Million Query Track 2009 Overview

Ben Carterette[*], Virgil Pavlu[†], Hui Fang[‡], Evangelos Kanoulas[§]

The Million Query Track ran for the third time in 2009. The track is designed to serve two purposes: first, it is an exploration of ad hoc retrieval over a large set of queries and a large collection of documents; second, it investigates questions of system evaluation, in particular whether it is better to evaluate using many queries judged shallowly or fewer queries judged thoroughly.

Fundamentally, the Million Query tracks (2007-2009) are ad-hoc tasks, only using complex but very efficient evaluation methodologies that allow human assessment effort to be spread on up to 20 times more queries than previous ad-hoc tasks. We estimate metrics like Average Precision fairly well and produce system ranking that (with high confidence) match the true ranking that would be obtained with complete judgments. We can answer budget related questions like how many queries versus how many assessments per query give an optimal strategy; a variance analysis is possible due to the large number of queries involved.

While we have confidence we can evaluate participating runs well, an important question is whether the assessments produced by the evaluation process can be reused (together with the collection and the topics) for a new search strategy—that is, one that did not participate in the assessment done by NIST. To answer this, we designed a reusability study which concludes that a variant of participating track systems may be evaluated with reasonably high confidence using the MQ data, while a complete new system cannot.

The 2009 track quadrupled the number of queries of previous years from 10,000 to 40,000. In addition, this year saw the introduction of a number of new threads to the basic framework established in the 2007 and 2008 tracks:

- Queries were classified by the task they represented as well as by their apparent difficulty.
- Participating sites could choose to do increasing numbers of queries, depending on time and resources available to them.
- We designed and implemented a novel *in situ* reusability study.

Section 1 describes the tasks for participants. Section 2 provides an overview of the test collection that will result from the track. Section 3 briefly describes the document selection and evaluation methods. Section 4 summarizes the submitted runs. In Section 5 we summarize evaluation results from the task, and Section 6 provides deeper analysis into the results.

For TREC 2009, Million Query, Relevance Feedback, and Web track ad-hoc task judging was conducted simultaneously using MQ track methods. A number of compromises had to be made to accomplish this; a note about the usability of the resulting data is included in Section 3.3.

1 Task Description

The basic task is ad hoc retrieval over a large set of queries. As with the Terabyte track before it, the challenge is running a very large number of queries over a very large index and still achieving good results. This aspect of the track has not changed since 2007.

[*] Department of Computer & Information Sciences, University of Delaware, Newark, DE, USA

[†] College of Computer and Information Science, Northeastern University, Boston, MA, USA

[‡] Department of Computer & Electrical Engineering, University of Delaware, Newark, DE, USA

[§] Information Studies Department, University of Sheffield, Sheffield, UK

This year we added two additional (optional) tasks. The first was prediction of *type* of intent behind a query. A large sample of queries represents a wide space of tasks and information needs; treating them all as traditional deep informational search tasks does not accurately reflect the needs of users working with a system. Thus we allowed participants to try to predict whether a query was more "precision-oriented" (i.e. looking for a small, well-contained set of facts) or more "recall-oriented" (i.e. looking for deeper, more open-ended information). Based on these predictions, participants could choose to select an alternate retrieval algorithm. In addition to intent-type predictions, we allowed participants to try to predict whether a query would be easy for their system or hard, and likewise to try alternate methods depending on the prediction.

The second was query prioritization. We assigned each of the 40,000 queries a priority level, and a site could choose to do only the highest-priority queries (of which there were 1,000), the 1st and 2nd highest priority (4,000), 1st, 2nd, and 3rd highest (10,000), or all queries (40,000). This was a concession to the fact that most sites had not previously indexed such a large corpus and were concerned about being able to run such a large number of queries.

2 Test Collection

2.1 Corpus

This year we adopted the new ClueWeb09 crawl of the general web. The full collection contains over one billion web pages and other documents in 10 different languages; since this was the first year using such a large corpus, we elected to use the smaller "Category B" set of 50 million English pages in roughly 2TB of disk space. More information about the corpus can be found at `http://boston.lti.cs.cmu.edu/Data/clueweb09/`.

2.2 Queries

Queries were sampled from two large query logs. To help anonymize and select queries with relatively high volume, they were processed by a filter that converted them into queries with roughly equal frequency in a third query log.

Assessors selected short keyword queries from lists of 10 randomly selected from the set of 40,000. The random selection was biased to ensure that more high-priority queries would be presented. After selecting a query, assessors "backfit" it to a TREC topic by writing a description and narrative. In all, 634 topics were developed for the track; since Web/RF topics 1–50 were included in judging, the final total is 684 topics.

Query intent types were assigned by assessors when backfitting their queries. Exactly one of the following six categories, based on work by Rose and Levinson [RL04], was chosen for each query; we then mapped the six categories into broader "precision-oriented" or "recall-oriented" classes:

- **Navigational** (precision): Find a specific URL or web page.
- **Closed** (precision) or directed information need: Find a short, unambiguous answer to a specific question.
- **Resource** (precision): Locate a web-based resource or download.
- **Open** (recall) or undirected information need: Answer an open-ended question, or find all available information about a topic.
- **Advice** (recall): Find advice or ideas regarding a general question or problem.
- **List** (recall): Find a list of results that will help satisfy an open-ended goal.

As an example of the classification task, the query "district of columbia department of motor vehicles" (from the 2008 1MQ track) might be classified as "precision-oriented", as its intent seems navigational in nature.

The query “vietnam vets and agent orange” might be classified as “recall-oriented”, as it seems to reflect a broad, open-ended informational intent.

The queries fell into the broad precision/recall classes with 43% precision and 57% recall. Most of the recall types (indeed, most of the queries) were classified as ”open or undirected information needs”, with “advice” and “list” types making up just 6% of all recall queries. Precision types were more evenly distributed, with 44% “closed or directed information need”, 33% “navigational”, and 23% “resource”. On average, among the 34 judgments per query in the MQ-only set 7.82 were judged relevant. The corresponding numbers for the 50 first queries are 70 relevant documents per query out of the 250 judged documents.

Judged queries were also assessed for difficulty by the Average-Average-Precision (AAP) score, the average of average precision estimates for a single query over all submitted runs. These were assigned automatically by partitioning the AAP score range into three intervals:

- **Hard**: $AAP \in [0, 0.06)$
- **Medium**: $AAP \in [0.06, 0.17)$
- **Easy**: $AAP \in [0.17, max]$

These intervals were chosen so that queries would be roughly evenly distributed. 38% of all queries were hard, 32% medium, and 30% easy; the “hard” class includes more queries because it includes every query for which no relevant documents were found.

2.3 Relevance Judgments

Assessors were shown a full web page, with images included to the extent possible by referencing the original site. They were asked to judge a document either “not relevant”, “not relevant but reasonable”, “relevant”, or “highly relevant”. Each topic received either 32 or 64 judgments to documents ranked by submitted systems and selected by alternating between two methods described below.

Some queries were designated for reusability study (see Section 5.1 below). For these, three of the eight participating sites were chosen to be held out of judgment collection. Each site was held out a roughly equal number of times, so each site contributed to and was held out from roughly the same number of queries (unless the site did queries beyond the high-priority set). All reusability-designated queries received 32 judgments; all non-reusability queries received 64.

In all, 684 queries received judgments, including the 50 that overlapped with the Web and Relevance Feedback tracks. On average, each query received 50 judgments, though the first 50 received many more due to the judging requirements of those tracks. The 634 MQ-only queries received an average of 34 judgments each. 26% of all judgments were “relevant” or “highly relevant”, though that dropped to 23% among the MQ-only queries.

3 Selection and Evaluation Methods

For a full description of the two methods, we refer the reader to the 2007 Million Query Track overview and the original work cited below. The descriptions below focus on estimates of average precision and other measures.

3.1 MTC

The Minimal Test Collections (MTC) method works by identifying documents that will be most informative for understanding performance differences between systems by some evaluation measure (in this case average precision). Details on the workings of MTC can be found elsewhere [CPK$^+$08, CAS06, ACA$^+$07]. Here we focus on MTC estimates of evaluation measures.

First, we consider each document i to have a distribution of relevance $p(X_i)$. If the document has been given judgment $j = 0$ or 1 (nonrelevant or relevant), then $p(X_i = j) = 1$; otherwise the probability that the document is relevant is $p(X_i = 1) = p_i$. Then we consider a measure to be a random variable expressed as a function of document relevances X_i. For example, precision can be expressed as a random variable that is a sum of document relevance random variables for those documents retrieved at ranks 1 through k.

