How Consistent Are Human Judgments of Whether an Open Resource is Educational Material?

Michael C. Harris **James A. Thom** **Falk Scholer**

School of Computer Science and Information Technology, RMIT University,
GPO Box 2476, Melbourne, Victoria 3001, Australia.
Email: {michael.c.harris,james.thom,falk.scholer}@rmit.edu.au

Abstract

Systems that filter web search results to return open educational resources need evaluation. The Cranfield method, which is widely used in information retrieval evaluation, can be used as the basis of a model for evaluating such systems. The Cranfield method requires a collection of resources with associated judgments. In this paper, we describe an experiment to collect judgments of whether open resources are educational material. We show experimentally that judges can agree on what resources are educational material, even in the absence of an educational context, and demonstrate that displaying the query used to retrieve the resource makes a judge less likely to rate a resource as educational.

Keywords: Open Educational Resources, Information Retrieval, Inter-rater Agreement

1 Introduction

The increase in the use of technology to support learners has led to a radical increase in the amount of digital teaching and learning material. Reusing existing material can increase the productivity of teaching staff, provide academics with more time to spend on value-added activities such as interacting with students, increase the return on investment of creating high-quality resources, and lead to the improvement of the quality of resources (Harris & Beiers 2005). However, the benefits of reuse cannot be realised if appropriate resources cannot be found.

This work explores a key issue related to the construction of collections appropriate for the evaluation of systems that filter web resources to identify learning material. The concept "likely to support learning" is not well defined, so complete agreement between judges rating resources according to that concept is unlikely. We investigate if people can broadly agree on what resources are likely to support learning in the face of this ambiguity.

This paper is organised as follows. We begin in Section 2 by discussing related research. In Section 3 previous work on agreement evaluation is detailed. In Section 4 we describe the user experiment conducted in this research, the results of which are analysed in Section 5. We then discuss what these results suggest about the ability of judges to agree on whether a resource is educational, and what impact this has on an evaluation methodology for systems that filter educational resources. We conclude by outlining possible future work.

2 Related work

Significant time and money has been spent in the private and public sectors developing and maintaining systems designed to manage educational material, often called learning objects. For example, CANARIE[1], the Canadian Advanced Network and Research for Industry and Education alone has spent almost $C30 million on online learning projects, with the design and development of e-learning repositories being a major research theme (MacLeod 2005).

These repositories often expose incompatible access interfaces and contain few resources (Neven & Duval 2002). Research on the effective retrieval of educational material has assumed that all resources being searched are learning objects. However, many educational resources are released on the World Wide Web, and clearly there is much more on the Web than just learning material.

When publicly released, these resources are known as Open Educational Resources (OERs) (Wiley 2007), which are generally defined as "technology-enabled, open provision of educational resources for consultation, use and adaptation by a community of users for non-commercial purposes" (UNESCO 2002).

Previous work suggests that when people want to find digital material to support learning, they prefer to use a public search engine, such as Google[2] (Harris & Beiers 2005, Griffiths & Brophy 2005). These users may be more satisfied with search engine results if only resources likely to support learning were presented. To provide satisfactory result sets, resources that are unlikely to support learning should not be present. A filter to detect material that is likely to support learning would therefore be useful.

To ensure effectiveness, filters should be systematically evaluated. This evaluation can be carried out by assessing filter performance on a dataset where each resource has been labelled according to whether it should be retained or filtered out. This labelled dataset is called a *ground truth* or *gold standard*. This paper examines the establishment of a ground truth for evaluating systems that filter web resources for educational material.

The notion of a ground truth is also used in information retrieval (IR) system evaluation, and this suggests a useful starting point for developing an evaluation methodology for learning material filtering. The ground truth for IR systems evaluation is typically constructed by having relevance judgments assigned to resources by human judges. Systems are

[1] http://www.canarie.ca
[2] http://www.google.com

measured based on their ability to approximate the human-assigned relevance judgments. This is known as the *Cranfield method* after experiments carried out at Cranfield University in the 1960s (Cleverdon 1967).

The Cranfield method requires a collection of documents, a set of queries, and a set of relevance judgments linking the documents and the queries (Hildreth 2001). It is the standard approach used in information retrieval evaluation, and is used for evaluation in, for example, the Text REtrieval Conference (TREC) (Buckley & Voorhees 2005) and the Cross Language Evaluation Forum (CLEF) (Braschler & Peters 2002). This evaluation methodology has also been adapted for use in other retrieval evaluation domains, such as for XML retrieval evaluation (Kazai et al. 2003).

We propose that, using the Cranfield method as a model, systems that filter learning material can be evaluated based on their ability to select those resources that have been categorised by human judges as educational.

Experiments based upon the Cranfield method make the assumption that relevance is a property of resources in relation to a query, independent of the user. Under this assumption, the user and the context of retrieval is completely represented by the query (Saracevic 2007). While there has been debate about the validity of this assumption, it has been a useful starting point for IR experiments in general (Buckley & Voorhees 2005). We make a similar simplifying assumption, that a resource can be judged educational or not, independent of the specific educational context. Though context obviously plays an important role in education, we believe making the assumption of context independence is appropriate for the development of an evaluation methodology for retrieving educational resources. Our experiment explores this assumption.

In IR systems, the ground truth is constructed by assessing relevance, a concept which is complex and multi-dimensional (Saracevic 2007). For filtering educational resources, the ground truth should have assigned judgments of whether resources are educational or likely to support learning, which is also a complex concept. We investigate whether people can broadly agree on what resources are likely to support learning in the face of this complexity.

