Retrieval Sensitivity Under Training Using Different Measures

Ben He
University of Glasgow
Glasgow, UK
ben@dcs.gla.ac.uk

Craig Macdonald
University of Glasgow
Glasgow, UK
craigm@dcs.gla.ac.uk

Iadh Ounis
University of Glasgow
Glasgow, UK
ounis@dcs.gla.ac.uk

ABSTRACT

Various measures, such as binary preference (bpref), inferred average precision (infAP), and binary normalised discounted cumulative gain (nDCG) have been proposed as alternatives to mean average precision (MAP) for being less sensitive to the relevance judgements completeness. As the primary aim of any system building is to train the system to respond to user queries in a more robust and stable manner, in this paper, we investigate the importance of the choice of the evaluation measure for training, under different levels of evaluation incompleteness. We simulate evaluation incompleteness by sampling from the relevance assessments. Through large-scale experiments on two standard TREC test collections, we examine retrieval sensitivity when training - i.e. if a training process, based on any of the four discussed measures has an impact on the final retrieval performance. Experimental results show that training by bpref, infAP and nDCG provides significantly better retrieval performance than training by MAP when relevance judgements completeness is extremely low. When relevance judgements completeness increases, the measures behave more similarly.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Performance, Experimentation

Keywords: Training, Optimisation, Evaluation, Relevance Judgements, Binary Preference, Mean Average Precision, Inferred Average Precision, Normalised Discounted Cumulative Gain

1. INTRODUCTION

In Information Retrieval (IR) evaluation, a central issue is how to cross-compare different weighting models and different IR systems. The Text REtrieval Conference (TREC) [28] is a forum for such a cross comparison of IR systems, designed around the Cranfield paradigm [7]. In this paradigm, the evaluation process involves the use of a test collection and a set of test topics/queries. The evaluated IR system creates indices for the test collection, and returns a set of documents for each test query. In addition to the test collections and queries, the returned documents by different IR systems need to be assessed for relevance by human assessors. This refers to the relevance assessment process, which results in a list of relevant documents for each test query.

When the test collection, queries and the relevance assessments are available, one or several evaluation measure(s) is/are used for the evaluation of the IR systems. The most commonly used evaluation measures in IR are based on precision and recall. *Precision* measures the percentage of the retrieved documents that are actually relevant, and *Recall* measures the percentage of the relevant documents that are actually retrieved. The list of relevant documents for each test query is specified by the relevance assessments. In the Cranfield experiments, it was assumed that the relevance assessments were complete - i.e. all documents in the collection were assessed for each topic [7]. However, with the increasing size of the recent test collections, such a full assessment would require an infeasible number of assessor man-hours.

Since its inception in 1992, TREC has been applying a *pooling* technique [26] that allows for a cross-comparison of IR systems using incomplete assessments for test collections [28]. For each test query, the top K returned documents (normally $K = 100$) from the participating systems are merged into a single pool. The relevance assessments are then done only for the pooled documents, instead of all the documents in the test collection. The evaluation measures in TREC are task-oriented. For example, the adhoc tasks in TREC use average precision as the evaluation measure. *Average precision* is the average of the precision values after each relevant document is retrieved. For a set of test queries, mean average precision (MAP), the mean of the average precisions for all the test queries, is used to evaluate the overall retrieval performance of an IR system. Recently, with the emergence of very large test collections such as .GOV2 (25 million documents), computing MAP requires an increasingly huge amount of human effort to get a good quality pool. In 2004, Buckley and Voorhees proposed the binary preference (bpref) evaluation measure [5]. The bpref measure takes into account the judged non-relevant documents, which is claimed to be more reliable than MAP when relevance judgements are particularly incomplete [5]. bpref was used as the official measure in the TREC 2006 Terabyte adhoc task, in which only the top 50 documents returned by participating systems were pooled [3]. In addition, other measures alternate to MAP have also been proposed, for example, the inferred average precision (infAP) [29] and the normalised discounted cumulative gain (nDCG) [13]. Eval-

SIGIR'08, July 20–24, 2008, Singapore.

uation work in this direction continues in the TREC Million Query track [1].