If the measure is a function of document random variables, then the measure has a distribution over possible assignments of relevance to unjudged documents. Note that this applies to measures averaged over queries as well as to measures for a single query. This produces a distribution over possible rankings of systems: there is some probability that system 1 is better than system 2, which in turn is better than system 3; some probability that system 1 is better than system 3, which in turn is better than system 2, and so on. It can be shown that the maximum a posteriori ranking of systems is that in which systems are ranked by the expected values of the evaluation measure of interest over the possible relevance assignments.

Calculating the expectation of an evaluation measure is fairly simple. Given the probability that document i is relevant $p_i = p(X_i = 1)$, we define:

$$\mathbf{E}prec@k = \frac{1}{k}\sum_{i=1}^{k} p_i$$

$$\mathbf{E}R\text{-}prec \approx \frac{1}{\mathbf{E}R}\sum_{i=1}^{\mathbf{E}R} p_i$$

$$\mathbf{E}AP \approx \frac{1}{\mathbf{E}R}\sum_{i=1}^{n} p_i/i + \sum_{j>i} p_i p_j/j$$

$$\text{and } \mathbf{E}R = \sum_{i=1}^{n} p_i.$$

Note that there is an assumption that document relevances are independent, or conditionally independent given features used to estimate the distribution $p(X_i)$.

Though MTC is designed for ranking systems (i.e. making decisions about relative differences in performance between systems), in this work we largely present expectations of evaluation measures for individual systems. Note that these expectations are *not* good estimates of actual values of evaluation measures; the most interesting quantities MTC provides are the probabilities of particular pairwise orderings of systems.

MTC evaluation is implemented in the `mtc-eval` package downloadable from the TREC web site or at `http://ir.cis.udel.edu/~carteret/downloads.html` (which will always have the latest version).

3.1.1 Estimating Probability of Relevance

The probability estimates p_i are the output of a model of the log odds of relevance expressed as a linear combination of feature values:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \sum_{j=1}^{v} \beta_j f_{ij}$$

where f_{ij} is the value of feature j for document i. Note that this is simply a logistic regression classifier.

We used two different sets of features to obtain relevance probabilities. One set, implemented by the `expert` utility in the `mtc-eval` distribution, consists of features extracted from the system rankings themselves. These are described in more detail by Carterette [Car07].

The other set, described by Carterette and Allan [CS07], uses similarities between documents as features: the value of f_{ij} is the cosine similarity between documents i and j. The “feature documents” are those that have been judged for relevance to the topic.

3.2 statAP

In statistical terms, average precision can be thought of as the mean of a population: the elements of the population are the relevant documents in the document collection and the population value of each element is the precision at this document for the list being evaluated. This principle is the base for several recently proposed evaluation techniques [YA06, APY06, AP, ACA+07]. StatAP is a sample-and-estimate technique defined by the following two choices.

Stratified Sampling, as developed by Stevens [BH83, Ste], is very straightforward for our application. Briefly, it consists of bucketing the documents ordered by a chosen prior distribution and then sampling in two stages: first sample buckets with replacement according to cumulative weight; then sample documents inside each bucket without replacement according to selection at the previous stage.

Generalized ratio estimator. Given a sample S of judged documents along with inclusion probabilities, in order to estimate average precision, *statAP* adapts the generalized ratio estimator for unequal probability designs [Tho92].(very popular on polls, election strategies, market research etc). For our problem, the population values are precisions at relevant ranks; so for a given query and a particular system determined by ranking r(.) we have (x_d denotes the relevance judgment of document d) :

$$statAP \quad = \quad \frac{1}{\widehat{R}} \sum_{d \in S} \frac{x_d \cdot \widehat{prec@r}(d)}{\pi_d}$$

where

$$\widehat{R} = \sum_{d \in S} \frac{x_d}{\pi_d} \; ; \qquad \widehat{prec@k} = \frac{1}{k} \sum_{d \in S, r(d) \leq k} \frac{x_d}{\pi_d}$$

are estimates the total number of relevant documents and precision at rank k, respectively, both using the Horwitz-Thompson unbiased estimator [Tho92].

Confidence intervals. We can compute the inclusion probability for each document (π_d) and also for pairs of documents (π_{df}); therefore we can calculate an estimate of variance, $\widehat{var}(statAP)$, from the sample, using the ratio estimator variance formula found in [Tho92], pp. 78 (see [AP, ACA+07] for details). Assuming the set of queries Q is chosen randomly and independently, and taking into account the weighting scheme we are using to compute the final MAP (see Results), we compute an estimator for the MAP variance

$$\widehat{var}(statMAP) = \frac{1}{(\sum_q w_q)^2} \sum_{q \in Q} w_q^2 \cdot \widehat{var}(statAP_q)$$

where w_q is distribution weight proportional with the number of judgments made on query q. Assuming normally distributed $statMAP$ values, a 95% confidence interval is given by $\pm 2std$ or $\pm 2\sqrt{\widehat{var}(statMAP)}$.

3.3 Note on TREC 2009 Judging

The MQ methods were also used to select documents for the Web track's ad hoc task and the Relevance Feedback track. Due to the varying interests of the tracks and the distinction between "Category A" runs that in principle could retrieve documents from the entire collection and "Category B" runs that were explicitly restricted to the smaller Category B subset of ClueWeb09, certain compromises had to be made in the application of these methods across tracks. The coordinators for these tracks, as a group, decided on the following approaches:

1. a depth-12 pool formed from all Category A runs would be judged for relevance to the 50 topics in common between all three tracks;

2. statAP and MTC would be used (jointly, alternating between the two for each topic) to select additional documents to judge from Category B runs for those 50 topics;

3. statAP and MTC would be used (jointly, alternating between the two for each topic) to select all documents to judge for all MQ track runs, which are all Category B runs by the track's guidelines.

International Institute of Information Technology (SIEL)	IRRA
Northeastern University	Sabir Research, Inc.
University of Delaware (Carterette)	University of Delaware (Fang)
University of Glasgow	University of Illinois at Urbana-Champaign

Table 1: Groups participating in the 2009 Million Query track

One result of these decisions is that statAP could not be used to evaluate any Category A run—it requires that sampling be done according to priors computed over all runs, which did not happen for Category A runs. A second result is that the estimates of document relevance required by MTC are different for the different groups of runs: the `erels.catA.1-50` file for Category A runs, the `erels.catB.1-50` file for Category B-only runs, and the `erels_docsim.20001-60000` file for the MQ runs. All three files can be downloaded from the MQ track site at `http://ir.cis.udel.edu/million/data.html`; this page also gives instructions on how to replicate and reuse data from the Web and MQ tracks.

4 Submitted Runs

The track received 35 runs from the 8 groups listed in Table 4. Every group was allowed to submit up to five runs. All of the submitted runs were included in judgment collection for a roughly equal number of queries. A new task introduced in this year's track is to classify queries based on underlying user intentions. Three groups, Sabir, NEU, and UIUC, submitted query class predictions.

The following is a brief summary of some of the submitted runs based on the information provided by the participating groups.

IIIT-Hyderabad submitted three runs: *iiithAuEQ*, *iiithAuthPN*, and *iiithExpQry*. The basic system was developed on top of Lucene using Hadoop infrastructure. They treat each document as a combination of separate fields such as title text, h1, image-alt, underlined, bold etc, but observed that a simpler approach of considering a document as document-url, title and body outperformed the earlier approach. To enhance recall, they expanded the queries using WordNet and also by combining the query with all possible subsets of tokens present in the query (excluding the stop words). To prevent query drift they experimented on giving selective boosts to different steps of expansion including giving higher boosts to sub-queries containing named entities as opposed to those that didn't. In fact, this run achieved highest precision among their other runs. Using simple statistics they identified authoritative domains such as wikipedia.org, answers.com, etc and attempted to boost hits from them, while preventing them from overly biasing the results.

IRRA submitted five runs: *irra1mqa*, *irra1mqd*, *irra2mqa*, *irra2mqd* and *irra3mqd*.

Northeastern University submitted five runs: *NeuSvmBase*, *NeuSvmHE*, *NeuSvmPR*, *NeuSvmPRHE* and *NeuSvmStefan*. They used the default retrieval function provided by indri to retrieve 2000 documents, extracted features from both document text and link information, and then trained svmlight to learn the ranking function and then rank documents based on the learned function. The differences among the submitted runs how to train and use the svmlight. In particular, (1) *NeuSvmBase* trains the svmlight on MQ08 data, (2) *NeuSvmHE* trains svmlight on hard queries from MQ08 data and uses it on predicted hard queries. It also trains svmlight on easy queries from MQ08 data and uses it on predicted easy queries. (3) *NeuSvmPR* trains svmlight on recall-oriented queries from MQ08 data and uses it on predicted recall-oriented queries. It also trains svmlight on precision-oriented queries from MQ08 data and uses it on predicted precision-oriented queries. (4) *NeuSVMPRHE* is similar to the other runs, except that four categories, i.e., hard recall/easy precision/hard precision/easy recall/ are considered. (5) *NeuSvmStefan* trains on a different query log from a major web search engine for overlap between the top 100 documents retrieved by Microsoft Bing and the top 2000 documents retrieved by indri. For the query class prediction, they predicted hear/easy using Jensen-Shannon divergence among ranking features, and predicted precision/recall using a SVM classifier trained on MQ08 query precision/recall tags.

