Repeatable and Reliable Search System Evaluation using Crowdsourcing

Roi Blanco
Yahoo! Research
Diagonal 177
Barcelona, Spain
roi@yahoo-inc.com

Harry Halpin
University of Edinburgh
10 Crichton St.
Edinburgh, UK
H.Halpin@ed.ac.uk

Daniel M. Herzig
Institute AIFB
Karlsruhe Institute of Technology
76128 Karlsruhe, Germany
herzig@kit.edu

Peter Mika
Yahoo! Research
Diagonal 177
Barcelona, Spain
pmika@yahoo-inc.com

Jeffrey Pound
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada
jpound@cs.uwaterloo.ca

Henry S. Thompson
University of Edinburgh
10 Crichton St.
Edinburgh, UK
ht@inf.ed.ac.uk

ABSTRACT

The primary problem confronting any new kind of search task is how to boot-strap a reliable and repeatable evaluation campaign, and a crowd-sourcing approach provides many advantages. However, can these crowd-sourced evaluations be repeated over long periods of time in a reliable manner? To demonstrate, we investigate creating an evaluation campaign for the semantic search task of keyword-based ad-hoc object retrieval. In contrast to traditional search over web-pages, object search aims at the retrieval of information from factual assertions about real-world objects rather than searching over web-pages with textual descriptions. Using the first large-scale evaluation campaign that specifically targets the task of ad-hoc Web object retrieval over a number of deployed systems, we demonstrate that crowd-sourced evaluation campaigns can be repeated over time and still maintain reliable results. Furthermore, we show how these results are comparable to expert judges when ranking systems and that the results hold over different evaluation and relevance metrics. This work provides empirical support for scalable, reliable, and repeatable search system evaluation using crowdsourcing.

Categories and Subject Descriptors

H.3.3 [**Information Storage Systems**]: Information Retrieval Systems

General Terms

Performance, Experimentation

Keywords

crowdsourcing, search engines, retrieval, evaluation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'11, July 24–28, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0757-4/11/07 ...$10.00.

1. INTRODUCTION

Advances in information retrieval have long been driven by evaluation campaigns using standardized collections of data-sets, query workloads, and most importantly, result relevance judgments. TREC (Text REtrieval Conference) [20] is a forerunner in IR evaluations, but campaigns also take place in specialized forums like INEX (INitiative for the Evaluation of XML Retrieval) [12] and CLEF (Cross Language Evaluation Forum). The main premises of these campaigns is that a limited and controlled set of human *experts* decide the correctness of a given set of results, which will be used as a ground truth for evaluating the performance of different systems [20]. Early evaluation campaigns targeted relatively narrow domains and used small collections, where evaluations using a small number of queries provided robust results. Moving to the open domain of the Web resulted in significantly larger heterogeneity of data sources and an increase in the potential information needs (and so diverse tasks) that need to be evaluated. Current research in campaigns (like TREC) and information retrieval evaluation in general focus primarily on the following goals:

Repeatability - As observed by Harter [11], there can be substantial variation among different expert judges performing the same task. If evaluation is to drive the next generation of search technologies, it is important to validate that relevance assignment is a repeatable process. This fundamental requirement exacerbates the scalability problem, because the agreement between assessors needs to be tested not only for each new search task, but also for each set of judges that have been employed (Agreement is a measure of the extent to which judges are interchangeable). However, outsiders who would like to validate an experiment will typically not have access to the original judges (or those judges may not be available or willing to repeat experiments at later times).

Reliability - The expert judges employed by campaigns such as TREC are expected to be sufficiently reliable to produce a ground truth for evaluation. However, setting up new "tracks" for novel search tasks is often not feasible or expedient, due to the time and effort it takes to set up such

tracks and the limited resources of the organizers. In such cases, researchers need to set up their own evaluation and seek replacements for experts, training others to be judges of their work, where training is often nothing more than providing a description of the task.

How can researchers create repeatable and reliable evaluation campaigns that scale over the number of new tasks brought about by the Web? An increasingly popular way of evaluating novel search tasks is the approach known as *crowdsourcing*. Crowdsourcing is a method of obtaining human input for a given task by distributing that task over a large population of unidentified human workers. In the case of building a search evaluation collection, crowdsourcing means distributing relevance judgments of pooled results over this crowd. The advantage of the crowd is that it is always available, it is accessible to most people at a relatively small cost, and the workforce scales elastically with increasing evaluation demands. Further, platforms such as Amazon Mechanical Turk[1] provide integrated frameworks for running crowdsourced tasks with minimal effort. We show how crowdsourcing can help execute an evaluation campaign for a search task that has not yet been sufficiently addressed to become part of a large evaluation effort such as TREC: ad-hoc Web object retrieval [17], for which we created a standard data set and queries for the task of object retrieval using real-world data, and the way we employed Mechanical Turk to elicit high quality judgments from the noise of unreliable workers in the crowd. The queries, index used, and results of the evaluation campaign are also publicly available for use in the evaluation of web-object retrieval systems[2].

There are two research questions that must be answered for crowdsourcing to be used systematically in evaluation campaigns. First, are evaluation campaigns with crowdsourced workers **repeatable**, such that the resulting ranking of systems is the same for different pools of crowdsourced judges over a period of time? Second, are crowdsourced workers **reliable**, such that differences between experts and crowdsourced workers do not change the resulting ranking of the systems? As our primary contribution, we experimentally demonstrate the repeatability of our search system evaluation experiment using crowdsourcing. We also test the reliability of judges who are not task or topic-experts, which has been questioned in previous work [3], as crowdsourced workers do not have access to the original information need and may lack specialized training or background knowledge possessed by experts. The case of Mechanical Turk provides an extreme where the judges are not only likely to be untrained and non-expert, but they also sign up for payment and so have an incentive to "cheat" in order to gain monetary reward. Therefore, we repeat our evaluation and assess whether the results from the original campaign can be reproduced after six months with a new set of crowdsourced judges, and whether those results correspond to what we would have obtained using a more traditional methodology employing expert judges. We also explore the effect of different numbers of judges per result on the quality of judgments. Finally, we analyse the robustness of three popular information retrieval metrics under crowdsourced judgments. The metrics studied are discounted cumulative gain (NDCG), mean average precision (MAP), and precision at k (P@k). To the best of our knowledge, we are the first to analyze the repeatability of crowdsourcing in a real-world evaluation campaign.