The level of agreement between raters will assist in deciding how many judgments are needed for each resource to establish an accurate ground truth. The use of a single assessor to judge the relevance of each resource has been a criticism of experiments based on the Cranfield method (Harter 1996). However, it is common in IR experiments to use a test collection where each resource has been assigned a relevance assessment by a single assessor, and this methodology has been shown to be adequate on small collections (Burgin 1992).

We further examine whether displaying the query used to retrieve a resource influences judgments and affects agreement.

Collaborative recommendation systems also use ratings that users assign to resources (Adomavicius & Tuzhilin 2005). In collaborative recommendation, ratings are collected from users in production systems and used to suggest resources to other users. Also related are information filtering systems, which find or remove resources from an incoming stream of data based on user profiles (Belkin & Croft 1992). However, these systems differ from judgments in IR evaluation experiments, and to what we propose in this paper, in that we collect ratings based on independent assessments of items; that is, the ratings have no relationship to users' general profiles.

3 Agreement evaluation methodology

In this section we provide some background on measures of agreement. Alongside our discussion of agreement measures, we present a worked example of each method, in the domain of judging educational material.

3.1 Overlap

The usefulness of larger test collections with single assessments was experimentally supported in relation to the TREC collections by Voorhees, who showed that, despite variability of individual relevance assessments, the relative ranking of systems is stable (Voorhees 1998). In this work, TREC collections were reassessed by additional judges, and the level of agreement between all judges was measured using *overlap* (Voorhees 1998).

Overlap is the mean of the size of the intersection of positive ratings divided by the size of the union of positive ratings for each resource. That is, the average of the number of times both judges rated the resource relevant (for our purposes, rated the resource educational) divided by the number of times either judge rated the resource relevant. Voorhees reports the overlap measures between each of the three judges, and overlap across the three judges. However, the value of this overlap calculated across all judges will decrease as the number of judges increases, as a single dissenting judge counts as disagreement. For this reason, mean pairwise overlap is a more useful measure.

We use the example data from Table 1 to illustrate all agreement measures. Let there be J judges and R resources. For our example we use three judges ($J = 3$) each rating five resources ($R = 5$). A value of 0 represents a judgment that a resource was not educational, and a value of 1 represents a judgment that a resource was educational.

Table 1: Example ratings of three judges on five resources

	Resources				
Judges	1	2	3	4	5
1	0	0	1	1	1
2	0	0	0	1	1
3	0	1	1	0	1

For our example, pairwise overlap can be calculated as follows. We have three judges, and three pairwise comparisons; judge 1 with judge 2, judge 1 with judge 3, and judge 2 with judge 3. Consider judges 1 and 2; they agree that resources 4 and 5 are educational, so the intersection is 2. However, judge 1 also rated resource 3 as educational, so the union is 3. Overlap for these two judges is therefore $\frac{2}{3} = 0.667$. Mean pairwise overlap is the average overlap across all pairs of judges, as shown in Table 2.

Table 2: Example mean pairwise overlap

Judges	intersect	union	pairwise overlap
1 & 2	2	3	0.667
1 & 3	2	4	0.500
2 & 3	1	4	0.250
mean pairwise overlap			0.472

3.2 Raw agreement

A further simple measure commonly used is *raw agreement*, which is the proportion of observed agreement to possible agreement. Uebersax says, "Much neglected, raw agreement indices are important descriptive statistics. They have unique common-sense value. A study that reports only simple agreement rates can be very useful; a study that omits them but reports complex statistics may fail to inform readers at a practical level." (Uebersax 2008)

In the context of assessments of whether resources are OERs, if a random resource is selected from a test collection, and we select a random rater who has judged the resource an OER, what is the probability that another random judge will agree? If the proportion of negative and positive judgments differ greatly, overall agreement will be biased towards the dominant judgments (Kundel & Polansky 2003). This often happens in relevance judgments, where there are likely to be far fewer documents that are relevant to a query than irrelevant documents. It is therefore important that the levels of positive and negative agreement are reported separately.

We derive the measure for raw agreement, limiting our discussion to the binary case, which we use in our experiment below. See Uebersax (2008) for calculation of raw agreement with an arbitrary number of categories.

We calculate raw agreement as follows. Let p_r be the number of times resource r was positively rated and n_r be the number of times resource r was negatively rated. To illustrate, we can see from Table 1 that resource 5 has three positive judgments, giving $p_5 = 3$, and resource 2 has two negative judgments, giving $n_2 = 2$. The number of times a rating was given to each resource in our example is shown in Table 3.

Table 3: Number of positive and negative ratings

	Resources				
Rating	1	2	3	4	5
n	3	2	1	1	0
p	0	1	2	2	3

There are p_r positive ratings for resource r. If we take one judge who rated resource r positively, there are $p_r - 1$ judges that agree. Raw agreements are bi-directional, so total positive agreement is calculated by $p_r(p_r - 1)$. Negative agreement is calculated in the same way, taking negative ratings instead of positive ratings.

From the example judgments, take resource 4 from Table 3. We see that $p_4 = 2$, meaning that 2 judges said resource 4 is educational. Thus, the total number of agreements that resource 4 is educational is $p_4(p_4 - 1) = 2(1) = 2$

Therefore, the observed agreement across all R resources can be expressed as follows, with A^-_{obs} representing observed agreement on negative judgments and A^+_{obs} observed agreement on positive judgments.

$$A^-_{obs} = \sum_{r=1}^{R} n_r(n_r - 1)$$

$$A^+_{obs} = \sum_{r=1}^{R} p_r(p_r - 1)$$

The number of possible agreements, A^-_{poss} for negative agreement and A^+_{poss} for positive agreement, can be calculated similarly, but instead of taking the number of judges that agreed with the original judge, we take the number of judges that *could* have agreed. Since a judge cannot agree with themself, this means possible agreement is one fewer than the total number of judges, or one fewer than the total number of ratings, $(p_r + n_r - 1)$, and thus positive agreement for a resource is $p_r(p_r + n_r - 1)$.

$$A^-_{poss} = \sum_{r=1}^{R} n_r(n_r + p_r - 1)$$

$$A^+_{poss} = \sum_{r=1}^{R} p_r(n_r + p_r - 1)$$

The observed and possible agreement for our example are shown in Table 4.