Sabir submitted five runs: *Sab9mq1bf1*, *Sab9mq1bf4*, *Sab9mq2bf1*, *Sab9mqBase1* and *Sab9mqBase4*. The last two runs are the base case SMART runs, i.e., ltu.lnu, with no expansion. The other three runs used blind feedback method with ltu.Lnu as the initial runs. Specifically, (1) *Sab9mq1bf1* assumed that top 25 documents are relevant and all other documents are non-relevant. 20 query terms are added to each query. The a,b,c weights for Rocchio feedback methods are respectively 32, 64, 128. (2) *Sab9mq1bf4* assumed that top 15 documents are relevant and all other documents are non-relevant. 5 query terms are added to each query. The a,b,c weights for Rocchio feedback methods are respectively 32, 64, 128. (3) *Sab9mq2bf1* assumed that top 25 documents are relevant and all other documents are non-relevant. 20 query terms are added to each query. The a,b,c weights for Rocchio feedback methods are respectively 32, 8, 0. They used Amchormap for query class prediction, ratio of sim at rank to sim at rank 0 for precision. Sort topics, and label top 25

University of Delaware (Carterette) submitted five runs: *udelIndDM*, *udelIndPR*, *udelIndRM*, *udelIndSP* and *udelIndri*. *udelIndri* is the baseline indri run. *udelIndDM* is the run with Metzler and Croft dependence modeling. *udelIndPR* is the run with a PageRank document prior. *udelIndRM* is the run with Lavrenko and Croft relevance models. *udelIndSP* is the indri run with a "domain trust" document prior. Domain trust is based on the frequencies of occurrence of a domain on external, public-available URL and send mail whitelists and blacklists. The parameters of all the runs are trained in a quasi-semi-supervised fashion using MQ08 data.

University of Delaware (Fang) submitted five runs: *UDMQAxBL*, *UDMQAxBLlink*, *UDMQAxQE*, *UDMQAxQEWP* and *UDMQAxQEWeb*. All of the five runs used the axiomatic retrieval models to rank documents. *UDMQAxBL* is the baseline run with the F2-LOG axiomatic retrieval function. *UDMQAxBLlink* ranks documents with F2-LOG based on both document content and anchor text. The other three runs used the semantic term matching method proposed in the axiomatic retrieval models. All of these three runs expanded original query terms with their semantically related terms. The related terms are selected from different resources: document collection (i.e., category B) for *UDMQAxQE*, Wikipedia pages in the document collection for *UDMQAxQEWP*, and snippets returned by Web search engines for *UDMQAxQEWeb*.

University of Glasgow submitted two runs: *uogTRMQdpA10* and *uogTRMQdpA40*. They used a MapReduce based indexer for the Terrier platform to index the ClueWeb09 collection. For retrieval, they investigated the application of a hyper-geometric parameter-free Divergence from Randomness weighting model.

University of Illinois at Urbana-Champaign submitted five runs: *uiuc09Adpt*, *uiuc09GProx*, *uiuc09KL*, *uiuc09MProx* and *uiuc09RegQL*. *uiuc09KL* is the baseline runs with KL-divergence retrieval model. *uiuc09GProx* and *uiuc09MProx* are two variaions of the positional relevance model (PRM) that exploits term proximity evidence so as to assign more weights to words closer to query words in feedback documents. *uiuc09MProx* estimates the PRM by first computing the joint probability of observing a word together with the query words at each position and then aggregating the evidence by summing over all the possible positions, and *uiuc09GProx* estimates the PRM by computing the association between each word and the query using documents and positions as "bridges". *uiuc09Adpt* used an adaptive retrieval model based on query classification results. And *uiuc09RegQL* is an improved document weighting in the relevance model by using a regression-based method to normalize query likelihood scores to approximate the probability of relevance.

5 Evaluation

The evaluation methods (MTC and statMAP) are described in Section 3. While both use Average Precision as a base metric, they are very different in both goal and approach. The MTC method is optimized for ranking runs by MAP; it does not attempt to directly estimate Mean Average Precision, but rather to estimate the correct ranking. The statMAP metric estimates Average Precision (per query) analogous to how election polls are conducted, using sampling. Overall we can expect MTC to produce a more accurate ranking of the runs, and statMAP a more accurate estimate of MAP.

Out of the 684 topics judged, 146 are common to all runs in the sense that all runs were involved in document selection. For the rest of the queries, several sites were held out in a round-robin fashion (details in the next section). Therefore the only consistent all-run comparison in terms of performance can be made

run	EMAP	MTC confidence	statMAP	statMAP conf interval
UDMQAxQEWeb	0.124	N/A	0.227	± 0.032
UDMQAxQE	0.099	0.793	0.149	± 0.027
UDMQAxQEWP	0.090	0.594	0.133	± 0.030
uogTRMQdph40	0.089	0.503	0.198	± 0.026
uogTRMQdpA10	0.087	0.575	0.195	± 0.025
uiuc09GProx	0.086	0.511	0.183	± 0.023
uiuc09MProx	0.082	0.682	0.179	± 0.024
UDMQAxBL	0.079	0.538	0.192	± 0.024
uiuc09Adpt	0.079	0.500	0.180	± 0.023
uiuc09RegQL	0.079	0.521	0.175	± 0.022
udelIndSP	0.075	0.566	0.169	± 0.022
udelIndRM	0.073	0.540	0.155	± 0.020
udelIndDM	0.072	0.522	0.173	± 0.022
Sab9mq1bf1	0.071	0.513	0.130	± 0.029
uiuc09KL	0.071	0.505	0.171	± 0.020
udelIndri	0.069	0.538	0.169	± 0.021
Sab9mq1bf4	0.067	0.521	0.122	± 0.023
irra1mqa	0.066	0.507	0.156	± 0.024
Sab9mq2bf1	0.062	0.556	0.127	± 0.022
UDMQAxBLlink	0.059	0.531	0.144	± 0.022
iiithExpQry	0.055	0.539	0.130	± 0.034
udelIndPR	0.054	0.511	0.123	± 0.015
Sab9mqBase1	0.052	0.522	0.111	± 0.021
Sab9mqBase4	0.052	0.500	0.111	± 0.021
irra2mqa	0.049	0.540	0.132	± 0.024
irra1mqd	0.048	0.520	0.103	± 0.018
irra3mqd	0.048	0.509	0.105	± 0.019
iiithAuEQ	0.045	0.535	0.097	± 0.022
iiithAuthPN	0.045	0.535	0.096	± 0.022
irra2mqd	0.040	0.558	0.097	± 0.015
NeuSvmStefan	0.034	0.569	0.077	± 0.017
NeuSvmBase	0.024	0.679	0.079	± 0.019
NeuSvmPR	0.023	0.530	0.076	± 0.018
NeuSVMHE	0.023	0.507	0.078	± 0.017
NeuSvmPRHE	0.022	0.540	0.073	± 0.015

Table 2: MQ09 runs evaluated by EMAP and statMAP over 146 topics to which all sites contributed judgments. EMAP estimates use relevance predictions based on document similarity features (the `erels_docsim.20001-60000` file).

using these common 146 queries only. These results are presented in Table 5. The best run (EMAP=0.12, statMAP=0.227) performs significantly worse than the performance of best runs in previous years ad hoc tracks; we speculate that the main reason for this is that most sites have not yet adapted their methods to the general web.

Table 5 presents confidence scores. The MTC confidences can be interpreted as the probability that the run performs better (in terms of MAP) than the run below given the provided judgments and model used

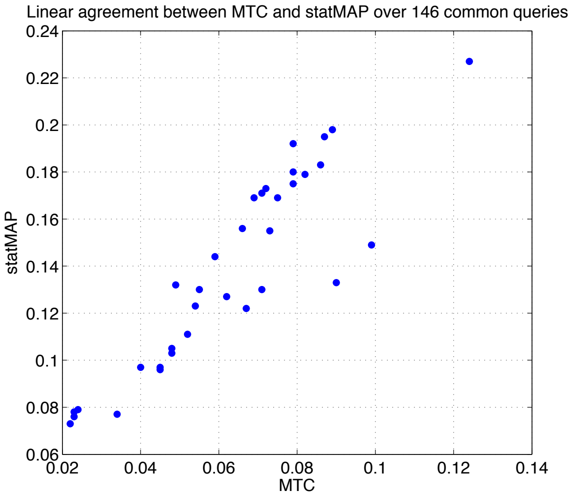

Figure 1: Common 146 queries: Linear agreement of the two measures. Kendall's $\tau = 0.80$, linear correlation coefficient ρ=0.90.

to estimate relevance. For statAP, the interval listed in the table is the 95% confidence interval computed under the assumptions explained in section 2.