[1] `http://www.mturk.com`

[2] `http://anonymized.for.review`

2. EVALUATING OBJECT RETRIEVAL

The broad class of search technologies that exploit semantic data encodings are often called *semantic search* systems [9]. Largely due to the Semantic Web and related efforts such as Linked Data,[3] there is an increasing amount of structured data on the Web about real-world objects and their relations, given in the RDF (Resource Description Framework) format. We focus in particular on the class of semantic search systems that apply traditional information retrieval techniques directly on Semantic Web data, so that we evaluate keyword search over data in RDF. We are motivated by the increasing need to locate particular information quickly and effectively and in a way that is accessible to non-expert users.

There are already a number of semantic search systems that crawl and index Semantic Web data such as [7, 16], and there is active research into algorithms for ranking in this setting [8]. Despite the growing interest, there has been no standardized evaluation campaign for semantic search. One of the principle reasons for the lack of a standardized evaluation campaign is the cost of creating a new and realistically sized "gold-standard" data-set and annual evaluation campaign was considered too high by the community. We believe crowdsourcing can solve this problem.

2.1 Ad-Hoc Object Retrieval

Arriving at a common evaluation methodology requires the definition of a shared task that is accepted by the community as the one that is most relevant to potential applications of the field. The definition of the task has also been a precondition for establishing a set of procedures and metrics for assessing performance on the task, with the eventual purpose of ranking systems [20]. For the field of text-based information retrieval, this task is the retrieval of a ranked list of (text) documents from a fixed corpus in response to free-form keyword queries, or what is known as the ad-hoc document retrieval (ADR) task.

For the field of semantic search, Pound et al. [17] defined the task of ad-hoc object retrieval (AOR), where the goal is to retrieve a ranked list of objects (resources in RDF parlance) from a collection of RDF documents in response to free-form keyword queries. The unit of retrieval is thus individual objects and not RDF documents, and so the task differs from classic textual information retrieval insofar as the primary unit is structured linked data rather than unstructured textual data. Pound et al. also proposed an evaluation protocol and tested a number of metrics for their stability and discriminating power. In our current work, we instantiate their methodology in the sense of creating a standard set of queries and data on which we execute the methodology using a crowdsourcing approach. As keyword search over RDF is broadly comparable to existing evaluation campaigns (such as keyword search over XML and textual ad-hoc document retrieval), so the general results of our research should hold over a variety of ad-hoc ranking tasks, especially if the information need is clearly specified about

[3] `http://linkeddata.org`

a single entity and the systems have significant performance differences.

2.2 Data Collection

Current semantic search engines have vastly different indices, with some specializing on only single data-sources with thousands of triples and others ranging over billions of triples crawled from the Web. Therefore, in order to have a generalized evaluation of the ranking of results, it is essential to normalize the index.

We required a data-set that would not bias the results towards any particular semantic search engine. The data-set that we wanted to use in the evaluation campaign needed to contain real data, sizeable enough to contain relevant information for the queries, yet not so large that its indexing would require computational resources outside the scope of most research groups. We have chosen the 'Billion Triples Challenge' 2009 data set, a data-set created for the Semantic Web Challenge[4] in 2009 and which is well-known in the community. The raw size of the data is 247GB uncompressed and it contains 1.4B triples describing 114 million objects. This data-set was composed by combining crawls of multiple semantic search engines. Therefore, it does not necessarily match the coverage of any particular search engine. We refer the readers to `http://vmlion25.deri.ie/` for more information on the data set and to `http://km.aifb.kit.edu/ws/semsearch10` for the details of the semantic search challenge.

2.3 Real-World Web Queries

As the kinds of queries used by semantic search engines vary dramatically (ranging from structured SPARQL queries to searching directly for URI-based identifiers), it was decided to focus first on keyword-based search. Keyword-based search is the most commonly used query paradigm, is supported by most semantic search engines, and often serves as the foundation for more complex searches and processing.

Clearly, the type of result expected, and thus the way to assess relevance depend on the type of the query. For example, a query such as *plumbers in mason ohio* is looking for instances of a class of objects, while a query like *parcel 104 santa clara* is looking for information for one particular object, in this case a certain restaurant. [17] proposed a classification of queries by expected result type, and for our first evaluation we have decided to focus on object-queries, i.e. queries demonstrated by the latter example, where the user is seeking information about a particular object. Note that for this type of queries there might be other objects mentioned in the query other than the main object, such as *santa clara* in the above case. However, it is clear that the focus of the query is the restaurant named *parcel 104*, and not the city of Santa Clara as a whole.

Our evaluation required a set of object-queries that would be unbiased towards any existing semantic search engine. First, although the search engine logs of various semantic search engines were gathered, it was determined that the kinds of queries varied quite a lot, with many of the query logs of semantic search engines revealing idiosyncratic research tests by robots rather than real-world queries by actual users. Since one of the claims of semantic search is that it can help general purpose ad-hoc information retrieval on the Semantic Web, we have decided to use queries from actual users of hypertext Web search engines. As these queries would be from hypertext Web search engines, they would not be biased towards any semantic search engine. We had initial concerns if within the scope of the data-set it would be possible to provide relevant results for each of the queries. However, this possible weakness also doubled as a strength, as the testing of a real query sample from actual users would determine whether or not a billion triples from the Semantic Web realistically could help answer the real-world information needs of ordinary users, as opposed to the research-driven queries in most semantic search query logs [10].

We used a sample of the publicly available *Yahoo! Search Query Log Tiny Sample v1.0*, released by Yahoo! as a part of their WebScope program[5], which contains 4,500 queries sampled from the company's United States query log from January, 2009. One constraint of this data-set is that it contains only queries that have been posed by at least three different (not necessarily authenticated) users, which removes some of the heterogeneity of the log, for example in terms of spelling mistakes. We expected a random sample of these queries to be realistic but difficult to satisfy. Given the well-known differences between the top of the power-law distribution of queries and the long-tail, we used an additional log of queries from the Microsoft Live Search containing queries that were repeated by at least 10 different users. We expected these queries to be easier to answer.

overeaters ananymous
imdb batman returns
aloha sol
the longest yard sale
sacred heart u
sagemont church houston tx
david suchet
NAACP Image Awards
mr rourke fantasy island
old winchester shotguns
la scala restaurant philadelphia
the quick lift

Table 1: Example queries from the Yahoo! log.