Table 4: Observed and possible agreement for each resource

	Resources					
	1	2	3	4	5	Total
A^-_{obs}	6	2	0	0	0	8
A^-_{poss}	6	4	2	2	0	14
A^+_{obs}	0	0	2	2	6	10
A^+_{poss}	0	2	4	4	6	16

Therefore, we can calculate specific agreement for both the positive (A^+) and negative (A^-) cases to be the proportion of observed agreement to possible agreement.

$$A^- = \frac{A^-_{obs}}{A^-_{poss}}$$

$$A^+ = \frac{A^+_{obs}}{A^+_{poss}}$$

For our example data, this means we have $A^- = \frac{8}{14} = 0.571$ and $A^+ = \frac{10}{16} = 0.625$.

Overall agreement can similarly be calculated by dividing the observed agreement from both positive and negative judgments by the number of possible agreements.

$$A = \frac{A^-_{obs} + A^+_{obs}}{\sum_{r=1}^{R}(n_r + p_r)(n_r + p_r - 1)}$$

Of course, $n_r + p_r$ is constant, the number of judgments made on a resource, and therefore the number of judges. We defined the number of judges earlier as J, so the overall agreement can simplified to the following.

$$A = \frac{A^-_{obs} + A^+_{obs}}{R \cdot J \cdot (J - 1)}$$

Using our example data, for overall agreement we have $A = \frac{8+10}{(5)(3)(2)} = \frac{18}{30} = 0.6$.

3.3 Kappa

The disadvantages of the overlap and raw agreement measures are that they are not corrected for chance, and it is not possible to estimate a confidence interval (Kundel & Polansky 2003). The index κ has been developed to address these issues; Cohen's κ for two raters (Cohen 1960) and Fleiss' κ for multiple raters (Fleiss 1971).

In calculating Fleiss' κ we ask the question, given that we have some set of observed judgments, what agreement would we expect by chance? The proportion of agreement expected by chance can be represented as $\bar{P}_e$. If we take this value away from perfect agreement, we have the best agreement possible, $1-\bar{P}_e$. If we take the chance agreement away from what was observed, which we can call $\bar{P}_o$, and divide it by the best possible agreement, we have the proportion of agreement that is not due to chance.

Therefore, κ can be defined as follows, with a value of 1 indicating complete agreement, and a value less than 0 representing agreement less than would be expected by chance.

$$\kappa = \frac{\bar{P}_o - \bar{P}_e}{1 - \bar{P}_e}$$

It is then possible to calculate the standard error, and a confidence interval. Calculating κ for our example data, we have $\kappa = 0.196$ and $p = 0.447$. Therefore, though there is some agreement above chance, our example data does not show statistically significant agreement.

The table developed by Landis and Koch (Landis & Koch 1977) is sometimes used as a way to interpret values of κ. However, the levels chosen are arbitrary, are not applicable across experiments (Sim & Wright 2003) and can lead to unreliable conclusions (Gwet 2001). For these reasons, we do not report our results in relation to the Landis and Koch table.

While κ is used as a measure of agreement, it is not a test of the effect of classifying resources using two methods. We may observe significant agreement both when judging resources with a query visible and with the query not visible, but we want to be able to compare the levels of agreement. For this we use Fisher's exact test (Agresti 1992), which tests the null hypothesis that there is no difference in the proportions that raters assign resources to different categories under each condition.

In the next section, we describe our experiment design. As explained in this section, we take the Cranfield method as a starting point for our work.

4 Experiment design

Evaluation of systems that filter e-learning material differs from evaluation of relevance using the Cranfield method in that it seeks to collect classifications of resources according to a concept (supporting learning) as opposed to drawing a relationship (relevance) between a query and a document. To investigate how such classifications should be collected, we conducted a user experiment. The experiment investigates whether human judges can agree on what resources are likely to support learning, and whether visibility of the query used to retrieve the resource has an effect on judgments.

Eight judges were recruited for the experiment. Participants were acquaintances of the first author, from diverse backgrounds, and all had some experience with using web browsers and web interfaces.

A total of 20 resources were judged by the eight judges under one of two conditions: the query used to retrieve the resource being visible (q) or not visible (q'). Each judge viewed ten resources under condition q and ten under q'. A latin squares design was used to control for ordering effects.

4.1 Resource selection

The Flexible Learning Toolboxes are a collection of OERs managed by e-Works[3] that comply with the Sharable Content Object Reference Model (SCORM). A log of queries submitted to the live repository of the e-Works collection was obtained, containing 21139 queries, 7764 of them unique. Queries were drawn at random from the unique queries. If it was judged improbable that submitting a particular query to a search engine would return educational resources, that query was discarded. For example, the queries "rte2606a" and "a*" were discarded. A total of 20 queries were selected for our experiments, and these are shown in the query terms column of Table 5.

While the queries were originally used for seeking resources from an educational repository, the resources used for our experiment were retrieved from a search across the entire web, with each query being submitted to the Yahoo! Search API[4]. Alongside each selected query, Table 5 shows the resources used in our experiment, which were selected as follows.

For the first ten queries, resources returned at rank position one were selected for judging. Call this set of resources R_A, resources 1 to 10 from Table 5.