A comparison of the two evaluation methods is presented in Figure 1. The two measure produce results on different scales, but linearly there is a reasonably strong correlation ($\rho = 0.90$, Kendall's $\tau = 0.80$) which, as in previous Million Query tracks, we take as an indication of validity for both measures in the absence of comparison with a strong "ground truth" measurement of MAP. We note two outliers that the two methods do not agree on: UDMQAxQE and UDMQAxQEWP. Removing these improves the correlations to $\rho = 0.95$ and $\tau = 0.90$, indicating high agreement. We have not yet determined why these two systems are different.

Table 5 presents the evaluation for each site independently over all topics for which the site participated in document selection. We call these sets the *site baselines*. It is difficult to make direct comparisons of runs between sites, because each site has its own set of baseline queries; but a within-site comparison is appropriate. Sites that chose to run more than just the first 1,000 highest-priority queries include those additional queries in their site baselines; those sites were UDel, SABIR, UOG, and IIITH, explaining why their totals are greater than the other sites.

5.1 Reusability

One of the goals of the 2009 track was to assess the reusability of a large set of queries with very shallow judgments. In general, because query sets that have (relatively) complete judgments tend to be small (usually 50 queries), there is no data available to study this problem. Data from previous years' MQ tracks provides a large sample of queries, but not the complete judgments needed to simulate reusability. Thus we have designed an experiment for *in situ* reusability testing: each site can be evaluated over a set of queries the site contributed judgments to as well as a set of queries the site did not contribute to.

In this section we describe the experimental design and analysis. More detail is presented by Carterette et al. [CKPF10].

5.1.1 Types of Reusability

We distinguish three types of reusability:

1. within-site reusability: a site uses a test collection to optimize and evaluate runs internally.

site (base topics)	run	mtc (EMAP)	statMAP
EceUDel (356 base topics)	UDMQAxQEWeb	0.115	0.271
	UDMQAxQE	0.096	0.205
	UDMQAxQEWP	0.082	0.133
	UDMQAxBL	0.081	0.255
	UDMQAxBLlink	0.059	0.192
UOG (409 base topics)	uogTRMQdph40	0.088	0.261
	uogTRMQdpA10	0.087	0.261
UIUC (358 base topics)	uiuc09GProx	0.076	0.218
	uiuc09MProx	0.075	0.209
	uiuc09RegQL	0.074	0.212
	uiuc09Adpt	0.072	0.217
	uiuc09KL	0.071	0.225
UDel (405 base topics)	udelIndri	0.072	0.212
	udelIndRM	0.071	0.173
	udelIndPR	0.058	0.169
	udelIndDM	0.080	0.236
	udelIndSP	0.078	0.211
SABIR (408 base topics)	Sab9mq1bf1	0.066	0.156
	Sab9mq1bf4	0.062	0.174
	Sab9mq2bf1	0.059	0.163
	Sab9mqBase1	0.053	0.150
	Sab9mqBase4	0.053	0.157
IIITH (369 base topics)	iiithExpQry	0.051	0.179
	iiithAuthPN	0.047	0.157
	iiithAuEQ	0.047	0.154
IRRA (357 base topics)	irra2mqa	0.051	0.152
	irra3mqd	0.049	0.151
	irra1mqd	0.049	0.144
	irra2mqd	0.043	0.121
	irra1mqa	0.061	0.193
NEU (358 base topics)	NeuSvmStefan	0.034	0.084
	NeuSvmBase	0.027	0.089
	NeuSvmHE	0.027	0.088
	NeuSvmPR	0.026	0.083
	NeuSvmPRHE	0.025	0.078

Table 3: MQ09 runs evaluated over site-baseline topics by MTC (EMAP) and statMAP. For each site, the baseline topics are those to which all of the site's runs participated in document selection: 146 topics common to all sites and runs, plus a round-robin share of the rest.

subset	topic	S_1	S_2	S_3	S_4	S_5	S_6
T_0	t_1	+	+	+	+	+	+
all-site	$\cdots$						
baseline	t_n	+	+	+	+	+	+
T_1	t_{n+1}	+	+	+	+	−	−
	t_{n+2}	+	+	+	−	+	−
	t_{n+3}	+	+	−	+	+	−
	t_{n+4}	+	−	+	+	+	−
	t_{n+5}	−	+	+	+	+	−
	t_{n+6}	+	+	+	−	−	+
	t_{n+7}	+	+	−	+	−	+
	t_{n+8}	+	−	+	+	−	+
	t_{n+9}	−	+	+	+	−	+
	t_{n+10}	+	+	−	−	+	+
	t_{n+11}	+	−	+	−	+	+
	t_{n+12}	−	+	+	−	+	+
	t_{n+13}	+	−	−	+	+	+
	t_{n+14}	−	+	−	+	+	+
	t_{n+15}	−	−	+	+	+	+
T_2	t_{n+16}	+	+	+	+	−	−
	$\cdots$	$\cdots$					
	t_{n+30}	−	−	+	+	+	+
T_3	$\cdots$						

Table 4: Illustration of proposed experimental design at the site level with $m = 6$ sites and $k = 2$ held out from each topic. Each column shows which topics a site contributed to. A + indicates that all of the sites' runs contributed judgments to the topic; – indicates that the sites' runs did not contribute judgments. Each subset $T_1 \dots T_b$ has the same contribution pattern as subset T_1.

2. between-site reusability: a site creates new runs for a test collection wishes to compare their system to that of another site using the same test collection (presumably by looking at that site's published work).

3. baseline reusability: a site reuses a test collection created at TREC and wishes to compare their system to the systems that originally participated in the TREC task.

5.1.2 Experimental Design and Analysis

The general form of the design is that there are n baseline topics to which all sites contribute judgments, and for each of the remaining $N - n$ topics, k of m sites are chosen to be held out of judgment collection. This produces b subsets of the topic set (numbered $T_1 \dots T_b$) in which all $\binom{m}{k}$ possible contribution patterns are represented. By choosing the k sites in a round-robin fashion over all $\binom{m}{k}$ possibilities, we obtain a full randomized block design, as illustrated in Table 4.

The design provides topics that can be used for each of the three types of reusability:

1. within-site: Within each subset T_i, each site contributes to $\binom{m-1}{k}$ topics and is held out from $\binom{m-1}{k-1}$ topics. Thus in addition to the n topics that all sites contribute to, each site contributes to $b\binom{m-1}{k}$ topics that can be used as a site baseline, and to $b\binom{m-1}{k-1}$ topics that can be used for testing reusability by comparing results on those topics to results on the site baseline topics. In Table 4, for instance, the within-site reusability set for site S_6 includes the first five topics in each subset, e.g. topics numbered $n + 1$ through $n + 5$ in subset T_1. The within-site baseline includes the first n all-site baseline topics along with the last 10 in each subset, e.g. those numbered $n + 6$ through $n + 15$ in subset T_1.

	baseline tests	
reuse tests	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	6	0
$p' \geq 0.05$	3	1

(a) Actual agreement.

	baseline expectation	
reuse expect.	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	7.098	0.073
$p' \geq 0.05$	2.043	0.786

(b) Expected agreement using power analysis.

Table 5: Example of actual vs. expected agreement of 10 paired t-tests among five runs evaluated over a baseline set of topics (that the runs contributed to) and a reusability set (that the runs did not contribute to). Comparing the two tables using a χ^2 goodness-of-fit test produces a p-value of 0.88, meaning we cannot reject the hypothesis that the topics and judgments are reusable.

2. between-site: Within each subset T_i, each *pair* of sites contributes to the same $\binom{m-2}{k}$ topics and is held out of the same $\binom{m-2}{k-2}$ topics. The $n + b\binom{m-2}{k}$ total topics those two sites contribute to form a baseline for comparisons between those sites. The $b\binom{m-2}{k-2}$ topics they were both held out from can be used to determine the between-site reusability. In Table 4, the first topic in each subset can be used for testing reusability between sites S_5 and S_6 against the last six that both contributed to, along with the first n in the baseline.

3. baseline: Within each subset T_i, there are $\binom{m-2}{k-1}$ topics that one site contributes to and another site does not. These topics can be used to evaluate comparing the non-contributing site to the contributing site. In Table 4, if S_5 is the "participant baseline" and S_6 is the "new system", topics numbered $n + 2$ through $n + 5$ are part of the set used to test reusability.

Once we have relevance judgments, we can compute evaluation measures, rankings, and statistical significance on topics that systems contributed to and topics systems were held out from. If these evaluations disagree, we can reject the hypothesis that the collection may be reusable.