For the many systems participating in TREC, which require prior training, a question arises: does it matter which evaluation measure should be used for training? In this paper, we assess whether any significant difference in retrieval performance can be observed when a system is trained using either MAP or an alternate measure such as bpref, infAP or nDCG[1], and when evaluated by either measure. We examine two separate test collections, providing us with insights on how training should be done for different parameters in IR models. In doing so, we gain a better understanding of how IR systems should be trained, and the level of completeness of the relevance judgements that is necessary to train an IR system successfully. This differs from previous studies, which only examined the stability of the evaluation measures under different levels of completeness, without concluding on their utility.

The main contributions of this paper are as follows. We conduct an extensive and systematic experimentation over two standard TREC test collections, which both have normal relevance judgements completeness[2] and a simulation of 16 different levels of incompleteness of the relevance judgements. According to our experiments, we conclude that the measure choice for training with normal relevance judgements completeness does not have any impact on the resulting retrieval effectiveness. However, the strength of this conclusion diminishes as the completeness of the relevance judgements decreases. Indeed, we confirm that it is significantly more stable to use the alternatives measures rather than MAP for training when there is a very low degree of relevance judgements completeness.

2. BACKGROUND

The retrieval performance of an IR system is a key feature of its ability to perform adequately in a realistic setting. The Cranfield methodology was introduced as a way of performing controlled experiments on retrieval performance [7]. However, the requirement that every document be judged for relevance with respect to every topic (known as complete judgements) is a burdensome requirement. Hence, TREC and other similar IR system evaluation forums (e.g. the Cross-Language Evaluation Forum (CLEF), and the NII-NACSIS Test Collection for IR Systems (NCTIR)) have successfully applied pooling as a means to reduce the amount of relevance judging required [10, 16, 28].

Table 1 details the collections and topics applied in the various TREC adhoc retrieval tasks. It is of note that while the number of documents in the collections has grown by several orders of magnitude over the 15 years that TREC has been running, the number of relevant documents being detected by the normal pooling process has not risen. In particular, for the recent TREC Terabyte track, using the GOV2 collection (25 million documents), the completeness of the relevance judgement is two orders of magnitude less than in the early TREC years. It follows that as IR systems are applied to larger and larger collections of documents, the techniques used in the evaluation of these IR systems need to evolve to address the issue of the incompleteness of the relevance assessments. Indeed, a central aim in the creation of the TREC Terabyte track was to define evaluation methodologies for Terabyte-scale test collection [25], and this work continues in the newer TREC Million Query track [1].

In particular, there are two ways of adapting the normal TREC evaluation process to be further resistant to incompleteness. The first of these is to modify the manner in which the pool of documents to be judged is created. For example, Cormack et al [8] proposed 'Move to Front Pooling' (which uses a variable number of documents from each source depending on its retrieval performance). Moreover, in the recent experimentation in the TREC Terabyte track, alternative pools were formed by random sampling of the systems to a higher depth or by starting pooling at a high depth [3]. A second strategy for dealing with high levels of incompleteness of relevance assessments is to use evaluation measures that are more resistant to error due to incomplete relevance assessment, e.g. bpref [5].

While various retrieval evaluation measures exist, many of the most frequently used measures are based in some way on recall and precision. The main evaluation measures used by TREC are mean average un-interpolated precision (MAP), precision at 10 documents retrieved (P@10), R-precision (R-prec) and binary preference (bpref). In this work, we focus on the most stable of these measures, MAP and bpref, because P@10 and R-prec are known to be less stable than MAP [4]. In particular, MAP is the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision of the relevant documents that are not retrieved [28].

In contrast, bpref was introduced by Buckley & Voorhees [5] because it is more stable (in terms of changes in rankings of systems) than MAP as the incompleteness of the relevance assessments used is increased. bpref is a function of the number of times documents judged non-relevant are retrieved before relevant documents. The bpref measure commonly applied is denoted as bpref-10 in [5], because this measure variant is guaranteed to use at least ten document pairs. This prevents the measure from becoming excessively coarse when the number of relevant documents is very small.

Other evaluation measures have recently been proposed as alternatives to MAP that are less sensitive to completeness. In particular, the inferred average precision (infAP) estimates average precision based on a random TREC pool subsample, which takes into account unjudged documents [29]. Next, the normalised discounted cumulative gain (nDCG) is commonly used for graded relevance assessments, and applies a logarithmic discounting function with respect to rank [13]. However, nDCG using binary relevance assessments has recently been shown to be less sensitive to completeness than MAP [24].