To ensure that the collection contained an adequate proportion of resources for which a positive judgment was probable, resources returned using the second ten queries were judged by one of the experimenters, in rank order, according to the same criteria that would ultimately be used by the participants. For each query, the highest ranked resource judged likely to support learning was added to the collection. These judgments were made without reference to the search query used to retrieve the resources. Call these resources R_B, resources 11 to 20 from Table 5.

4.2 Presentation and user interface

Five resources from R_A and five from R_B were combined and their order randomised to form the first pool, P_1. The remaining resources were combined and their order randomised to form P_2. The judgment pools contained the following resources.

$$P_1 = [4, 11, 15, 5, 3, 13, 2, 12, 1, 14]$$

$$P_2 = [19, 6, 16, 18, 10, 8, 17, 20, 9, 7]$$

Resources were presented to judges in four ways, as described below.

Group 1) P_1 with the query followed by P_2 without the query.

Group 2) P_1 without the query followed by P_2 with the query.

Group 3) P_2 with the query followed by P_1 without the query.

Group 4) P_2 without the query followed by P_1 with the query.

For example, Group 1 were presented the resources from P_1 with the original search query displayed, and then the resources from P_2 were presented without the original search query displayed. The ten resources in P_1 and P_2 were always presented in the same order.

Two judges were randomly assigned to each group.

Resources were presented sequentially via a web interface. Judges were asked to classify resources as likely or unlikely to support learning. Specifically,

[3] http://www.eworks.edu.au

[4] http://developer.yahoo.com/search/web/

Table 5: Resources in collection.

resource	query terms	rank	URL (http://)
R_A			
1	Communicate with colleagues and clients in an office environment	1	www.visualdatallc.com/clients.aspx
2	food safety practices	1	www.ehow.com/how_2075133_practice-food-safety.html
3	cash flow	1	en.wikipedia.org/wiki/Cash_flow
4	Maintain equipment for activities	1	www.govliquidation.com/list/c7484/lna/1.html
5	safe beauty	1	www.safeinternetshops.com/beauty.htm
6	search internet	1	kaftos.com
7	software development	1	en.wikipedia.org/wiki/Software_development
8	lifting safely	1	www.smarter.com/---se--qq-Lifting+Safely.html
9	cash budget	1	www.investopedia.com/terms/c/cashbudget.asp
10	Process customer complaints	1	www.bcuc.com/Complaint.aspx
R_B			
11	costing ingredients	4	www.ces.ncsu.edu/depts/poulsci/tech_manuals/ingredient_sampling.html
12	prepare cook and serve food	3	www.cfsan.fda.gov/~dms/hret2-2.html
13	computing skills	3	www.cnn.com/TECH/computing/9806/03/autism.idg
14	treat weeds	1	www.blm.gov/weeds/FAQs/FAQs.htm
15	library skills	31	www.rock-hill.k12.sc.us/schools/high/sphs/Media/libraryskills.htm#Mod1
16	(plan and conduct and meetings)	3	www.azskillsusa.org/Teachers/meetings.htm
17	administer projects	8	sais-jhu.edu/cmtoolkit/issues/evaluation/index.html
18	Coordinate implementation of customer service strategies	50	en.wikipedia.org/wiki/Customer_relationship_management
19	write simple documents	1	nadoka.vipnet.org:8080/doc/user/08_is1.htm
20	Arts administration	3	en.wikipedia.org/wiki/Arts_administration