Carterette et al. present a detailed approach to testing agreement between sets of paired t-tests [CKPF10]: if t-test significance results between pairs of runs evaluated over one of the baseline sets disagree with t-test significance results between the same pairs evaluated over one of the reusability sets, then we reject the hypothesis that the collection is reusable. Table 5 illustrates the process of comparing results from 10 paired t-tests over two different sets of queries.

5.1.3 Diagnostics

In this section we investigate some simple diagnostics to verify that there were no serious flaws in our experimental design. The main factor we will look at is whether any particular run or site had an undue effect on the judgments collected or the resulting evaluation results.

Did any site, when held out, cause fewer (or more) relevant documents to be found? On average, 24.13% of documents were judged relevant. It is the case that when the two best overall sites were held out, fewer relevant documents were found—queries for which UOG or EceUDel were held out had 22% relevant documents on average. It is also the case that when the worst overall site was held out, more relevant documents were found—queries for which NEU was held out had 25% relevant documents on average. No other site substantially changed the proportion of relevant documents when held out. Thus we can answer the question in the positive, but note that in practice these differences are miniscule, amounting to less than one relevant document per query for the number of judgments we made.

Did any site, when held out, have a strong effect on estimates of the number of relevant documents? To answer this, we looked at the statMAP estimate of $\widehat{R}$, the number of relevant documents (defined in Section 3.2). The average estimate of $\widehat{R}$ is 85.77 relevant documents, with standard error 5.09. The average estimates for UOG and EceUDel are lower (76.08 and 79.74 respectively), but still within the 95% confidence interval of the mean. Interestingly, when UIUC is held out, $\widehat{R}$ is 101.79 (± 9.16). It is surprising that holding out a well-performing site could *increase* $\widehat{R}$, though we note that the confidence intervals still overlap substantially.

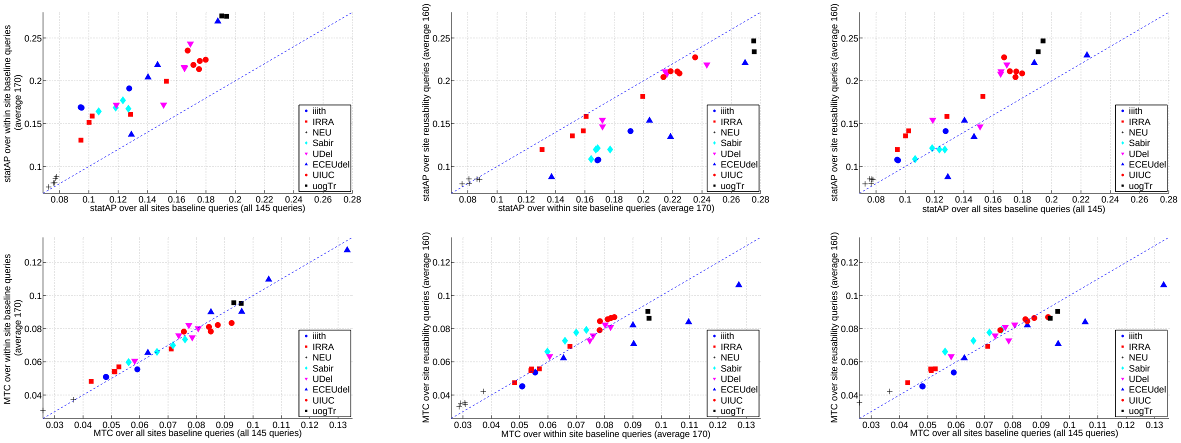

Figure 2: StatMAP and MTC EMAP scores of systems over (a) 145 baseline against 170 site baseline topics, (b) 170 site baseline against 160 site reuse topics, and (c) 145 baseline against 160 site reuse topics.

Based on the above, it seems that there were no major problems with the experimental design as applied to the MQ track. Carterette et al. present deeper analysis on the design in general in a separate work [CKPF10].

5.1.4 Results

Reusability results for MQ are illustrated in Figure 2, which shows statMAP (top) and MTC EMAP (bottom) scores of runs over (a) 145 baseline against 170 site baseline topics (left), (b) 170 site baseline against 160 site reuse topics (center), and (c) 145 baseline against 160 site reuse topics (right). Each run was evaluated over all the topics it contributed to and all the topics it was held out from, but since different sites contributed to different topics, no two sites were evaluated over exactly the same set of topics.

Differences in mean scores over baseline topics and mean scores over reusability topics for a given site may be due to a number of different effects: (1) the baseline and reuse topics are two different topic sets of different size; (2) apart from the site under study there are two other sites that did not contribute documents to each reusability topic; (3) the site under study itself did not contribute documents to the reuse topics (this is the actual effect we would like to quantify); and finally, (4) for this particular study the fact that both methods evaluate runs with a very small number of documents introduces some variability even in the baseline topics.

The plots in Figure 2 attempt to separate the second and third effects. Essentially, the comparison of the mean scores between the 145 baseline topics and the 160 site reuse topics (right) summarizes the results of the reusability experiment, and it is what an actual new site would observe by using the MQ 2009 collection. StatMAP scores over the reuse topics are positively correlated with the statMAP scores over the baseline topics, though the correlation is rather weak. MTC EMAP scores over these two sets of topics are well correlated. One can consider the other two plots as the decomposition of the effects seen in the right plot. The left plot illustrates the effect of holding out sites other than the site under study. For the statMAP case this has a rather strong effect on the scores computed, though it is minimal for the MTC scores. The middle plots try to isolate the effect of holding out the site under study. As can be seen, this also has a strong effect on the statMAP scores, while the effect is mild in the case of the MTC scores.

The plots give a visual sense of reusability, suggesting within-site may be acceptable at the level of rank agreement if not score agreement, but between-site is likely not acceptable. To quantify this, we computed three correlation statistics. First we computed the overall Kendall's *tau* between the ranking induced by the scores in the two topic sets. This is a rough estimate of the between-site reusability. For statMAP scores this is 0.7643, while for MTC EMAP scores this is 0.8350, both of which are rather low. Next we computed the Kendall's τ among the runs of each individual site to estimate within-site reusability; Table 7 shows

	baseline tests	
reuse tests	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	257	41
$p' \geq 0.05$	133	100

(a) Actual agreement.

	baseline expectation	
reuse expect.	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	302.5	26.2
$p' \geq 0.05$	85.1	117.2

(b) Expected agreement using power analysis.

Table 6: Observed versus expected agreement in significance results for between-site reusability aggregated over all Million Query 2009 sites. The χ^2 p-value is 0, indicating sufficient evidence to reject reusability.

		iiith	IRRA	NEU	Sabir	UDel	ECEUdel	UIUC	uogTr
within-site τ	statAP	0.333	1.000	0.200	0.333	0.800	0.800	-0.600	1.000
	MTC	0.333	0.800	1.000	1.000	0.600	0.800	0.800	1.000
participant comparison	statAP	0.750	0.547	1.000	0.987	0.573	0.773	0.773	0.939
	MTC	0.938	1.000	1.000	0.840	0.933	0.707	0.947	0.909

Table 7: Rank correlations based on Kendall's τ for site baseline to site reusability (top) and for comparison of site reusability to the "original" TREC runs excluding those treated as new (bottom).

these. Note that the values are not comparable across sites since the number of runs compared affects the Kendall's τ values. Finally, we computed a τ-like correlation to quantify the ability to compare "new" runs to contributing participants. For each site, we count the number of its reusability runs that are correctly ordered against the baseline runs and the number that have been swapped with a baseline run. Every comparison involves exactly one run for that site; for this measure we do not compare two runs from the same site or two runs from a different site. The final value is determined identically to Kendall's τ; the set of values can be seen in Table 7.

The significance test agreement procedure, when applied to this data, suggests that there is not enough evidence to reject within-site reusability ($p > 0.5$), but there is more than enough to reject between-site reusability ($p < 0.01$). To explain how within-site reusability holds despite some of the low τ correlations in Table 7, we note that τ is not able to capture anything about whether swaps are "reasonable". The lowest τ is -0.6 for UIUC, but by inspection (Fig. 2) UIUC's systems are all very close to each other. It is perfectly reasonable that they would be ordered differently over another set of topics, and thus the low τ is not a concern. For between-site reusability, however, we have seen that it is unlikely; that the χ^2 test confirms this is a point in its favor. The full contingency table for between-site reusability is shown in Table 6.

5.1.5 Reusability Conclusions

Our conclusions are as follows:

1. We do not reject the hypothesis that it is possible to reuse MQ09 topics and judgments for within-site comparisons, that is, comparisons between new runs that are developed by the same sites that contributed to the track.

2. We *do* reject the hypothesis that it is possible to reuse MQ09 topics and judgments for between-site comparisons, that is, comparisons between new runs developed by sites that did not contribute to the track.

