benchmark for relevance in the first experiment.[8]

In this experiment, we use our click-based metrics to calculate scores for individual hyperlinks. We hypothesise that these scores will be higher for hyperlinks that are judged relevant by assessors than for those that are judged non-relevant, because we believe users will be inclined to click topically relevant links more often than others. We test this by taking the set of hyperlinks for which both click log data and manual assessments are available, and splitting it into two subsets — the *relevant* set and the *non-relevant* set — according to the manual assessments. We find that the average click-based score in the *relevant* set is indeed higher than in the *non-relevant* set, regardless of which of our four click-based metrics is used.

Because of the small sample size for this experiment (311 hyperlinks), we considered the possibility that the difference in click-based scores was due to chance. However, unpaired one-tailed *t*-tests (the results of which are shown in Table 1) show that the differences are significant to the 1% level.

The use of four separate click-based metrics necessitates a familywise adjustment to the *p*-values in these *t*-tests. Popular adjustments such as the Bonferroni correction are too conservative for our purposes because they assume that the experiments are statistically independent, which is not true in our case since each click-based metric is based upon the previous one. Instead, we use a correction suggested by Blakesley [1, pp. 37–38], which accounts for statistical dependence by using correlation coefficients to calculate the adjusted *p*-value.

These results suggest that our metrics are in general agreement with manual assessment: they can distinguish relevant links from non-relevant links as accurately as manual assessors can, while also providing a graded measure of hyperlink quality that manual assessors cannot provide.

4.3 Treatment of INEX Runs

Because the INEX runs specify multiple targets (a maximum of 5) for each anchor, the entire list of anchor/target pairs for a given document is not intended to be treated as a ranked result list: instead, the anchors are ranked, and the targets for each anchor are ranked separately. To reflect this, our click-based score for each anchor is calculated as the mean click-based score over all of the targets of that anchor. The list of anchors is then treated as a ranked list.

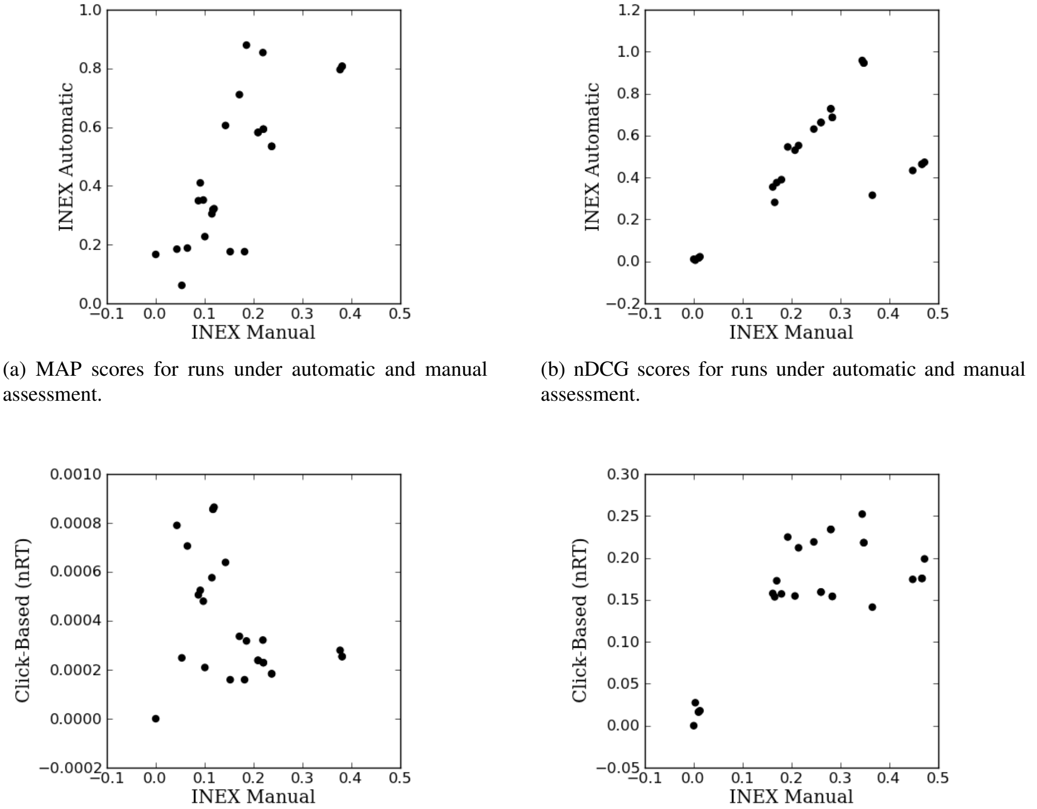
No matching is performed between the anchor positions (in the document) specified by INEX runs and the anchor positions in Wikipedia: instead, the quality of a link is calculated based on all clicks that have the same source and target documents as that link, regardless of anchor text. This is equivalent to INEX *file-to-file* evaluation, which is ideal for our purposes because it avoids the problems of matching anchors in different versions of documents when the click log does not record the necessary information to do so.