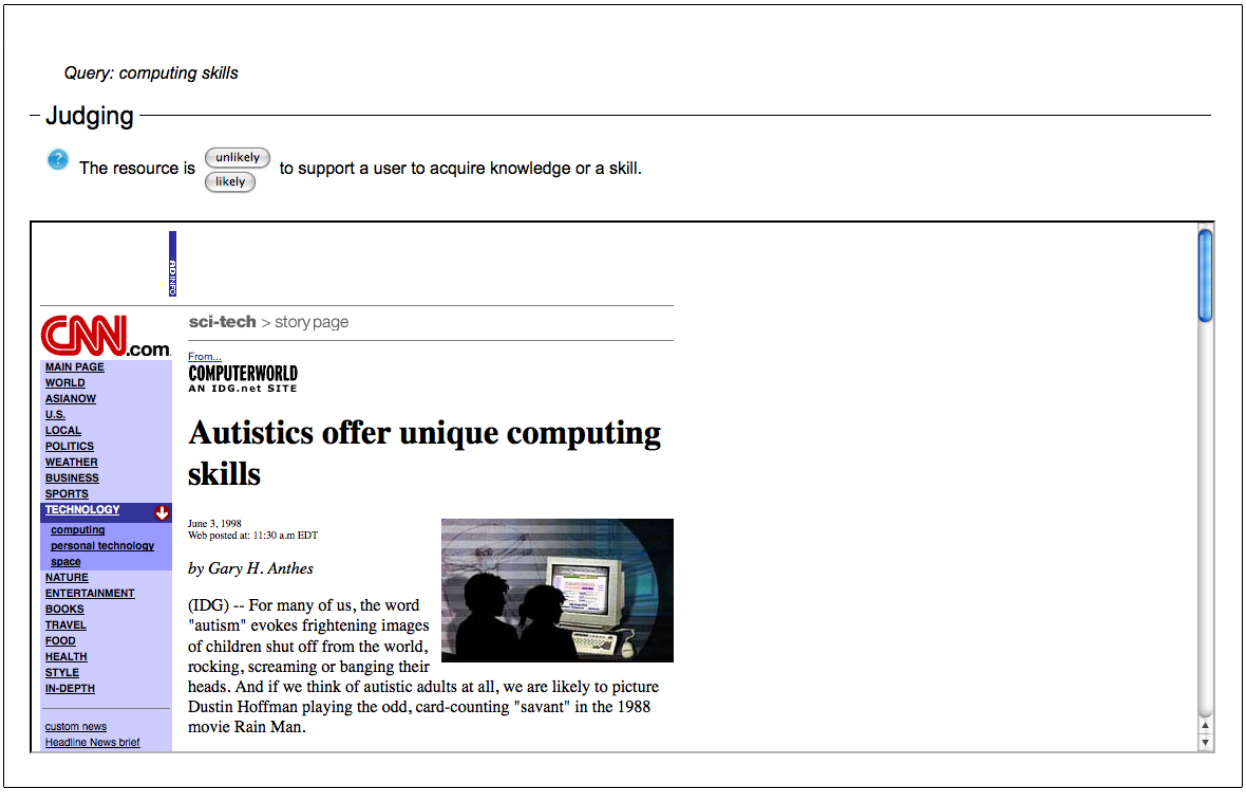

Figure 1: Judgment interface with query.

they were presented with the statement, "The resource is likely/unlikely to support a user to acquire knowledge or a skill," where the words "likely" and "unlikely" were buttons that recorded the judgment. The judgment interface with the query visible is shown in Figure 1.

An HTML iframe element was used to embed single page resources in the judgment interface. All links within the resources were disabled, and raters evaluated the resources without reference to other web pages.

The resource displayed in Figure 1 is one of the resources used in our experiment, resource 13 from Table 5. The judgment interface without the query visible was identical, save for the removal of the query. Overall, judgments for this resource were split, with four judges believing it was educational material and four believing it was not. When the query used to retrieve the resource was not displayed, three of the four judges who assessed this resource said it was educational material. However, when judges could see the query, only one of the four assessors judged the resource as educational material.

Each participant answered several questions after each judgment, and after they completed all judgments, including being asked for comments about the resource they had just judged.

5 Analysis

This section reports on the results of our experimental evaluation of rater agreement, and provides analysis of these results. To give a general picture of the judgments, the number of times a resource was judged educational is shown in Figure 2. For our analysis, we begin by investigating rater agreement across all judgments regardless of other factors. Second, we discuss the impact of query visibility on rater agreement. We then report on the influence of resource type, that is, whether the resource was included in the judgment pool as a first ranked resource, or as a resource that was pre-selected to be a likely educational resource. Finally, we conclude by providing a discussion of comments that raters made after judging each resource.

Table 6: General agreement

Overlap	0.595
Negative	0.633
Positive	0.749
Overall	0.702
κ	0.382 ($p < 0.001$)

5.1 General rater agreement

The frequency with which a number of raters judged resources as educational is shown in Figure 3. The leftmost bar represents the number of resources that no rater judged as educational, and the rightmost bar represents the number of times that all raters judged a resource educational.

We see a bimodal distribution, with higher frequencies at the extremes. This is as expected if there is a high level of agreement.

The agreement measures between the eight judges observed across all resources are presented in Table 6. All measures suggest a high level of agreement, and the value of κ is highly significant. The calculated mean pairwise overlap measure between the eight judges is 0.595, compared with the mean pairwise overlap measure between three assessors of 0.447 shown in Voorhees' work that justified the use of a single judge in relevance assessments.

5.2 Query visibility

When building a collection for assessing systems using the Cranfield method, an assessor makes a judgment about the relevance of a document to a query. The query is therefore central to the process,

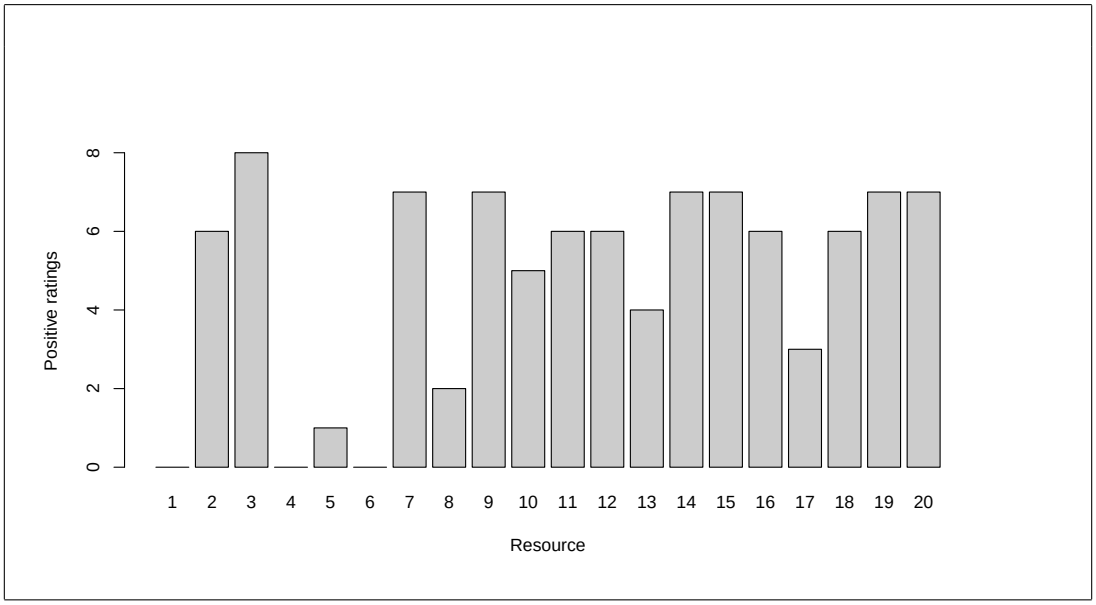

Figure 2: Positive OER judgments by resource and query visibility.