Future work should investigate precisely why reusability failed and what could be done (e.g. more topics, more judgments, better MTC relevance models) to rectify the situation.

5.2 Query Class Evaluation

An optional task was to predict query intent type on two axes (precision-type vs recall-type, and hard vs easy), and potentially use an appropriate search strategy for each particular query type. Few sites attempted

		assessed PREC(306)	assessed RECALL(356)
NeuSvmPR	predicted PREC	81	108
	predicted RECALL	176	235
	unpredicted	49	13
Sab9mq1bf1	predicted PREC	64	71
	predicted RECALL	60	79
	unpredicted	182	206
Sab9mq1bf4	predicted PREC	66	66
	predicted RECALL	61	81
	unpredicted	179	209
Sab9mq2bf1	predicted PREC	50	47
	predicted RECALL	53	60
	unpredicted	203	249
uiuc09Adpt	predicted PREC	112	109
	predicted RECALL	146	234
	unpredicted	48	13
uiuc09KL	predicted PREC	80	82
	predicted RECALL	178	261
	unpredicted	48	13
uiuc09RegQL	predicted PREC	138	151
	predicted RECALL	120	192
	unpredicted	48	13

Table 8: Query Categorization: Precision vs recall confusion matrix for runs performing query precision-recall prediction. The precision-recall assessment was performed by the topic assessor, by choosing exactly one out of 6 predefined query categories.

this. While we let the participants explain how such query-type strategy affected their performance, we are providing a "truth assessment" of these categories for the judged queries.

We present the effectiveness of the runs at predicting query category as a confusion matrix, separately for precision-recall (Table 8) and hard-easy (Table 9). The precision-recall assessment was performed by the relevance judge, by choosing exactly one out of 6 predefined query categories. Overall the sites that make precision-recall predictions did so on 68% of the queries. In the cases where a prediction was made, it was right 55% of the time.

The hard-easy assessment was obtained by partitioning the query-AAP (by statAP) score range into 3 intervals: $\text{hard} = [0, 0.06), \text{medium} = [0.06, 0.17), \text{easy} = [0.17, max]$. Overall the sites made hard-easy predictions on 60% of the queries. On these, counting as errors only assessed-hard-predicted-easy and assessed-easy-predicted-hard (that is, excluding the queries assessed "medium", which could be interpreted either way), site predictions were wrong 16% of the time. The system predictions (hard or easy) were assessed as "medium" in 20% of the cases.

As it can be viewed in both tables predicting the query intent and the query hardness appears to be an enormously hard task. In particular, systems choose to not make predictions on a large number of queries.

6 Query Analysis

The aim of this section is to answer the following question: what is the number of queries needed to guarantee that, when systems are run over this many queries, their effectiveness scores reflect their actual performance? To answer this question we conduct a study based on the correlation of the systems rankings when a subset

		assessed HARD(262)	assessed MEDIUM(201)	assessed EASY(199)
NeuSvmHE	predicted HARD	169	145	119
	predicted EASY	53	49	65
	unpredicted	40	7	15
NeuSvmPRHE	predicted HARD	115	91	63
	predicted EASY	107	103	121
	unpredicted	40	7	15
Sab9mq1bf1	predicted HARD	61	45	29
	predicted EASY	44	53	54
	unpredicted	157	103	116
Sab9mq1bf4	predicted HARD	67	44	27
	predicted EASY	45	52	57
	unpredicted	150	105	115
Sab9mq2bf1	predicted HARD	53	29	15
	predicted EASY	38	39	52
	unpredicted	171	133	132

Table 9: Query Categorization: Hard vs easy confusion matrix for runs performing query hard-easy prediction. The hard-easy assessment was obtained by partitioning the query-AAP (by statAP) score range into 3 intervals.

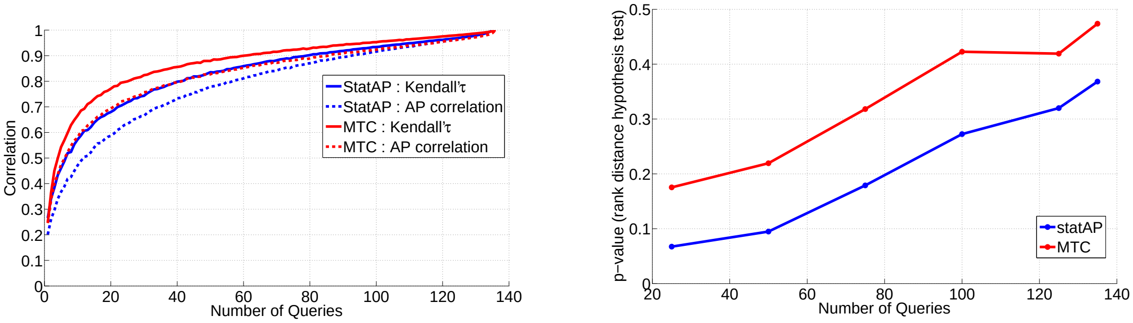

Figure 3: Correlation between rankings.

of the original queries is used and the systems rankings when all available queries are used.

For the system performance comparisons presented in Table 5 the 149 common to all systems queries were used. Out of these, 14 queries had no relevant documents found. In the analysis that follows we discard these queries[1] and use the remaining 135 queries – 87 with 64 relevance judgments and 48 with 262 relevance judgments. The next step is to determine the extent to which the number of queries can be further reduced, and how to sample the categories to achieve similar results with less overall effort.

In the correlation studies that follow we employ three correlation statistics, (a) the Kendall's τ that corresponds to the minimum number of pairwise adjacent interchanges needed to convert one ranking into the other, (b) the AP-correlation [YAR08] that more heavily weights errors (discordant pairs of systems between the two rankings) towards the top of the rankings and thus identifies how much two different rankings agree over the effectiveness of good systems, and (c) d-rank [Car09], a distance metric between two rankings that does not treat all errors equally; instead it accounts only for the significant ones.

6.1 Correlation studies over subsets of queries

In this first study we sample queries without replacement from the pool of all 135 available queries, for different sample sizes. For each query subset size, 1000 query subsets are created. The systems mean average precision over each query subset is computed and all three correlation metrics (Kendall's τ, AP-correlation and d-rank) between the induced ranking of systems and the ranking of systems over the 135 queries are calculated.

The left-hand side plot in Figure 3 shows the Kendall's τ and AP-correlation values on the y-axis with the number of queries in the query subset on the x-axis. Note that both metrics increase logarithmically with the number of queries. The solid red line indicates that MTC reaches a Kendall's τ of 0.9 after 60 queries and approximately 3,800 judgments, while statAP reaches the same Kendall's τ value after 80 queries and approximately 5,100 judgments. The AP-correlation values (dashed lines) are always lower than the Kendall's τ values indicating the difficulty in identifying the best performed systems. In particular, MTC reaches an AP-correlation of 0.9 after 80 queries and approximately 7,000 judgments, while statAP reaches the same value of AP-correlation after 92 queries and approximately 5,800 judgments.

Given that d-rank is in principle unbounded and the fact that its magnitude is highly dependent on the number of systems being ranked, the number of topics systems are evaluated over and the overall independence of the systems by the effectiveness metric, we perform a hypothesis test on the basis of d-rank to test whether the two rankings are the same [Car09] and report the p-values. A p-value sufficiently low (e.g. below 0.05) indicates that one can reject the hypothesis that the two rankings are the same with 95% confidence. When the p-values are greater than 0.05 the hypothesis cannot be rejected. The right-hand side plot in Figure 3 shows the p-values of the rank distance hypothesis test on the y-axis with the number of queries on the x-axis. The blue line corresponds to statAP and the red to MTC. Note that opposite to Kendall's τ and AP-correlation the p-values increase approximately linearly with the number of queries. Furthermore, even with a subset of as low as 25 queries the hypothesis that the system rankings over that subset of queries and the rankings over all 135 queries are the same cannot be rejected. The p-values for MTC appear to be larger than the ones for statAP indicating that MTC leads faster to stable system rankings.

6.2 Correlation studies for different query categories

We further consider whether a Kendall's τ and an AP-correlation of 0.9 may be reached with less effort when query types are sampled in different proportions. Out of the 135 common queries 87 were judged with respect to user intent (precision-oriented vs. recall-oriented) and query hardness (hard vs. medium vs. easy). We repeat the afore-described analysis by sampling queries at different proportions from each query category. Note that the empirical data to generate different splits is limited; as a result the maximum number of queries varies with the sampling proportion.