Source/target pairs found in a run but not in Wikipedia are considered "unassessed", and removed from the ranked list before assessment.[9] Source-target pairs that exist, but are never clicked, are given a click-based score of zero for all four metrics.[10]

[8] We do not use automatic assessment for this experiment because the set of links that it would judge non-relevant is identical to the set of links for which click log data is unavailable.

[9] To determine whether this removal would bias the results, we performed the same processing on the INEX manual assessments and compared the resulting ranking with the original ranking. The Spearman's rank correlation coefficient between the two rankings is 0.853, which we consider high enough to indicate a negligible difference.

[10] All of the links that we assess occur in documents that are visited at least once in the log. This means that although the click



(a) MAP scores for runs under automatic and manual assessment.

(b) nDCG scores for runs under automatic and manual assessment.

(c) MAP scores for runs under INEX and click-based assessment, using nRT metric.

(d) nDCG scores for runs under INEX and click-based assessment, using nRT metric.

Figure 1: Scatter plots showing the difference between the automatic/manual comparison and the manual/click-based comparison. (Experiment 2)

Of the 32 runs submitted to INEX 2008, not every run includes all 50 assessed topics. To avoid the unfairness that would occur if we used incomplete topics in our evaluation, we remove 4 of the runs[11] (because they include too few topics) and only assess according to a particular set of 22 topics,[12] chosen so that: first, each of the remaining runs includes all of the topics in the set; and second, each of the topics in the set includes links for which both manual and click-based assessments were available.

4.4 Experiment 2: correlation

We use two performance metrics to aggregate the per-link click-based scores (and relevance judgements, for the INEX assessments) across the entire set of topics used. These metrics are Mean Average Precision (MAP) and Normalised Discounted Cumulative Gain (nDCG). We use MAP because it is well understood in the information retrieval community, and nDCG because it is designed specifically for use with graded judgements such as ours. To use MAP with graded judgements, we modify the standard formula for average precision as follows (where L is the set of links specified by a run for a given topic):

$$\mathrm{AvgP} = \frac{1}{|L|} \sum_{l \in L} \frac{\text{Total of all scores down to } \mathrm{rank}(l)}{\mathrm{rank}(l)}$$

The formula for nDCG is used without modification, and is defined as follows. (iDCG is the *ideal DCG*: the DCG that a run would score for a given topic if it returned all possible relevant/useful links in the best possible order.)

$$\mathrm{DCG} = \mathrm{rel}(l_1) + \sum_{l \in L, l \neq l_1} \frac{\mathrm{rel}(l)}{\log_2 \mathrm{rank}(l)}$$

$$\mathrm{nDCG} = \frac{\mathrm{DCG}}{\mathrm{iDCG}}$$

Because several institutions participated in the Link-the-Wiki track in 2008, we are able to evaluate a wide variety of link discovery strategies using our

count of a link may be zero, none of the normalisations of our click-based metrics are ever undefined due to division by zero.

[11] A fifth run is also removed due to improper specification of anchors.

[12] The list of topics is as follows: *Air conditioning*, *Anne Rice*, *Boston Celtics*, *Boston Tea Party*, *Chamonix*, *Data mining*, *De Stijl*, *Greenhouse gas*, *Information retrieval*, *Love*, *Mainz*, *Mouse*, *Natural gas*, *Near and far field*, *New Zealand Labour Party*, *Personal name*, *Ratan Tata*, *Search engine*, *Ski jumping*, *Studio Ghibli*, *Symphony*, and *Togo Heihachiro*.