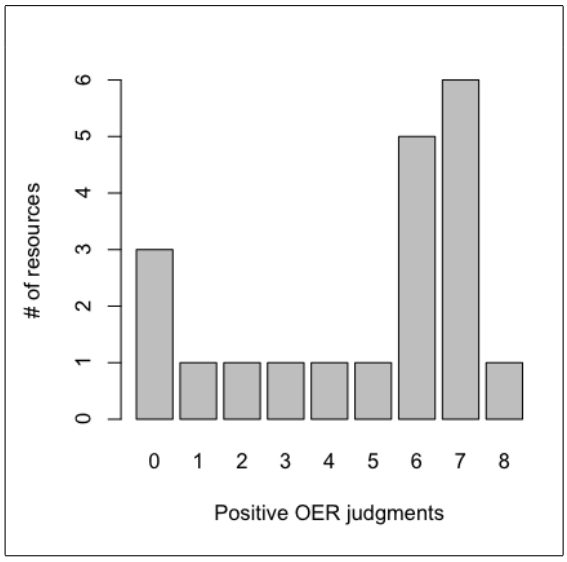

Figure 3: Overall frequency of positive OER judgments.

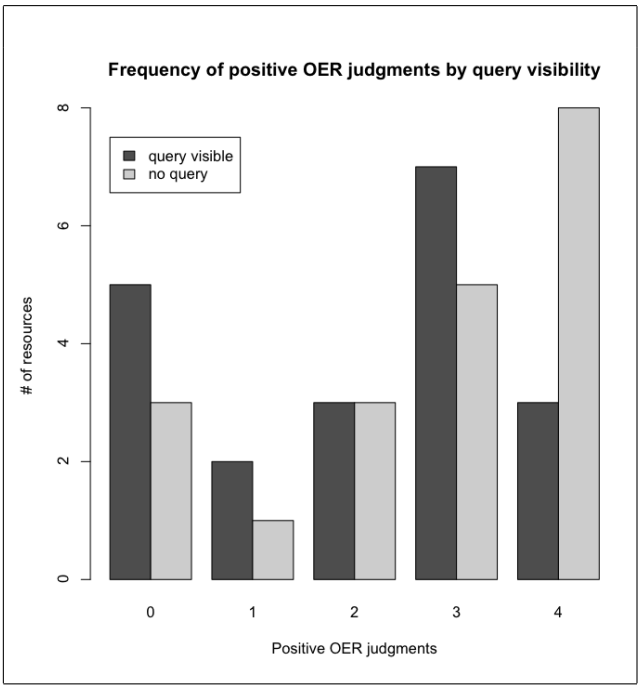

Figure 4: Frequency of positive OER judgments by query visibility.

and different queries will cause the resource to be judged differently. However, when judging whether a resource is educational, the judgment criteria is stable, and it is unclear what effect query visibility would have on the judging process. Here we report the results of varying query visibility.

Each resource has an *a priori* probability of being judged likely to support learning. In this case, we are interested in the conditional probability, that is, the probability a resource will be judged likely to support learning given query visibility.

Figure 4 shows the frequency with which a number of raters judged a resources as educational, separated by query visibility. Each resource was judged by four judges under each condition. The leftmost bar shows that, with the query visible five resources received no positive ratings, while without the query visible three resources received no positive ratings. The rightmost bar indicates that all raters judged a resource educational on three occasions when the query was visible and on eight occasions when the query was not visible.

As with Figure 3, bimodal distributions indicate a high level of agreement. The distributions of frequency with and without the query being visible do appear to be generally bimodal, however, inspection suggests that displaying the query makes it less likely that a resource will be judged educational.

The agreement measures when split by query visibility are presented in Table 7. We can see that on all measures except negative agreement, agreement is noticeably higher when the query is not visible, though κ is significant in both cases. There is a very high level of positive agreement when the query is not visible, meaning that when the query is not visible raters very often agree that a resource is educational. It appears that judges use different criteria to rate a resource when the query is visible.

Fisher's exact test indicates that query visibility has a weakly significant effect on judgments ($p = 0.053$).

Table 7: Agreement by query visibility

	query	no query
Overlap	0.516	0.685
Negative	0.667	0.615
Positive	0.683	0.815
Overall	0.675	0.750
κ	0.350 ($p < 0.001$)	0.430 ($p < 0.001$)

Table 8: Agreement by resource group

	R_A	R_B
Overlap	0.622	0.583
Negative	0.805	0.272
Positive	0.762	0.741
Overall	0.786	0.618
κ	0.567 ($p < 0.001$)	0.013 ($p = 0.827$)

5.3 Resource rank

Figure 5 shows the positive ratings made on each resource, that is, ratings where raters judged the resource educational, separated by query visibility. As described in Subsection 4.1, resources 1 through 10 were included in the judgment pool because they were returned at rank position one in response to a search for the first 10 queries (R_A), and the resources 11 through 20 were included in the judgment pool because they were the highest ranked resource judged educational from the results returned in response to the second 10 queries (R_B).

The agreement measures when split by resource group are presented in Table 8. On all measures, agreement is lower for resources in R_B, and though we see similar values for overlap, positive agreement and overall agreement, κ does not show significant agreement for R_B. Negative agreement is particularly low for R_B when compared with R_A.

Fisher's exact test indicates that how a resource was added to the judgment pool has a significant effect on judgments ($p < 0.001$). This means that the proportions of negative and positive judgments are different depending on whether the resource was a first ranked resource or was the highest ranked resource judged to be educational.