Figure 4 shows the Kendall's τ (top row) and AP-correlation (bottom row) values for different number of queries and different sampling proportions from the hard, medium and easy categories both for statAP (left column) and MTC (right column). The limited number of queries per category does not allow us to calculate how many queries are needed to reach a 0.9 correlation. Given that extrapolation may not be statistically correct we only observe the rate of increase in the correlation scores. As it can be observed for statAP a query set biased towards easy queries (blue line with circular markers, green line with circular markers and black line plus markers) seem to lead faster to high Kendall's τ values. It is also apparent that a query set biased towards hard queries do not lead to rankings similar to the ones over all 135 queries. Similar conclusions can be drawn from the AP-correlation plot for statAP. Opposite to statAP, MTC appears to reach faster well correlated system rankings when the query set consists of queries of medium difficulty. This is particularly clear in the case of AP-correlation.

Given that the mean average precision scores are dominated by high the average precision values obtained over easy queries we repeated the same analysis by using the geometric mean of average precision values ($GMAP = exp\frac{1}{n}\sum_i log(AP_i)$). Figure 5 shows the correlation scores as a function of the number of queries and the sampling proportions from the hard, medium and easy categories. Regarding statAP no clear

[1] statAP does not produce estimates of AP values for queries that have no relevant documents

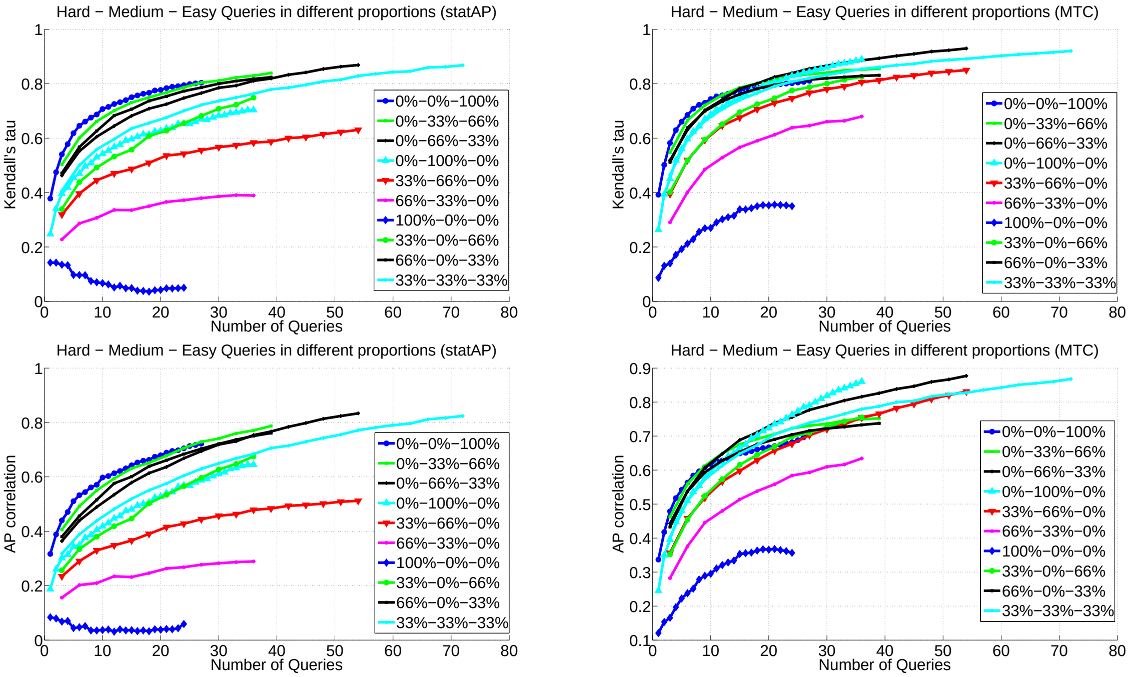

Figure 4: Kendall's tau (top row) and AP-correlation (bottom row) for statAP (left column) and MTC (right column) for different number of queries and different proportions from the hard/medium/easy categories.

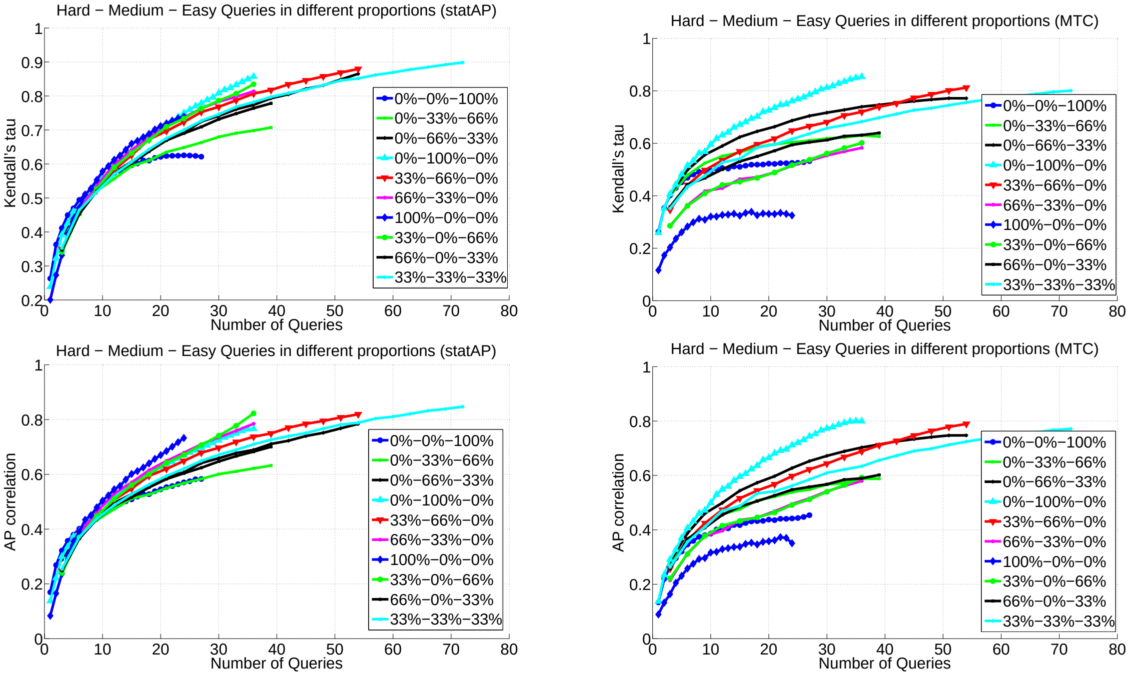

Figure 5: Kendall's tau (top row) and AP-correlation (bottom row) for statAP (left column) and MTC (right column) for different number of queries and different proportions from the hard/medium/easy categories when GMAP is employed.

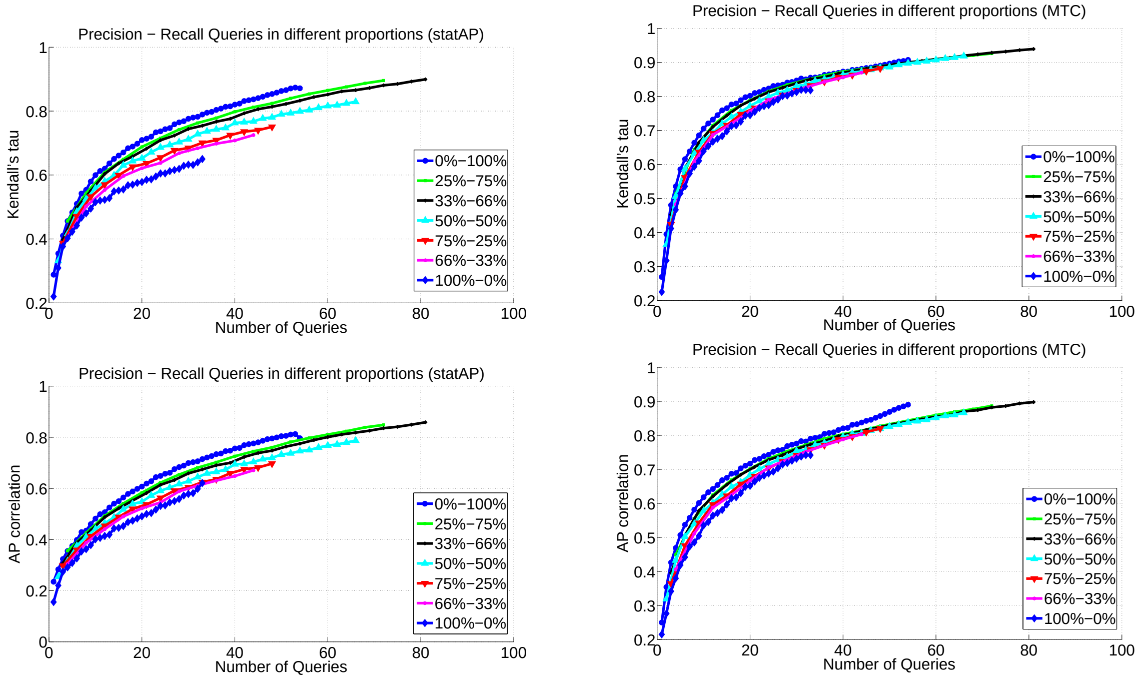

Figure 6: Kendall's tau (top row) and AP-correlation (bottom row) for statAP (left column) and MTC (right column) for different number of queries and different proportions from the precision-/recall-oriented categories.