	Auto	Manual	CV	pCV	bpCV	nRT
Auto	1.00					
Manual	0.74	1.00				
CV	0.49	-0.03	1.00			
pCV	0.24	-0.27	0.88	1.00		
bpCV	0.09	-0.34	0.59	0.75	1.00	
nRT	0.05	-0.36	0.58	0.75	0.99	1.00

(a) Rank correlation matrix for MAP scores under different assessment methods.

	Auto	Manual	CV	pCV	bpCV	nRT
Auto	1.00					
Manual	0.61	1.00				
CV	0.72	0.39	1.00			
pCV	0.72	0.39	1.00	1.00		
bpCV	0.72	0.39	1.00	1.00	1.00	
nRT	0.77	0.57	0.89	0.89	0.89	1.00

(b) Rank correlation matrix for nDCG scores under different assessment methods.

Table 2: Correlation matrices for assessment metrics. (Experiment 2)

evaluation method. This includes the University of Amsterdam's algorithm, based on *link likelihood ratio*, and QUT's algorithm, based on *title matching*.

Most notably, however, it includes a number of implementations of the Itakura & Clarke [6] algorithm: from the University of Waterloo, the University of Otago, and Lycos. This algorithm was particularly popular in 2008 because of its high performance in 2007.

The purpose of our experiment is to examine the relationship between the existing assessment methods and our own. Because any change in assessment methodology is prone to causing changes in the rank order of runs, we compare the nature of these changes from automatic to manual assessment with those from manual to click-based assessment, rather than examining the latter in isolation.

The scatter plots in Figures 1a and 1b show the relationship between automatic and manual assessment (under MAP and nDCG, respectively). This relationship is mostly linear, albeit with a prominent group of outliers that perform well under manual assessment but not as well under automatic assessment when measured using nDCG.

Figures 1c and 1d show the relationship between manual and click-based assessment. At first glance, it appears that there is no correlation. However, closer inspection reveals a linear correlation up to a certain point, after which two lines appear, each encompassing runs that perform well under one assessment method but not the other. This effect is particularly noticeable under MAP, where the top-right quadrant of the scatter plot is conspicuously sparse.

This suggests that the relationship between high performance under the two assessment methods is not positive or negative, but is mutually exclusive in the existing set of link discovery algorithms. This does not mean that an algorithm cannot satisfy the criteria of both — only that no current algorithm does so.

We analyse the results presented in these scatter plots using Spearman's rank correlation coefficient.[13] Tables 2a and 2b show the correlation matrices for all assessment methods (automatic, manual and the four click-based metrics) under MAP and nDCG. Predictably, the click-based metrics correlate well with each other, especially under nDCG. Also, automatic assessment correlates well with manual assessment, but neither correlates as highly with the click-based metrics as the click-based metrics do with each other. Surprisingly, automatic assessment correlates better with click-based assessment than with manual assessment.[14]

5 Discussion

The differences in rank order of runs between INEX manual assessment and click-based assessment are not surprising, given that click-based assessment uses large quantities of user data rather than the judgement of a few assessors.

However, the implications of our click-based method range further than simply ranking link discovery runs. Using click logs as a data source enables assessment to be performed in ways that were not possible in past, such as by automatically checking the quality of links as they are added to and removed from an online corpus such as Wikipedia, and adjusting linking over time.

Furthermore, our work suggests that the established order of link discovery strategies is not final, but rather that it may be affected by criteria of hyperlink quality other than topical relevance. This is an important result because link discovery has recently come to be regarded as a solved problem, owing to the excellent performance of the Itakura & Clarke [6] algorithm under automatic and manual assessment.

Since this algorithm selects new links based on their probability of occurring in the existing hypertext, it disregards context and simply mimics the surface structure of the hypertext. Itakura et al. [7] acknowledge that it is surprising that their algorithm should perform so well given its simplicity. Our introduction of the click-based metrics is promising for the continuation of link discovery as an open field of research.

There are also practical advantages to using a click log based method rather than manual assessment. Click

[13] We choose Spearman's coefficient (rather than Pearson's coefficient) because it is non-parametric: i.e. it disregards the curvilinearity of the line and only takes into account the rank order of the data points. It has previously been established in information retrieval that the rank order of performance scores such as MAP and nDCG is more important that their absolute values.