3. NECESSITY FOR TRAINING OF IR SYSTEMS

Most IR systems have some parameters which affect the selection and ordering of results that are returned to the user, and hence affect the overall retrieval performance of the system. While these parameters can be left at their default value, there is often an improvement in terms of retrieval performance to be gained by tuning these parameters using a supervised or unsupervised training method [30]. For instance, training a retrieval system using the PL2 weighting model [2] on TREC 10 adhoc queries, gives an MAP

[1]In the rest of this paper, we denote one of bpref, infAP and nDCG as M.

[2]We describe the used test collections as having normal levels of completeness.

Table 1: Completeness of various TREC adhoc test collections. The TREC tasks marked with asterisks are used in our experiments. Ave Rels. stands for the average number of relevant documents per query, and #Doc stands for the number of documents in the collection.

	Documents			Topics		
TREC Task	Coll.	#	GB	#	Ave Rels.	Ave Rels / #Docs
1	disk1&2	741856	2.0	50	327.72	$4.4e^{-4}$
2	disk1&2	741856	2.0	50	232.90	$3.2e^{-4}$
3	disk1&2	741856	2.0	50	196.10	$2.6e^{-4}$
6*	disk4&5	528155	1.9	50	92.22	$1.7e^{-4}$
7*	disk4&5	528155	1.9	50	93.48	$1.8e^{-4}$
8	disk4&5	528155	1.9	50	94.60	$1.8e^{-4}$
8	WT2G	247491	2.0	50	45.58	$1.8e^{-4}$
9*	WT10G	1692358	10.0	50	52.34	$3.1e^{-5}$
10*	WT10G	1692358	10.0	50	67.26	$3.9e^{-5}$
13	GOV2	25205179	425.0	49	216.67	$8.6e^{-6}$
14	GOV2	25205179	425.0	50	208.14	$8.3e^{-6}$
15	GOV2	25205179	425.0	50	117.86	$4.7e^{-6}$

of 0.2397 on the TREC 9 queries, compared to an MAP of 0.2174 using the default ($c = 1$) setting. This difference is statistically significant ($p <= 0.0009$). Indeed, many IR systems are moving to retrieval models that use many parameters, as training can be reasonably used to weight and combine many features (for example [17]).

Moreover, there has been much research done over recent years to develop new methods for training models with many parameters, for instance by attempting to directly optimise rank-based evaluation measures such as MAP [18]. In contrast, the RankNet technique described in [6, 27] avoids generating entire rankings of documents, by using a cost function calculated on document pairs, that correlates with the effectiveness of an approach that directly maximises nDCG. This work is now seen as part of the wider Learning to Rank combined field of machine learning and information retrieval [14]. Furthermore, various studies have investigated automatic unsupervised methods for setting the parameters of retrieval models automatically without the need for relevance assessments [12], and the sensitivity of retrieval models to their parameters [20].

Traditionally, training to find a setting of the parameters involves maximising the retrieval performance of the retrieval system, using a suitable measure, on a set of training queries using the corresponding relevance assessments. The performance of the system can then be measured using a set of unseen queries - known as the test set.

It is of note that retrieval evaluation measures are often non-smooth with respect to the parameter space, so finding a best setting for the parameter is often non-trivial, involving many system evaluations. Moreover, a parameter setting trained on one set of queries (or collection of documents) does not necessarily transfer to give an optimal setting on another set of queries (or collection) [12].

In this work, we are interested in the practical problem of how training with different evaluation measures affects an IR system. In particular, Section 2 demonstrated that as the size of corpora increases, the completeness assumption of the Cranfield evaluation methodology is violated to a greater extent. To address this, various IR evaluation measures that are less sensitive to relevance judgements completeness have been developed. To this end, this work compares evaluation measures, to answer the following two research questions: Firstly, how does the choice of evaluation measure affect the training of an IR system, assuming a normal level of incompleteness in the relevance judgements? Lastly, given further incompleteness, how does the choice of evaluation measure affect the training of an IR system?

We address the first research question by training various parameters across two standard TREC test collections, in Section 4. Furthermore, in Section 5, we simulate various levels of incompleteness in the relevance assessments used for training, and determine whether this has any effect on the conclusions drawn earlier.

4. EXPERIMENTS WITH NORMAL RELEVANCE JUDGEMENTS

In this section, we conduct experiments with a normal degree of relevance judgements completeness, which uses all documents from the standard TREC pools. We introduce the experimental methodology in Section 4.1, and the experimental setting in Section 4.2. Finally, we present the experimental results in Section 4.3.