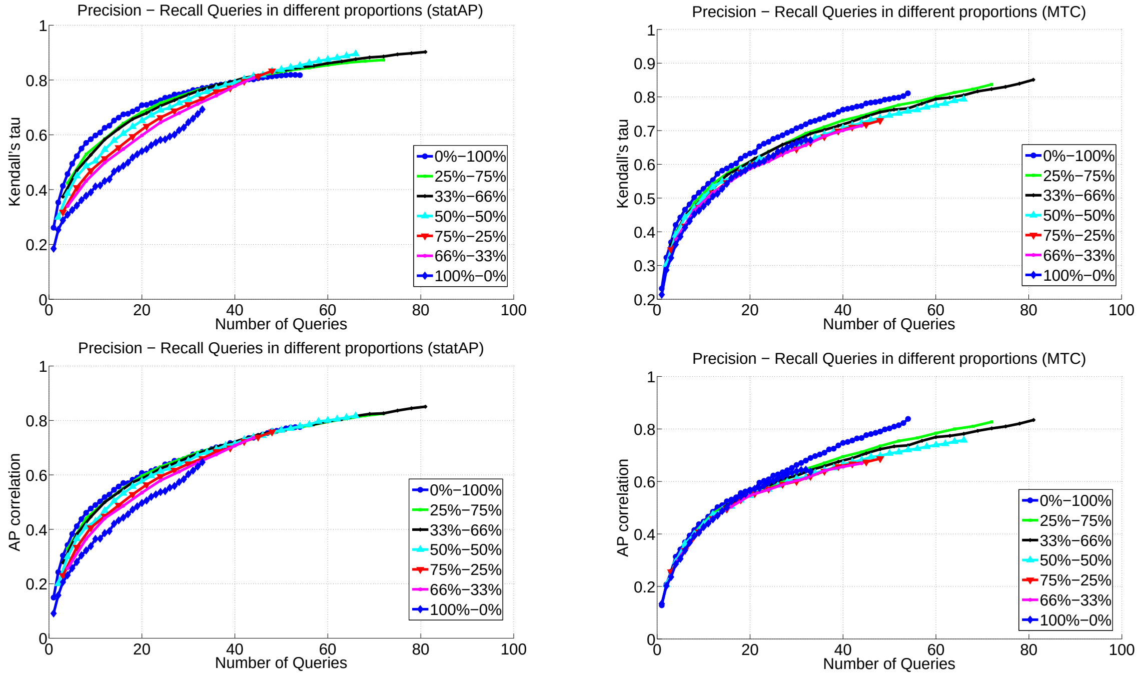

Figure 7: Kendall's tau (top row) and AP-correlation (bottom row) values for statAP (left column) and MTC (right column) for different number of queries and different proportions from the precision-/recall-oriented categories when GMAP is employed.

conclusions can be drawn, except that query sets biased towards easy queries demonstrate a slow increase in the ranking correlation. On the other hand medium difficulty queries seem to outperform any other query mixture when MTC is employed.

Figure 6 shows the Kendall's τ (top row) and AP-correlation (bottom row) values for different number of queries and different sampling proportions from the precision-oriented and recall-oriented queries. As it can be observed both in the case of statAP and in the case of MTC, correlation values increase faster when the query set mostly consists of recall queries. Note that out of the 87 queries 33 are precision-oriented while 54 are recall-oriented (This particular proportion of queries approximately corresponds to the black line with the cross marker in the plots). Further note that the distribution of the precision-oriented queries over the query hardness categories is 36%, 36% and 28% over the hard, medium and easy categories, respectively. On the other hand, the distribution of the recall-oriented queries is 22%, 44% and 33% over the query hardness categories. This indicates that the recall-oriented queries are in general easier than the precision-oriented ones, which may explain the plots in Figure 6. Due to this we repeated the same analysis utilizing GMAP. The results of this analysis are illustrated in Figure 7. As it can be observed, the original query proportion along with the one with half of the queries being precision and the other half recall-oriented appear to lead faster to high correlation scores. On the other hand, MTC still appears to lead to high correlation scores when the query set consists of recall-oriented queries.

7 Conclusions

The Million Query Track ran for the third time in 2009. The track was designed to serve two purposes: first, it was an exploration of ad hoc retrieval over a large set of queries and a large collection of documents; second, it investigated questions of system evaluation, in particular whether it is better to evaluate using many queries judged shallowly or fewer queries judged thoroughly. The aspects of retrieval evaluations investigated was the reliability of the evaluation under this MQ evaluation setup and the reusability of the resulting test collection. A query intent and hardness prediction task investigated the ability of retrieval systems to correctly classify queries and possibly adapt the retrieval algorithms respectively.

Query Prediction Task. Even though only three sites attempted to classify queries based on their intent and hardness, the task appeared to be excessively hard. None of the sites that participated in the task did significantly better than random prediction.

Reusability. Conducting *in situ* reusability experiments we concluded that we cannot reject the hypothesis that it is possible to reuse MQ09 topics and judgments for within-site comparisons, that is, comparisons between new runs that are developed by the same sites that contributed to the track. However, we *do* reject the hypothesis that it is possible to reuse MQ09 topics and judgments for between-site comparisons, that is, comparisons between new runs developed by sites that did not contribute to the track. Future work should investigate precisely why reusability failed and what could be done (e.g. more topics, more judgments, better MTC relevance models, larger effectiveness differences) to rectify the situation.

Reliability. Conducting correlation studies based on three different statistics, Kendall's τ, AP-correlation and d-rank we concluded that to rank systems similarly to our 135 ground truth queries and 18,400 total judgments MTC needs about 60 to 80 queries and 3,800 to 5,100 judgments and statAP needs about 80 to 92 queries and 5,100 to 5,800 judgments. The d-rank analysis indicates that the system rankings are not significantly different with as low as 25-50 queries and 1,600 to 3,200 judgments. Furthermore, recall queries appear to be more useful in evaluation than the precision ones (both in the correlation studies and in the variance decomposition study). The analysis over query hardness when statAP is used heavily depends on the evaluation metric used. MTC appears to reach high correlation scores when medium hardness queries are used.

References

[ACA$^+$07] James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. Overview of the TREC 2007 Million Query Track. In *Proceedings of TREC*, 2007.

[AP] Javed A. Aslam and Virgil Pavlu. A practical sampling strategy for efficient retrieval evaluation, technical report.

[APY06] Javed A. Aslam, Virgil Pavlu, and Emine Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of SIGIR*, pages 541–548, 2006.

[BH83] K. R. W. Brewer and M Hanif. *Sampling With Unequal Probabilities.* Springer, New York, 1983.

[BL07] David Bodoff and Pu Li. Test theory for assessing ir test collection. In *Proceedings of SIGIR*, pages 367–374, 2007.

[BOZ99] David Banks, Paul Over, and Nien-Fan Zhang. Blind men and elephants: Six approaches to trec data. *Inf. Retr.*, 1(1-2):7–34, 1999.

[Bre01] Robert L. Brennan. *Generalizability Theory.* Springer-Verlag, New York, 2001.

[Car07] Ben Carterette. Robust test collections for retrieval evaluation. In *Proceedings of SIGIR*, pages 55–62, 2007.

[Car09] Ben Carterette. On rank correlation and the distance between rankings. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 436–443, New York, NY, USA, 2009. ACM.

[CAS06] Ben Carterette, James Allan, and Ramesh K. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of SIGIR*, pages 268–275, 2006.

[CKPF10] Ben Carterette, Evangelos Kanoulas, Virgil Pavlu, and Hui Fang. Reusable test collections through experimental design. In *Proceedings of SIGIR*, 2010.

[CPK$^+$08] Ben Carterette, Virgil Pavlu, Evangelos Kanoulas, James Allan, and Javed A. Aslam. Evaluation over thousands of queries. In *Proceedings of SIGIR*, pages 651–658, 2008.

[CS07] Ben Carterette and Mark Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of CIKM*, pages 643–652, 2007.

[RL04] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 13–19, 2004.

[Ste] W. L. Stevens. Sampling without replacement with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 20, No. 2. (1958), pp. 393-397.*

[Tho92] Steven K. Thompson. *Sampling.* Wiley Series in Probability and Mathematical Statistics, 1992.

[YA06] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of CIKM*, pages 102–111, 2006.

[YAR08] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 587–594. ACM Press, July 2008.