We have selected a sample of 42 entity-queries from the Yahoo! query log by classifying queries manually, which filtered for ambiguous queries. A sample of these queries are given in Table 1. We have selected a sample of 50 queries from the Microsoft log. In this case we have pre-filtered queries automatically with a named entity recognizer, a gazetteer and rule-based named-entity recognizer that has shown to have very high precision in competitions, but these queries were not filtered for ambiguity. Both sets were combined into a single, alphabetically ordered list, so that participants were not aware which queries belonged to which set, or in fact that there were two sources of queries. We distributed the final set of 92 queries to the participants two weeks before the submission deadline.

[4] `http://challenge.semanticweb.org`

[5] `http://webscope.sandbox.yahoo.com/`

3. CROWDSOURCING JUDGMENTS

In this Section, we report how we used Amazon Mechanical Turk to assess the relevance of search results and describe the different sets of assessments we obtained for the evaluation. Using Mechanical Turk, tasks - called Human Intelligence Tasks (HITS) - are presented to a pool of human judges known as 'workers' who do the task in return for very small payments. Amazon provides a web-based interface for the workers that keeps track of their decisions and their payments. Because *anyone* can sign up to be a worker, we had to present each result for judgement in a way comprehensible to non-expert human judges. It was not an option to present the data in the native syntactic format of RDF such as RDF/XML or N-Triples, because they are too complex for average users, especially with the use of URIs as opposed to natural language terms for identifiers in RDF. In practice, semantic search systems use widely varying presentations of search results, sometimes tailored to particular domains. However, the rendering of results could possibly affect the valuation given by a judge. Allowing each participant to provide their own rendering would make it difficult to separate the measurement of ranking performance from effects of presentation, and would also eliminate the ability to pool results which reduces the total number of judgments needed.

For the purpose of evaluation, we have created a rendering algorithm to present the results in a concise, yet human-readable manner without domain-dependent customizations (see Figure 1). First, for each subject URI, all properties and objects were retrieved. Then the last rightmost hierarchical component of the property URI was used as the label of the property after tokenization. For example, the property `http://www.w3.org/1999/02/22-rdf-syntax-ns/type` was presented to the judge simply as `type`. A maximum of twelve object properties were displayed to the judge, with a preference being given to a few well-known property types defined in the RDF and RDF Schema namespaces, followed by custom-defined properties presented in the order retrieved from the data-set. In order to keep the amount of information given constant across judges and facilitate timely completion of the task, the URIs were not clickable and the judges were instructed to assess using only the information rendered, as to make the task of ad-hoc object retrieval directly comparable to tasks such as ad-hoc document retrieval. During the evaluation, we encountered the problem that some of the retrieved URIs only appear as objects, resulting in an empty display. Of the 6,158 URIs, a small minority of URIs (372) had triples only in the object position. For the current evaluation, we have ignored these results. Workers were given three options to judge each result: "Excellent - describes the query target specifically and exclusively", "Not bad - mostly about the target", and "Poor - not about the target, or mentions it only in passing." Note that we used the human-friendly labels "Excellent", "Not bad" and "Poor" for relevant, somewhat relevant and irrelevant results. We did not provide instructions to emphasize any particular properties (such as the "categories" in Figure 1), leaving the judgment to be based on general purpose judgment combining background knowledge about the entities and all of the displayed information.

In order to ensure quality in the presence of possible low-quality workers, each HIT consisted of 12 query-result pairs for relevance judgments. Of the 12 results, 10 were real results drawn from the participants' submissions, and 2 were gold-standard results randomly placed in the list of results. These gold-standard results were results from queries distinct from those used by the workers and have been manually judged earlier by an expert in RDF and information retrieval as being obviously 'relevant' or 'irrelevant'. For each HIT, there was both a gold-standard relevant and gold-standard irrelevant result included. These gold-standard results enabled the detection of workers who were not properly doing their task, as can be done by monitoring the average performance of judges on the gold-standard results hidden in their HITs. It is a common occurrence when using paid crowdsourcing systems for bogus workers to try to 'game' the system in order to gain money quickly without investing effort in the task, either by using automated bots or simply answering uniformly or randomly. Note that while we chose our gold-standards manually since we were evaluating a new task, one could in future campaigns use result with high inter-annotator agreement as new gold standards. Amazon Mechanical Turk allows payment to be withheld at the discretion of the creator of the HIT if they believe the task has not been done properly.

Before publishing the final tasks, we had done small-scale experiments with varying rewards for the workers. Mason and Watts have already determined previously that increased financial incentives increase the quantity, but not the quality, of work performed by participants [14]. Thus our approach was to lower the payment to workers down to the price where the speed of picking up the published tasks was still acceptable. When our results were published via Amazon Mechanical Turk, workers were paid $0.20 per HIT. In the first experiment reported here 65 workers in total participated in judging a total of 579 HITs or 1737 assignments (3 assignments per HIT), covering 5786 submitted results and 1158 gold-standard checks. (Note that of these only a subset of 4209 results and 842 checks is relevant here, being those which were also evaluated in MT2 and EXP, see below). Three workers were detected to be answering uniformly or randomly, and their work (a total of 95 assignments) was rejected and their assignments returned to the pool for another worker to complete. Two minutes were allotted for completing each HIT. On average the HITs were completed in 1 minute, with only two complaints that the allotted time was too short. This means that workers could earn $6-$12 an hour by participating in the evaluation. The entire competition was judged within 2 days, for a total cost of $347.16. We consider this both fast and cost-effective.