5.4 Rater comments

In the judgment interface raters were invited to make comments about their judgments. In total, raters made 91 comments from the 160 judgments. Seven of the eight raters made at least one comment after judging a resource. Approximately a third of the comments make reference to the query used to retrieve the resource. For example, after judging a resource with the query visible one rater said, "Query asking general question; resource for much more specific request which is likely irrelevant. Therefore, easy to judge," and another said, "query not specific enough," and "if it was autism and computing skills this would be a useful resource."

In some cases, the rater stated that they found the resource difficult to judge because the query was not known, "Specific resource and without search terms, difficult to determine whether relevant to query; therefore, difficult to judge."

These comments appear to suggest that raters find it more difficult to judge whether a resource is educational in the absence of the context given by a query. It might be expected that a more difficult judgment decision would take longer to make; however timing measurements reveal that there is no significant interaction between judging times and query visibility.

6 Discussion and future work

The methodology presented in this paper produces a ground truth that can be used in the evaluation of systems that filter web search results for educational resources. This methodology is based upon the Cranfield method, which is commonly used in IR experiments. Further, we establish how many judges should rate each resource, and whether the queries used in the retrieval of resources should be presented to assessors as part of the judgment interface.

We present a user experiment in which participants judged whether web resources were educational, in a manner similar to the way relevance assessments are collected when building test collections for use in experiments using the Cranfield method.

Our results show a high level of general agreement. Indeed, our results show a level of agreement higher than that used in the IR literature to justify the use of a single assessor. We conclude that, given this high level of agreement, an appropriate methodology for building a test collection for the evaluation of systems that filter for educational resources would involve having resources categorised by a single judge rather than have multiple judges categorise each resource. In particular, for fixed time and number of judges, it would be more useful to judge a larger number of resources than have multiple judgments on fewer resources.

In relation to query visibility, our results show a high level of agreement both with and without the query visible. The level of agreement is higher when the query is not visible, though this is only weakly significant overall. While we found nothing to suggest that the query needs to be displayed, judges did report that they found the task difficult. The task may be made simpler by asking judges to rate resources in an artificial context and then to make a judgment as to whether the resource is educational in other contexts.

This effect is significant when the resource is not the first ranked result. This result is intuitively reasonable, as the search engine used for retrieving the resource has rated the resource as less relevant than other resources, as reflected in its ranking. It may be that the query distracts raters from the task of judging whether a resource is likely to support learning, and causes them to judge relevance instead.

When people use search engines generally, they issue a query and judge how well the documents returned meet their information need. That is, they judge the relevance of returned resources to their query. Therefore, when presented with a resource to judge, and the query that was used to retrieve it, it is unsurprising that their judgments reflect relevance. As we are interested in filtering educational resources, relevance is handled by the search engine, and therefore should be factored out for our purposes.

Our experiment used a fixed description when asking judges to rate resources. Future work could examine what instructions should be given to judges and what effect this may have on agreement.

The selection of resources appropriate for inclusion in a test collection must also be addressed. We contend that it is appropriate to submit queries to a search engine and select returned resources for the collection. In this work, initial queries came from a log of queries submitted to a repository for e-learning material. This is appropriate in that the users were searching for the type of material in which we are interested. However, a user's search behaviour may be different when searching a specific repository as opposed to the wider web, and thus the queries may not be representative of the sorts of queries that would be submitted to a filtering system. Equally, queries

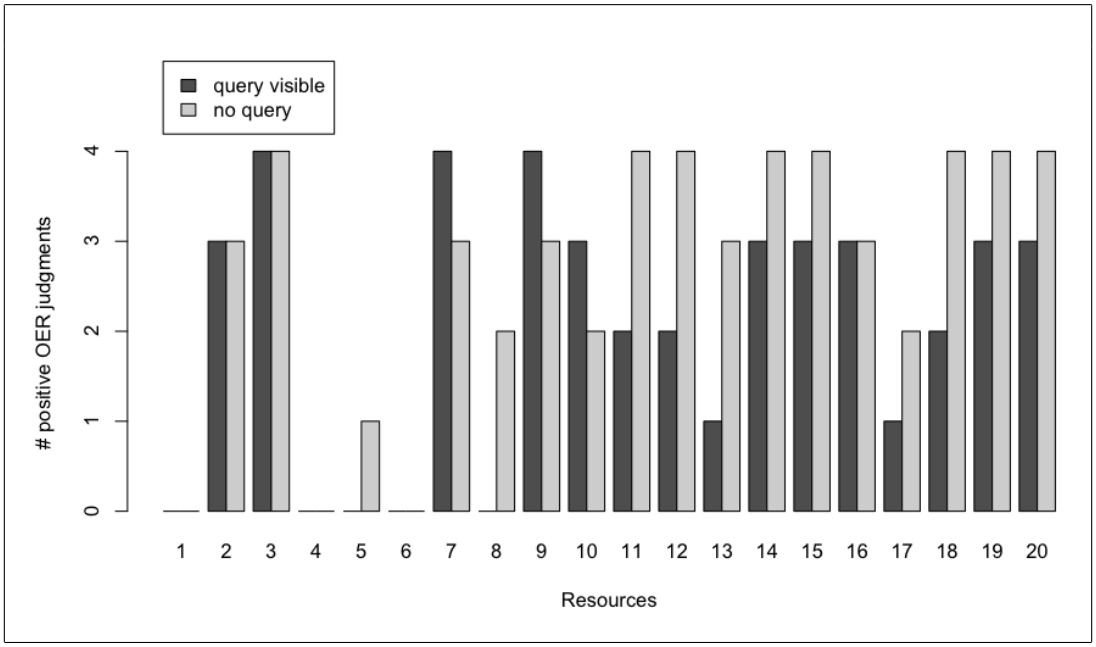

Figure 5: Positive OER judgments by resource and query visibility.

selected from a general query log, without knowledge of user intent, are likely to be inappropriate.