4.1 Experimental Methodology

To investigate the effect that the choice of evaluation measure for training has on retrieval performance, we develop a methodology that allows the detection of any difference between training measures, as follows:

Given two evaluation measures M and MAP, we train a set of parameters of a retrieval system using a set of training queries and their corresponding relevance judgements. Then using a separate set of test queries, we use the settings derived using training measures M and MAP, and evaluate on the test queries using MAP. We denote these MAP(M) and MAP(MAP), respectively. If the retrieval performance of the parameter settings obtained using MAP(M) and MAP(MAP) do not differ by a statistically significant amount, then we conclude that there is no significant difference between training by either M or MAP.

We use the proposed methodology to test the first research question described in Section 3 above. In particular, the following sections test this question across 4 tasks from two adhoc test collections, and when training two statistically different weighting models.

4.2 Experimental Setting

In our experiments, we aim to determine the ability of various evaluation measures used for training to provide an

effective retrieval system as measured by MAP on a separate test set of queries. In this work, we experiment with four measures for training, namely MAP, bpref, infAP, nDCG. In all cases, each measure is calculated by version 9.0 of the commonly available *trec_eval* evaluation tool[3].

The experiments in this paper are conducted using the Terrier platform [21]. Stopwords are removed and Porter's stemming is applied. In our experiments, we use two different TREC test collections, namely the disk4&5, and the WT10G collections (denoted with asterisks in Table 1 above). These two collections represent two different collection types, namely a news wire collection (disk4&5) and a Web collection (WT10G). Moreover, both collections exhibit levels of completeness deemed acceptable by TREC (These levels of completeness allow the simulation of incompleteness later in this paper). We use the 100 topics from the TREC 6 & 7 adhoc tasks on disk4&5, and the 100 topics from the TREC 9 & 10 Web adhoc tasks on WT10G. Each of the involved TREC adhoc tasks is associated with 50 topics. Each topic has three different topic fields, namely title, description and narrative. We use two different combinations of the topic fields as follows:

- Title-only queries: only the title field is used.
- Full queries: all the three fields are used.

Title-only queries usually contain only few keywords, while full queries are much longer than the title-only ones. As reported in [31], in the context of language modelling, query length has an important impact on the setting of parameter values. Therefore, we experiment with two different types of queries to check if query length affects the training process.

We use two weighting models, namely the classical BM25 probabilistic model [23], and the PL2 model based on the Divergence from Randomness (DFR) probabilistic framework [2]. In addition, we apply the Bo1 term weighting model for query expansion [2]. The formulae for BM25, PL2 and Bo1 are given in Appendix (Equations (1) - (4)). Each weighting model used has a parameter (i.e. b of BM25 and c of PL2) in its term frequency normalisation component (the component that smooths term frequency in the document with respect to the document's length). Using Bo1, the exp_term most informative terms are extracted from the exp_doc top-ranked documents as expanded query terms. In applying query expansion, we add more parameters that have to be trained, allowing us to check how training with different evaluation measures affects retrieval performance with various parameters. As recommended in [2], exp_doc is optimal from 3 to 10, and the setting of exp_term depends on the characteristics of the collection used.

The parameters involved in our experiments with normal relevance judgements are BM25's b or PL2's c term frequency normalisation parameters, and in addition, the exp_doc and exp_term parameters of the Bo1 term weighting model for query expansion. These parameters are just examples for the purposes of our study - many other parameters exist in countless other retrieval strategies with which this study can be repeated. Each collection used in our experiments is associated with two TREC adhoc test topic sets, as mentioned above. On each collection, we optimise these four parameters on one topic set, by bpref, infAP, nDCG and MAP, respectively. On the other topic set, we study to which degree the training by different measures leads to different retrieval performances. We then swap the training and testing queries and repeat the experiments.

[3]Available from `http://trec.nist.gov/trec_eval/`

Table 2: Experimental results using BM25. The best result in each row is in bold.