To study repeatability of our evaluation campaign we have re-evaluated the relevance of the search results returned by our test systems using a second set of workers. This second experiment has been performed six months after the initial evaluation using the exact same procedure. In the following, we will refer to the original set of assessments as MT1 and the repeated set of assessments as MT2.For MT1 there were 64 judges in total. The top four judges did 131 HITs and did not differ from the experts on the gold-standard items, with the overall percentage of mistakes over the 2176 gold-standard items in those 1088 HITs was 3.2%. For MT2 there were 69 judges in total. The top five judges did 165 HITs and did not differ at all from experts on the gold-standard items, and the overall percentage of mistakes with regards the 1662 gold-standard items in those 831 HITs was 4.5%. For future

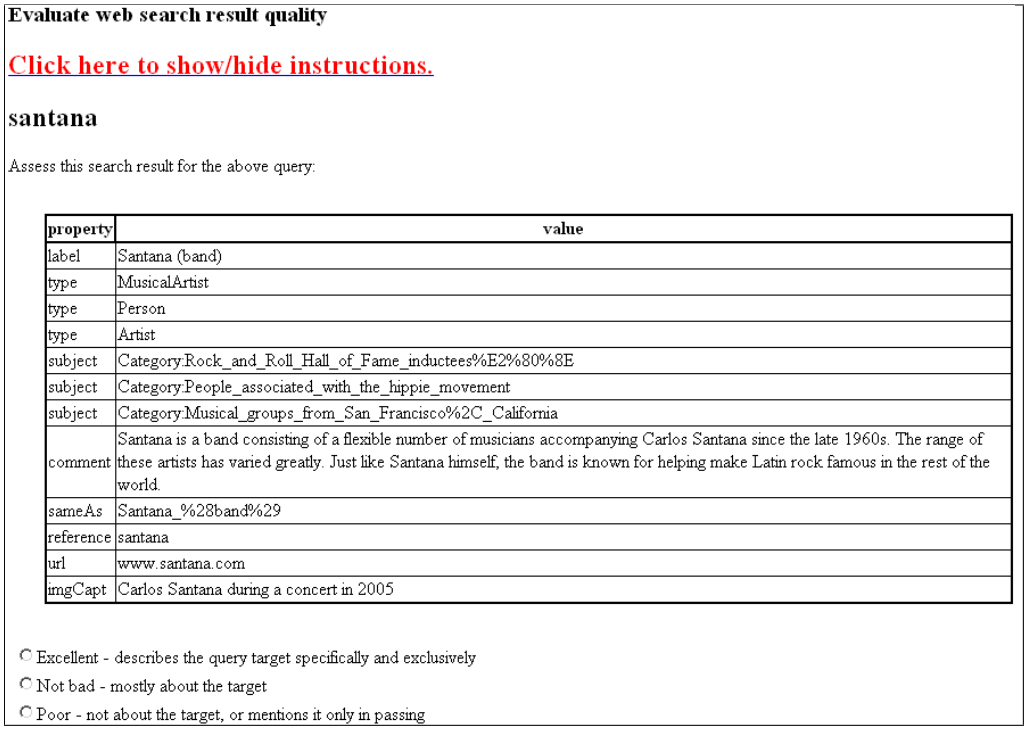

Figure 1: A sample HIT for semantic search evaluation.

campaigns items with a high inter-annotator reliability could be used to chose more gold-standard items.

To study the reliability of our crowdsourced judgments, we also created an "expert" set of relevance judgments over standard HITs that were not gold-standard items. Unlike repeatability, reliability concerns the ability of Mechanical Turk to reproduce a ground truth provided by experts. In our case, the authors of this paper have provided the ground truth by re-evaluating the same subset used in MT2. As this is a significant effort, we have used only one judge per HIT for re-evaluating the entire set of 4209 results, in 421 HITs of 10 results (leaving out the known-good and known-bad gold-standard check items). The resulting dataset is referred to as EXP herein.

For all of MT1, MT2, and EXP, we report here on the exact same set of queries and results. Some participants submitted more than one set of results (outputs from their system in differing configurations), of which we used the best submission of each of the competitor systems for testing repeatability. In total there were 6 competing systems with one submission each. Each result of every submission was judged by 3 crowdsourced workers, with systems results being judged to a depth of 10, given that it was a new unstudied task. We broke ties by taking the majority vote, except where the three judges each gave a different judgment, in which case we chose the middle, "Not Bad" assessment. In EXP, as mentioned above, each result was judged by a single expert, but a subset of 30 results were judged by three experts to determine intra-expert reliability.

Although the procedure for MT2 was the same as for MT1, the intervening six months appear to have seen a significant change in the worker pool: monitoring worker time-to-complete and performance on the known-good and known-bad gold-standard results revealed a total of 14 bogus workers for MT2, who completed a total of 1471 assignments between them before they were detected and blocked and their assignments returned to the pool. This change from 5% of assignments rejected in MT1 to 54% of assignments rejected in MT2 may indicate a significant increase in the number of bogus workers, and underlines the importance of including known-good and known-bad data in every HIT.

4. ANALYSIS OF RESULTS

We seek to answer the following in our experiments:

- **Repeatability** Are judges really inter-changeable?
  - Can we expect anonymous crowdsourced workers to agree on judgments?
  - Can we expect repeated experiments to produce the same results in terms of relevance metrics and the rank-order of the evaluated systems?

This requires also confirming previous results [2]:

- **Reliability** Can crowdsourced workers reliably reproduce the results we would have obtained if we were using expert judges?
  - Are the same items scored similarly by workers and experts?
  - Can worker evaluations produce the same results in terms of our relevance metrics and the rank-order of the evaluated systems?

We will use as parameters both the evaluation metric, the number of assessors per item and the relevance scale used. In particular, we would like to find out the following:

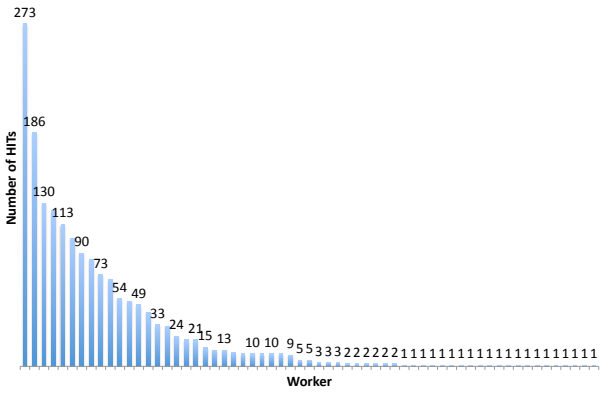

Figure 2: Workers ordered by decreasing number of items assessed.