[14] This appears to be a result of the very low-scoring cluster visible in the bottom-left corners of the nDCG graphs.

logs provide large quantities of data essentially for free, whereas manual assessment involves a high cost even to assess a small number of documents. Manual assessment typically uses a small number of assessors each performing a large number of assessments, whereas click logs record a large number of users each performing small tasks. According to the findings of Sanderson et al. [9] (that an assessor's assessment criteria tend to "shift" over time), manual assessment can also be expected to be less reliable than click log based assessment. Although it is too early to say that our method can replace manual assessment, it is clearly desirable that it be made able to do so. The results of our first experiment, showing that our metrics are sensitive to topical relevance, suggest that this may be possible.

6 Further work

The simplest improvement to our work would be to find larger and more diverse click logs to use. The click log that we use here does not have very wide coverage of Wikipedia articles, and it omits some information that could reasonably have been recorded, such as the *referrer* of each request. The best way to implement a click-based assessment system would be to use the click logs of the website that was being assessed (e.g. Wikipedia). However, even a large proxy log such as that of an ISP would be useful.

Also, by making some assumptions about what kinds of links are likely to share similar click-based scores, it would be possible to use a relatively sparse set of click-based judgements — such as those that we have generated from the University of Otago click log — to assess an entire corpus. This could be done simply by using known click-based scores to estimate the click-based scores of similar links (for example, links that shared the same anchor text and target document). However, this technique would have to be used carefully to avoid diluting the specificity of our metrics to the point where they was no longer dependent on users' information needs, as inferred from the context of links.

The availability of good click log data would also be improved by using "live" click logs. Click logs are typically produced by web servers and proxy servers as part of their normal operation, and they are therefore constantly being updated. However, research into click logs tends to be done on portions that are recorded over fixed time periods (for example, our click log is from 2008). If an assessment methodology similar to the one used here were instead used on a continuously updating click log — perhaps through a program installed on the web server for the website under assessment — its assessments could be used to automatically prune low-quality hyperlinks, and add new hyperlinks that were similar to existing high-quality ones. In conjunction with an automated link discovery system (which would also be applied continuously) this would automate much of the time-consuming task of hyperlinking new documents, while avoiding the low precision that has previously made automated link discovery impractical for live websites.

7 Conclusion

In this article, we have introduced a series of metrics of hyperlink quality which are based on user behaviour. These metrics can be used to calculate per-link scores (which could help hypertext authors decide which links to remove) or they can be used to evaluate entire hypertext collections according to standard information retrieval performance measures such as Mean Average Precision (MAP) and Normalised Discounted Cumulative Gain (nDCG).

We have justified these metrics theoretically, and verified that they produce similar relevance classifications to a baseline set of results from INEX manual assessment. We have also examined the relationship between INEX assessment (both automatic and manual) and our click-based metrics when applied to INEX runs. We have seen that runs from the currently available set diverge into specialised subsets — one for each assessment method — as their performance improves. This suggests that our criteria should be considered alongside the existing criteria when link discovery strategies are developed, and that the task of meeting these criteria presents substantial opportunities for further work in link discovery.

References

[1] R.E. Blakesley. *Parametric control of familywise error rates with dependent P-values*. Ph.D. thesis, University of Pittsburgh, 2008.

[2] C. Buckley and E.M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR*, pages 25–32. ACM, 2004.

[3] N. Craswell, O. Zoeter, M. Taylor and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM*, pages 87–94. ACM, 2008.

[4] L. Denoyer and P. Gallinari. The wikipedia xml corpus. In *ACM SIGIR Forum*, Volume 40, pages 64–69. ACM, 2006.

[5] W.C. Huang, A. Trotman and S. Geva. The importance of manual assessment in link discovery. In *SIGIR*, pages 698–699. ACM, 2009.

[6] K. Itakura and C. Clarke. University of Waterloo at INEX2007: Adhoc and Link-the-Wiki tracks. *Focused Access to XML Documents*, pages 417–425, 2008.

[7] K.Y. Itakura, C.L.A. Clarke, S. Geva, A. Trotman and W.C. Huang. Topical and Structural Linkage in Wikipedia. *ECIR*, 2011.

[8] S. Malik, A. Trotman, M. Lalmas and N. Fuhr. Overview of inex 2006. *Comparative Evaluation of XML Information Retrieval Systems*, pages 1–11, 2007.

[9] M. Sanderson, F. Scholer and A. Turpin. Relatively Relevant: Assessor Shift in Document Judgements. *ADCS 2010*, pages 60, 2010.