Testing	MAP(MAP)	MAP(bpref)	MAP(infAP)	MAP(nDCG)
Title-only queries, no query expansion				
TREC 6	**0.2424**	0.2420	**0.2424**	0.2411
TREC 7	**0.1937**	**0.1937**	**0.1937**	0.1935
TREC 9	**0.2038**	**0.2038**	**0.2038**	0.1934
TREC 10	0.1956	0.1941	**0.1958**	0.1945
Title-only queries, query expansion				
TREC 6	**0.2694**	0.2678	0.2674	0.2667
TREC 7	0.2495	0.2460	0.2494	**0.2508**
TREC 9	0.2314	**0.2134**	0.2319	0.2319
TREC 10	0.2328	0.2277	**0.2354**	0.2250
Full queries, no query expansion				
TREC 6	**0.2649**	**0.2649**	**0.2649**	0.2647
TREC 7	0.2440	0.2370	0.2390>	**0.2462**
TREC 9	0.2403	0.2366	**0.2410**	0.2409
TREC 10	0.2134	0.2134	0.2134	0.2134
Full queries, query expansion				
TREC 6	0.2817	0.2819	0.2906	**0.2981**>
TREC 7	0.2900	0.2904	0.2925	**0.2945**
TREC 9	**0.2666**	0.2425	0.2647	0.2648
TREC 10	**0.2729**	0.2593	0.2723	0.2482>

In all our experiments, the term frequency normalisation parameters are optimised using Simulated Annealing [15]. exp_doc and exp_term are optimised by a two-dimensional data sweeping for $3 \leq exp\_doc \leq 10$ with interval=1, and for $10 \leq exp\_term \leq 100$ with interval=5.

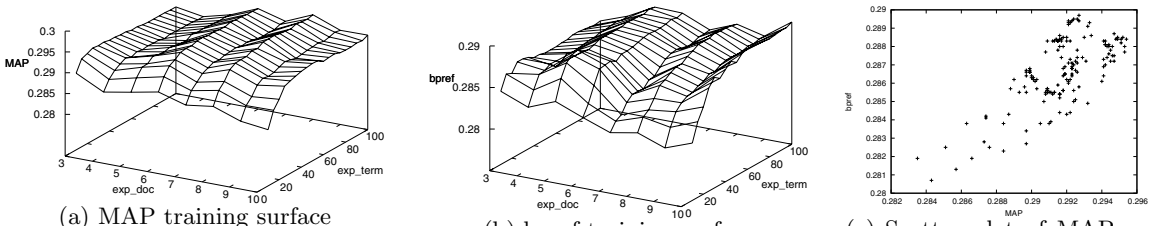
4.3 Experimental Results

In this section, we provide the experimental results with normal relevance judgements for the TREC tasks. In Tables 2 & 3, we provide the experimental results obtained using BM25 and PL2, respectively. In the tables, a MAP(M) value marked by > indicates a statistically significant difference between MAP(M) & MAP(MAP) at 0.05 level according to the Wilcoxon matched-pairs signed-ranks test. M can be bpref, infAP or nDCG. The results obtained on each test topic set are given by the parameter values trained on the other topic set associated to the collection. For example, the MAP(bpref) value obtained for TREC 6 is given by the parameter value trained by bpref on the TREC 7 task. From Tables 2 and 3, we find no major difference between MAP(M) & MAP(MAP) for the title-only queries. The difference is statistically insignificant in all cases. Moreover, for full queries, we observe the same in most cases, where the difference between MAP(M) & MAP(MAP) is minor. However, we also find three exceptions where the use of two different evaluation measures for training can possibly result in large significant difference in MAP (see the MAP(M) values marked with > in Table 2).

In addition, Figures 1 (a) & (b) show the training surface for TREC 6 topics for the MAP and bpref measures respectively. We can observe that they are similar, and the scatter plot (Figure 1 (c)) confirms this similarity (Spearman's $\rho = 0.625$). For brevity, we omit plots for infAP and nDCG, although similar observations are made.

The above experimental results suggest that for title-only queries, it does not seem to matter which of the four evaluation measures is used for training on the collections used, regardless of the parameters being trained. We observe the same for full queries, although there are three exceptions where there are significant differences between MAP(M) and MAP(MAP). Indeed, we expect full queries to be more sensitive to the parameter setting [31].

Overall, with a normal level of relevance judgement completeness, we conclude that the choice of evaluation measure



(a) MAP training surface (b) bpref training surface (c) Scatter plot of MAP vs bpref ($\rho = 0.625$)

Figure 1: Comparison of the training surface for QE parameters *exp_term* and *exp_doc* on the TREC 6 topics.

Table 3: Experimental results using PL2. The best result in each row is in bold.