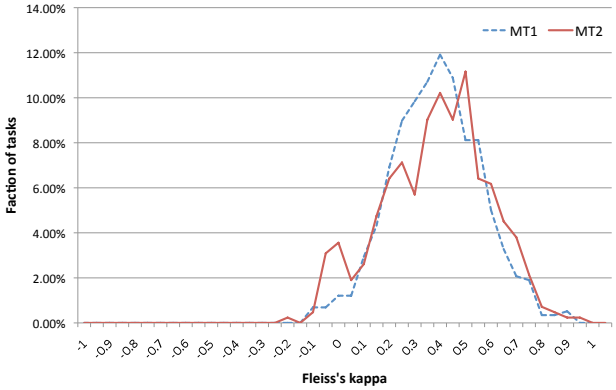

Figure 3: Agreement between workers.

- Which of our three evaluation metrics (MAP, NDCG, P@10) are more robust to changing the pool of workers, and when replacing experts with workers?
- Do we obtain better results with increasing number of assessments per item?
- Do our results hold for both binary and ternary scale assessment?

4.1 Repeatability

As previously discussed, in IR evaluation the notion of repeatability is tied to measuring the extent to which judges are inter-changeable. The argumentation goes that if we show that judges from a particular pool of assessors are inter-changeable, the experiment can be repeated with any subset of judges from the pool: the judges will agree on the relevancy of items to be judged, which will be reflected in the metrics to be computed, and the eventual ranking of the competing systems.

The most common measures of inter-annotator agreement in IR evaluations are Cohen's κ for the case of two judges, and Fleiss's κ for the case of multiple judges, which has a free-marginal version [15]. While we report inter-annotator agreement, we note that the applicability of standard metrics to the case of crowdsourced workers can be questioned. The reason is that although we have a fixed number of workers for each HIT, in the crowdsourcing scenario the workers select the tasks, and thus they are not necessarily the *same* workers who assess each item. Figure 2 shows the number of items judged by each worker in our first experiment with Mechanical Turk. In the case of traditional expert-based evaluation, this distribution would be flat as each expert would assess the same items. In our case, each worker may assess a different number of the total set of HITs. Some workers assess a large number of HITs, with the most diligent worker going through 273 HITs, while a long tail of workers worked on a single task only. This long tail is especially problematic since there is much less data about these workers on which to base reliability tests.

Based on our knowledge of the related work, it seems that there is not yet consensus as to how to account for this deficiency [6] and the question of reliability is sometimes ignored altogether [8]. We believe the most prudent way to proceed is to report the distribution of Fleiss' κ values considering

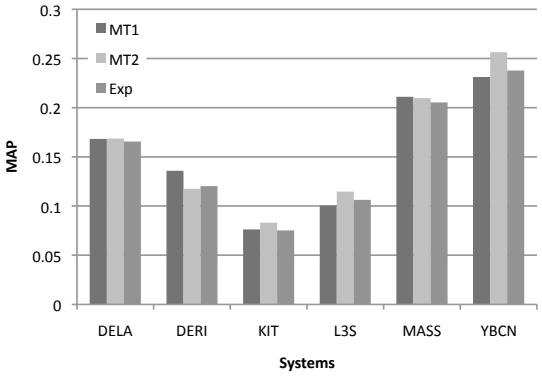

Figure 4: Mean average precision (MAP) for the systems using different test sets.

all HITs as individual assessments of a small number of 12 items. In Figure 3 we show this distribution for our first and second experiment. As the Figure shows, the level of agreement is very similar. The average and standard deviation are 0.36±0.18 for the first experiment (MT1) versus 0.36±0.21 for MT2. In fact, the difference between the average agreement appears at the fourth digit, strongly supporting the idea of a homogeneous pool of workers. We achieve slightly higher levels of agreement for binary relevance (with somewhat relevant and relevant judgments counted both as relevant), 0.44±0.22 and 0.47±0.25. There is thus no marked difference between a three-point scale and a binary scale, meaning that it was feasible to judge this task on a three-point scale.

Agreement numbers are not easy to interpret even in the context of related work, and agreement is only a proxy for a repeatable evaluation: what we are ultimately after is whether different pools of workers used in different experiments lead to the same results in terms of evaluation metrics, and ultimately the same ordering of the evaluated systems. Figure 4 shows Mean Average Precision (MAP) scores for the different systems using the two different evaluation sets obtained via Mechanical Turk (MT1 and MT2). The results are also included in Table 3. We can see that the scores are close in value, and in fact there is no change to the rank-order of the systems. The result holds for both binary and

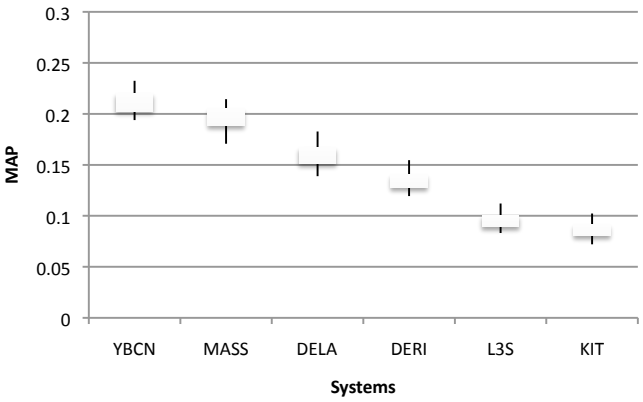

Figure 5: Mean average precision (MAP) for the systems using different test sets and a single worker.

ternary scale, and for both MAP, P@10 and NDCG. Broadly, this confirms our hypothesis that crowdsourced ad-hoc evaluation is repeatable. The relative change in scores across the two sets, for all systems in average, is 7.85% for MAP, 4.24% for NDCG and 6.87% for P@10. This gives us a first indication that two systems would need to be very close in performance in order to change places in the ranking produced by repeated experiments.

In fact, Mechanical Turk gives surprisingly robust results with just a single assessment per item. We have tested this by subsampling, i.e. selecting randomly a single assessment for each item from the six assessments we have collected in total. We have repeated this a 100 times and computed the min, max, mean and standard deviation of our metrics. Figure 5 shows the min, max, and the range of one standard deviation from the mean for each system, using MAP as the metric. This figure furthermore shows that even one standard deviation intervals provide different ranges for the different systems and effectively separate them. Though the score of a system in a particular sample may surpass the score of an overall inferior system, such cases would be rare. Note that there is a particular robustness to Mechanical Turk. Though conventional wisdom would certainly be against running an evaluation with a possibly unreliable single judge, in the case of crowdsourcing the assessments will come from not a single expert judge for all the results, but multiple workers. These workers may be individually unreliable, but each will judge a small number of items. When considering three judges, see Figure 6, the intervals around the mean get even tighter.