Resources must then be selected from the ranked resources returned by a search engine in response to a query. Table 5 shows that most of the resources R_B (those included in the pool because they were judged likely to support learning) were ranked in the top 10 results. The mean rank was 7.7 and the median was 3. Also, half the resources in R_A (those included as the first ranked result) were judged educational by a majority of the judges. This suggests that a reasonable percentage of highly ranked resources will be judged likely to support learning, and that we need not have taken the precaution of pre-judging some resources. Therefore, it is appropriate to include the first N results from the returned ranked results in the collection. The results of this preliminary study suggest that $N = 10$ is sufficient.

Our future work will involve the construction of a test collection appropriate for the evaluation of educational resource filters. We will construct such filters, using some of the features from the resources judged in this work as starting points, and evaluate them using the methodology outlined here. For example, resources not judged educational often include a large proportion of links when compared to content, whereas resources judged educational often have large amounts of text separated by headings and a higher proportion of internal links.

One shortcoming of this work, and an opportunity for future work, is that we have only considered single page resources. However, it is likely that multi-page resources are more interesting from a learning point of view. This is related to the concept of *granularity* as used in reference to reusable learning objects.

References

Adomavicius, G. & Tuzhilin, A. (2005), 'Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions', *IEEE Transactions on Knowledge and Data Engineering* **17**(6), 734–749.

Agresti, A. (1992), 'A survey of exact inference for contingency tables', *Statistical Science* **7**(1), 131–153.

Belkin, N. J. & Croft, W. B. (1992), 'Information filtering and information retrieval: two sides of the same coin?', *Communications of the ACM* **35**(12), 29–38.

Braschler, M. & Peters, C. (2002), CLEF methodology and metrics, *in* C. Peters, M. Braschler, J. Gonzalo & M. Kluck, eds, 'Cross-Language Information Retrieval and Evaluation', Springer-Verlag, Lecture Notes in Computer Science, Vol. 2406, pp. 394–404.

Buckley, C. & Voorhees, E. (2005), *Retrieval System Evaluation*, MIT Press, Cambridge, MA, USA, pp. 53–75.

Burgin, R. (1992), 'Variations in relevance judgments and the evaluation of retrieval performance', *Information Processing and Management* **28**(5), 619–627.

Cleverdon, C. (1967), 'The Cranfield tests on index languages devices', *Aslib Proceedings* **19**(6), 173–194.

Cohen, J. (1960), 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* **20**(1), 37–46.

Fleiss, J. L. (1971), 'Measuring nominal scale agreement among many raters', *Psychological Bulletin* **76**(5), 378–382.

Griffiths, J. & Brophy, P. (2005), 'Student searching behaviour and the web: use of academic resources and Google', *Library Trends* **53**(4), 539–554.

Gwet, K. (2001), *Statistical tables for inter-rater agreement*, STATAXIS Publishing Company.

Harris, M. C. & Beiers, H. (2005), Barriers to the reuse of learning objects, *in* 'EdMedia 2005 - World Conference on Educational Multimedia, Hypermedia & Telecommunications', pp. 482–489.

Harter, S. P. (1996), 'Variations in relevance assessments and the measurement of retrieval effectiveness', *Journal of the American Society for Information Science* **47**, 37–49.

Hildreth, C. R. (2001), 'Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study', *Information Research* **6**(2).

Kazai, G., Gövert, N., Lalmas, M. & Fuhr, N. (2003), The INEX evaluation initiative, *in* H. Blanken, T. Grabs, H.-J. Schek, R. Schenkel & G. Weikum, eds, 'Intelligent search on XML data', Springer-Verlag, Lecture Notes in Computer Science, Vol. 2818, pp. 279–293.

Kundel, H. L. & Polansky, M. (2003), 'Measurement of observer agreement', *Radiology* **228**(2), 303–308.

Landis, J. R. & Koch, G. G. (1977), 'The measurement of observer agreement for categorical data', *Biometrics* **33**, 159–174.

MacLeod, D. (2005), Learning object repositories: Deployment and diffusion, E-learning report, CANARIE Inc., Ottawa, Ontario.
URL: `http://http://www.canarie.ca/funding/elearning/2005_LOR_final_report.pdf`
Accessed: 12 September 2008.

Neven, F. & Duval, E. (2002), Reusable learning objects: a survey of LOM-based repositories, *in* 'Proceedings of the tenth ACM international conference on Multimedia', pp. 291–294.

Saracevic, T. (2007), 'Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance', *Journal of the American Society for Information Science and Technology* **58**(13), 1915–1933.

Sim, J. & Wright, C. C. (2003), 'The kappa statistic in reliability studies: use, interpretation, and sample size requirements', *Physical Therapy* **85**(3), 257–268.

Uebersax, J. (2008), 'Raw agreement indices'.
URL: `http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm`
Accessed: 10 February 2009.

UNESCO (2002), 'UNESCO promotes new initiative for free educational resources on the Internet'.
URL: `http://www.unesco.org/education/news_en/080702_free_edu_ress.shtml`
Accessed: 18 February 2009.

Voorhees, E. M. (1998), Variations in relevance judgments and the measurement of retrieval effectiveness, *in* 'SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM, pp. 315–323.

Wiley, D. A. (2007), On the sustainability of open educational resource initiatives in higher education, Technical report, Organisation for economic co-operation and development (OECD).
URL: `http://www.oecd.org/dataoecd/33/9/38645447.pdf`
Accessed: 17 February 2009.