Testing	MAP(MAP)	MAP(bpref)	MAP(infAP)	MAP(nDCG)
Title-only queries, no query expansion				
TREC 6	**0.2486**	**0.2486**	**0.2486**	0.2474
TREC 7	**0.1911**	0.1906	**0.1911**	0.1910
TREC 9	**0.2031**	0.2028	**0.2031**	0.2009
TREC 10	**0.1955**	0.1952	**0.1955**	0.1952
Title-only queries, query expansion				
TREC 6	**0.2510**	**0.2510**	**0.2510**	0.2508
TREC 7	**0.2499**	**0.2499**	**0.2499**	0.2492
TREC 9	0.2137	0.2111	**0.2198**	0.2151
TREC 10	**0.2219**	0.2167	0.2169	0.2133
Full queries, no query expansion				
TREC 6	0.2523	0.2523	0.2672	**0.2676**
TREC 7	**0.2440**	**0.2440**	**0.2440**	0.2437
TREC 9	0.2445	0.2430	**0.2446**	0.2445
TREC 10	**0.2249**	0.2202	0.2244	0.2232
Full queries, query expansion				
TREC 6	0.2634	0.2617	**0.2751**	0.2750
TREC 7	0.2838	0.2822	0.2826	**0.2839**
TREC 9	0.2414	**0.2476**	0.2424	0.2375
TREC 10	**0.2790**	0.2526	0.2782	0.2626

used for training does not often have a significant effect on the retrieval performance. In the next section, we continue our study with a simulation of partial relevance judgements.

5. EXPERIMENTS WITH PARTIAL RELEVANCE JUDGEMENTS

The similarity of the the obtained retrieval performances when trained using different training measures suggests that the relevance judgements in the used test collections are indeed complete. In this section, we randomly remove judgements from the training relevance assessments, to examine the effect this has on the retrieval performance of the resulting trained system. We introduce the experimental methodology in Section 5.1, and present the experimental results in Section 5.2.

5.1 Experimental Methodology

In the second part of our experiments, we conduct experiments with partial relevance judgements, instead of full relevance judgements. We only conduct experiments for full queries because the parameter settings are more sensitive for full queries [31]. In this section, we study if the choice of the measure used for training, still does not affect the resulting retrieval performance, with further incompleteness of relevance judgement. We simulate different degrees of relevance judgement completeness as follows.

Following Buckley & Voorhees who sampled relevance judgments to simulate incompleteness [5] for measure evaluation, we experiment with 16 different degrees of completeness of relevance judgements as follows (in percentage): 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80 and 90. For each degree x of relevance judgements completeness, we randomly remove $(100 - x)$ percent of relevant and non-relevant documents from the complete pool. We repeat this random sampling 10 times so that we create 10 different partial relevance judgements for each degree x. While these partial relevance judgements could have been sampled by creating 'shallower' pools, performing the sampling from the complete pool ensures that there is no bias in the partial relevance assessments towards likely retrieved documents.

For each of the 10 partial relevance judgements, of each degree x, we optimise BM25 and PL2's tf normalisation parameters (b and c) by each of bpref, infAP, nDCG and MAP. We do not use query expansion to limit the number of parameters to a feasible combinatorial space. Firstly, we would like to check if the parameter values, optimised by either of M and MAP, are stable with respect to different degrees of relevance judgements completeness. M can be any of bpref, infAP and nDCG. For each degree x, using each weighting model, on each TREC adhoc task, we compute the inverse standard deviation ($1/\sigma_{opt}$) of the parameter values optimised by each evaluation measure. As standard deviation is widely used for measuring the variation of a variable over samples, a high $1/\sigma_{opt}$ value indicates a high stability of the parameter values, optimised by the evaluation measure used, over a degree x of partial relevance judgements.

Secondly, we would like to investigate whether the training with the different evaluation measures would lead to a consistent retrieval effectiveness on the test data set. On each of the two collections used, we use one TREC topic set for training, and use the other TREC topic set for testing. We then swap the training and test topics sets and repeat the experiments. For each pair of evaluation measure M and MAP, where M can be bpref, infAP or nDCG, we have MAP(M), MAP(MAP), M(M) and M(MAP), the retrieval effectiveness evaluated by MAP or M, given by the parameter value trained using measure M or MAP, respectively. The retrieval effectiveness is evaluated on the test topic set and the parameter value is trained on the training topic set. For MAP (resp. M), we compute $1/\sigma_{diff.}$, the inverse standard deviation of the absolute difference $diff. = |MAP(MAP) - MAP(M)|$ (resp. $diff. = |M(MAP) - M(M)|$) . A high $1/\sigma_{diff.}$ value indicates a high stability of the resulting retrieval effectiveness, evaluated by M or MAP, given by parameter values trained by different measures.