The decrease of standard deviation around the mean is also shown in Figure 7. This Figure shows the standard deviation on the y-axis, for different numbers of workers (x-axis), and using different metrics. We see that P@10 benefits the most from increasing the number of workers and that adding more workers decreases the standard deviation between workers.

4.2 Reliability

Repeatable evaluations require that each evaluation be reliable, and while work such as Alonso et al. [2] has shown that crowdsourced judges can be reliable in information retrieval tasks, we should show that this reliability holds over

Set	Total items	Irrelevant	Somewhat R.	Relevant
MT1	4209	2593	970	646
MT2	4209	2497	975	737
EXP	4209	2847	640	722

Table 2: Scoring patterns in different evaluation sets.

repeated experiments. We measured the agreement between expert judges on a subset of the items (30 HITs). In this case, the average and standard deviation of Fleiss's κ for the two- and three-point scales are 0.57±0.18 and 0.56±0.16, respectively. The level of agreement is thus higher for expert judges, with comparable deviation. For expert judges, there is practically no difference between the two- and three-point scales, meaning that expert judges had much less trouble using the middle judgment.

The most basic statistic we can look at is the difference in scoring patterns of experts and non-experts. Moving on to comparing expert reliability with crowdsourced judgements from MT1 and MT2, Table 2 shows that again different sets of workers behave very similarly, though different from the experts on the whole. Fleiss's κ is similar with 0.412 between MT1 and experts, and 0.417 between MT2 and experts. In particular, experts are more pessimistic in their scoring, marking irrelevant many of the items that the workers would consider somewhat relevant.

This effect is also visible in Figure 8, which shows the assessments of the two worker sets compared to the assessments of the experts for the three assessment options. Whereas the two worker sets display similar behaviour compared to each to other, the difference towards more positive assessments compared to the experts can be observed. This may suggest that crowdsourced judgments cannot replace expert evaluations. Based on comments and the data, the source of this effect is likely the fact that experts understood "describes the query target specifically and exclusively" to be much of a more sharp distinction about objects than workers. An expert would note that the IMDB article about a movie featuring actor David Suchet would not be considered 'relevant', while workers would often judge that result as relevant if the query asked for David Suchet.

Looking at agreement rate in other settings, such a κ of 0.55 at TREC 2005 on sentence relevance at TREC 2004 Novelty Trac [18], our experts are clearly reliable, with agreement ratings of 0.57 (binary scale) and 0.56 (ternary scale). Yet then the reliability of non-expert crowdsourced judges of 0.36 in our experiment then appears to be less than ideal. However, does it change the ranking of the systems? This would be the ideal test of how far reliability has to degrade in order to impact an evaluation campaign.

Even if the level of agreement is higher amongst expert judges, if the ranking of the systems does not change when non-experts are employed, then a crowdsourcing approach is still reliable enough for the task (even if their reliability is strictly speaking relatively lower than expert judges). The relative change in scores when going from experts to workers (moving from EXP to three-samples of MT1 and MT2), for all systems in average, and using three judgments, is 1.8% for MAP, 3.5% for NDCG and 12.8% for P@10 (see also Table 3). These are comparable changes to what we have seen when moving from one worker set to another, but

	MAP			NDCG			P@10		
System	MT1	MT2	EXP	MT1	MT2	EXP	MT1	MT2	EXP
YBCN	0.23	0.26	0.24	0.35	0.38	0.33	0.48	0.54	0.45
MASS	0.21	0.21	0.21	0.34	0.34	0.33	0.48	0.51	0.40
DELA	0.17	0.17	0.17	0.29	0.27	0.28	0.41	0.43	0.35
DERI	0.14	0.12	0.12	0.24	0.24	0.22	0.39	0.36	0.30
L3S	0.10	0.11	0.11	0.20	0.21	0.20	0.28	0.30	0.24
KIT	0.08	0.08	0.08	0.15	0.14	0.15	0.26	0.28	0.23

Table 3: Evaluation results using different evaluation sets and metrics.

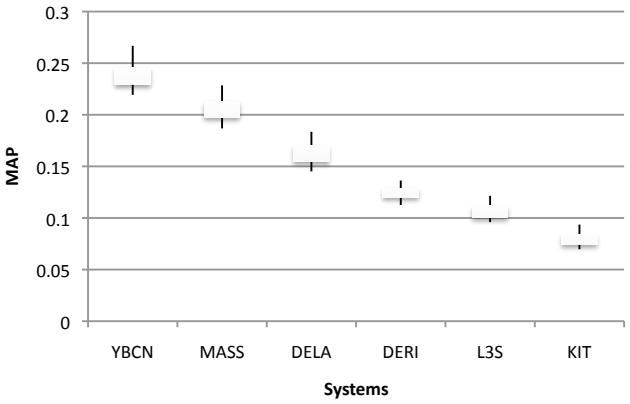

Figure 6: Mean average precision (MAP) for the systems using different test sets and three workers.

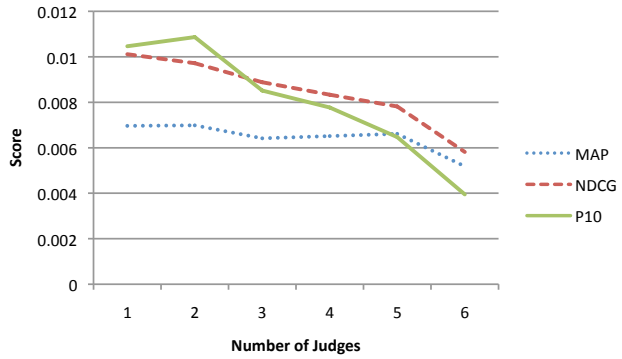

Figure 7: Average standard deviation around the mean for different numbers of workers and using different metrics.

the changes are mostly positive, with notable increases in P@10 when changing from experts to workers. In particular, the increase in somewhat relevant scores explains the increase of the binary P@10 measure. Somewhat relevant results (counted as relevant for the binary measures) that are coming in at lower ranks boost P@10 more than MAP and NDCG, which are less sensitive to changes in the lower ranks. While the reliability of non-expert judges is lower than expert judges, the reliability of non-expert judges is still sufficient for ranking systems in the evaluation.

Figure 4 illustrates visually the performance values for MAP for the different systems using the two MT evaluation sets and the expert judgments. The values are not only close, but in fact again the obtained values for the experts produce the same rank-order of the systems as with any of the MT evaluation sets.

As in the case of repeatability, we might ask whether crowdsourced assessments become more reliable when adding more judges. We have already shown in Figure 7 that increasing the number of workers decreases their standard deviation and increases the reliability of workers, and this trend seems to continue beyond 6 workers. Figure 9 shows the deviation resulting from using the workers' assessments instead of the expert assessments, in particular the average relative change in our metrics for subsamples, for different numbers of workers. We can see a clear benefit to using three workers instead of 1 or 2 workers, but there is comparatively less benefit from employing more than three judges. Figure 10 shows the same for MAP and NDCG using the average values of Kendall's τ between the subsamples of worker judgments and the expert assessments. This value of τ is already very close to one for three judges independent of the metric. While intra-worker reliability increases as the number of workers increase, adding more than three workers will lead to a higher number of disagreements with expert judges.

5. RELATED WORK

The main difference in using crowdsourcing to "gold standard" evaluation data-set creation in campaigns like TREC [4] is that human judges are no longer a relatively small group of professional expert judges who complete an equal-sized number of assessments, but large group of non-experts who may complete vastly differing numbers of assessments and may not actually have the required skill-set (such as command of English) to complete the task or be completing the task honestly. Earlier work in using crowdsourcing for information retrieval demonstrated quick turn-around times and the ability to have a much higher number of judges than previously thought possible [1]. This has led to a rapidly-expanding number of applications of crowdsourcing evaluation data sets to a wide range of information retrieval tasks such as XML-based retrieval [2]. Crowdsourcing has also been expanded successfully to related areas, such as machine translation [5].

In this vein, our primary contribution is in demonstrating the repeatability crowdsourcing judgments in creating evaluation data sets, even when entirely different sets of judges are used on the same task over long periods of time, a necessary feature for running large-scale campaigns for novel information retrieval tasks on an annual basis. Previous work

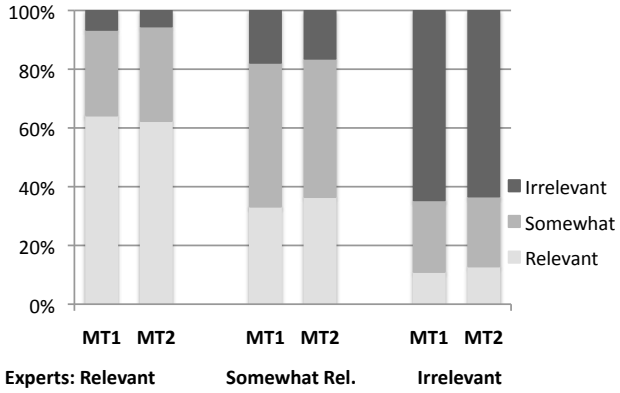

Figure 8: Assessments of the two workers sets compared to the experts' assessments for the three assessment options.

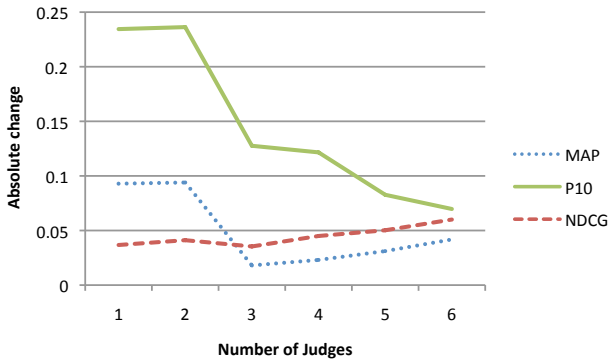

Figure 9: Average deviation of sample means from the expert assessments.

on crowdsourcing evaluation campaigns, such as work on replicating image labelling in ImageCLEF[15], has focused on determining the reliability of the judges over small subsets of the original campaign, but has not tested whether the evaluation campaign is repeatable over large time intervals (i.e., months or years), only inspecting differences over small amounts of time (4 days) and not comparing the judges performance over time to each other, but aggregating all judgments.

Previous work [1, 15] in general has focused on comparing crowdsourcing judgments to that of experts on existing campaigns with well-known "gold standards," not bootstrapping new evaluation campaigns for new search tasks where there are multiple competing but unevaluated search systems, such as in semantic search. Another goal of our work is to demonstrate the use of crowdsourcing for a large-scale evaluation campaign for a novel search task, which in our case is ad-hoc object retrieval over RDF. Many semantic search systems of this type, such as [9, 16, 19], have appeared in the past few years, but none have been evaluated against each other except on a very small scale. Semantic search systems are a subset of information retrieval systems, and thus it would be natural to apply existing IR benchmarks for their evaluation in a large-scale campaign.

There are two difficulties in applying the ad-hoc document retrieval methodology directly to our object retrieval problem as identified in [17]. The first and most apparent problem is that not all semantic search engines perform document retrieval, but rather retrieve knowledge that is already encoded in RDF, where factual answers may be found by aggregating or linking knowledge across RDF data. This is a clear difference to 'entity search' tracks such as the TREC Entity Track [4] or the INEX Entity Ranking Track [12]. With respect to addressing keyword retrieval on structured data, there is also existing work in the database literature (e.g., [13]), but this field of research has not produced a common evaluation methodology that we could have adapted. Second, in semantic search the unit of retrieval and thus the way to evaluate the results is dependent on the type of query. In turn, the types of queries supported may vary from search engine to search engine. By reducing the broad problem of semantic search to that of keyword-based ad-hoc object retrieval (i.e. retrieving objects given in RDF with relevant factual assertions connected as a property by a single link), we could invite multiple systems to our campaign, as most semantic search systems have this base-line feature. More complex query and result processing relies upon first retrieving a baseline of relevant RDF graphs, and so this baseline should be evaluated first.

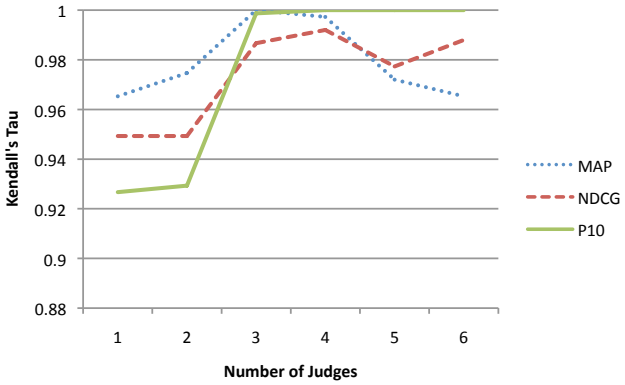

Figure 10: Kendall's Tau between workers and experts for different number of assessments per item.

6. CONCLUSIONS

With the advent of crowdsourcing platforms like Amazon Mechanical Turk, creating a "gold standard" evaluation data set of relevance judgments for new kinds of search tasks is now cheap, scalable, and easy to deploy. We have shown how to quickly boot-strap a repeatable evaluation campaign for a search task that has not previously been systematically evaluated, such as the object information retrieval task in semantic search, using Mechanical Turk. However, are such crowdsourced evaluation campaigns trustworthy? Are the relevance judgments of crowdsourced judges both reliable compared to experts and can such judgments be repeated with entirely different crowdsourced judges over time?

Regarding the **repeatability** of such crowdsourced judgments, we have shown that the level of agreement is the same for two pools of crowdsourced judges even when the evaluation is repeated after six months. Repeating an evaluation using crowdsourcing after six months led to the same result in evaluation metrics and the rank-order of the systems being unchanged. As regards the **reliability** of the

crowdsourced judgments, while there were differences between expert judgments and crowdsourced judgments, with experts in general rating more results negative than crowdsourced judges. This is likely due to the object retrieval task and the time pressure on workers, as expert were more adept at discriminating between queries exclusively about an object to ones simply mentioning an object given time limits. However, the rank ordering of systems does not change when moving from experts to crowdsourced workers. Three judges seems to be a sufficient number and, surprisingly, increasing the number of crowdsourced judges has little effect unless the systems are particularly close. As regards evaluation metrics, *P@10* is more brittle than measures such as *MAP* and *nDCG* and so benefits most from collecting additional judgments.

We have successfully shown how a number of real-world and research semantic search systems can be evaluated in a repeatable and reliable manner via creating a new evaluation campaign using crowdsourcing. While the study here as focused on agreement between judges and workers over time and holding the items (queries and results) constant, future research needs to study the agreement between judges and workers on a per-item basis. So, for example, how does the ambiguity of entity queries effect reliability and repeatability? So the next study should also take into account if these results hold over different kinds of items, so the "Semantic Search" evaluation campaign will be broadened to deal with new kinds of semantic search tasks featuring different keyword queries and more expressive and complex queries beyond keywords. Of course, the methodology demonstrated in this work should be repeated for these new tasks if the task goes beyond object retrieval. Crowdsourced evaluation can lead to new tasks being evaluated quickly with reliable and repeatable evaluations. It also aids in having much larger corpora and query workloads for these campaigns. Most importantly, as the crowdsourced results are reliable and repeatable for this task at any time, evaluation campaigns can now run *continuously* (by using a standard community-driven evaluation web service) rather than annually. Our results support fast and scalable "just-in-time" evaluation of new search tasks, with empirically demonstrated repeatability and reliability.

7. ADDITIONAL AUTHORS

Thanh Tran Duc, Institute AIFB, Karlsruhe Institute of Technology, 76128 Karlsruhe, Germany ducthanh.tran@kit.edu

8. REFERENCES

[1] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[2] O. Alonso, R. Schenkel, and M. Theobald. Crowdsourcing assessments for XML ranked retrieval. In *ECIR*, pages 602–606, 2010.

[3] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *SIGIR*, pages 667–674, New York, NY, USA, 2008. ACM.

[4] K. Balog, A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the TREC 2009 Entity Track. In *NIST Special Publication: SP 500-278*, 2009.

[5] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore, August 2009. Association for Computational Linguistics.

[6] B. Carpenter. Multilevel bayesian models of categorical data annotation. technical report. Technical report, Alias-I, 2008. http://lingpipe.files.wordpress.com/2008/11/carp-bayesian-multilevel-annotation.pdf.

[7] L. Ding, T. Finin, A. Joshi, R. Pan, S. R. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: a search and metadata engine for the semantic web. In *CIKM*, pages 652–659, 2004.

[8] S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum. Language-model-based Ranking for Queries on RDF-graphs. In *CIKM*, pages 977–986. ACM, 2009.

[9] R. Guha, R. McCool, and E. Miller. Semantic Search. In *WWW*, pages 700–709. ACM, 2003.

[10] H. Halpin. A query-driven characterization of linked data. In *Proceedings of the WWW Workshop on Linked Data on the Web*, Madrid, Spain, 2009.

[11] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *J. Am. Soc. Inf. Sci.*, 47(1):37–49, 1996.

[12] J. Kamps, S. Geva, A. Trotman, A. Woodley, and M. Koolen. Overview of the INEX 2008 Ad Hoc Track. *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008*, pages 1–28, 2009.

[13] Y. Luo, W. Wang, and X. Lin. SPARK: A Keyword Search Engine on Relational Databases. In *ICDE*, pages 1552–1555, 2008.

[14] W. Mason and D. J. Watts. Financial Incentives and the "Performance of Crowds". In *Human Computation Workshop (HComp2009)*, 2009.

[15] S. Nowak and S. M. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Multimedia Information Retrieval*, pages 557–566, 2010.

[16] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: a document-oriented lookup index for open linked data. *IJMSO*, 3(1):37–52, 2008.

[17] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc Object Ranking in the Web of Data . In *Proceedings of the WWW*, pages 771–780, Raleigh, USA, 2010.

[18] I. Soboroff and D. Harman. Novelty detection: the trec experience. In *HLT '05*, USA, 2005. ACL.

[19] T. Tran, H. Wang, and P. Haase. SearchWebDB: Data Web Search on a Pay-As-You-Go Integration Infrastructure, 2008.

[20] E. Voorhees. The philosophy of information retrieval evaluation. In *In Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370. Springer-Verlag, 2001.