5.2 Experimental Results

In this section, we present the experimental results using partial relevance judgements. As mentioned above, in the first step of our experiments, we investigate if the parameter values, optimised by M and MAP, are stable over different degrees of relevance judgements completeness. We compute the inverse standard deviation ($1/\sigma_{opt}(X)$) of the parameter values optimised by evaluation measure X, which is either M or MAP. Table 4 provides the results of the sign-test for comparing $1/\sigma_{opt}(M)$ with $1/\sigma_{opt}(MAP)$ over different degrees of relevance judgements completeness. From Table

Table 4: The number (a/b) of cases when the parameter values, optimised by M or MAP are more/less stable than the other. The values marked with $>$ (resp. $\gg$) indicate statistical significance at 0.05 (resp. 0.01) level according to the sign-test. x is the degree of relevance judgements completeness in percentage.

Model	bpref	infAP	nDCG
All 16 x degrees			
BM25	34/30	49/15$\gg$	45/19$\gg$
PL2	36/28	52/12$\gg$	52/12$\gg$
$0 < x \leq 10$			
BM25	18/6$>$	22/2$\gg$	21/3$\gg$
PL2	13/11	21/3$\gg$	19/5$\gg$
$x >$10			
BM25	16/24	27/13$>$	24/16
PL2	23/17	31/9$\gg$	33/7$\gg$

4, we observe the following:

Using both BM25 and PL2, the $1/\sigma_{opt}(X)$ values of bpref and MAP are quite close to each other over different degrees of relevance judgements completeness. $1/\sigma_{opt}(X)$ represents the stability of X when parameters are trained using different evaluation measures. For example, when BM25 is trained using bpref or MAP, over all of the 16 different degrees of relevance assessments completeness, the parameter values trained using bpref were more stable in 34 out of 64 cases[4]. In particular, according to the sign-test for the comparison between $1/\sigma_{opt}(bpref)$ and $1/\sigma_{opt}(MAP)$ (see the column bpref in Table 4), we find that bpref provides a higher $1/\sigma_{opt}(X)$ value in an insignificant number of cases, until the relevance judgements completeness is extremely low (smaller than 10 percent), when BM25 is used.

Using PL2, bpref is not more stable than MAP in terms of the resulting optimised parameter values with different relevance judgements completeness. $1/\sigma_{opt}(bpref)$ is higher than $1/\sigma_{opt}(MAP)$ in an insignificant number of cases in all circumstances.

In contrast to bpref, infAP and nDCG are more stable than MAP in most cases, particularly when relevance judgements completeness is lower than 10 percent. Indeed, on the two collections used, infAP and nDCG seem to be less sensitive to relevance judgement incompleteness than bpref and MAP. However, infAP and nDCG are comparable for the purposes of training a system with incomplete relevance assessments.

In the second step of our experiments, we investigate if the parameter values, trained by M and MAP, lead to similar retrieval effectiveness. In other words, we check if the measure choice for training affects the resulting retrieval effectiveness with various relevance judgements completeness degrees. While this should be expected, it is in fact dependent on the retrieval sensitivity of the parameters. In particular, for a pair of evaluation measure M and MAP, we compute $1/\sigma_{diff.}(M)$ (resp. $1/\sigma_{diff.}(MAP)$), the inverse standard deviation of the absolute difference $diff. = |M(M) - M(MAP)|$ (resp. $diff. = |MAP(M) - MAP(MAP)|$). X(Y) is the retrieval performance evaluated by X, when training is done by Y. A high $1/\sigma_{diff.}(X)$ value indicates a high similarity of the resulting retrieval performance, evaluated by X, of training by the two different evaluation measures. Table 5 provides the results of the sign-test for comparing $1/\sigma_{diff.}(M)$ with $1/\sigma_{diff.}(MAP)$. From Table 5, we find that the retrieval effectiveness measured by M is significantly more stable than that measured by MAP. In all but one cases, the sign-test shows a significant difference between $1/\sigma_{diff.}(M)$ and $1/\sigma_{diff.}(MAP)$. The $1/\sigma_{diff.}(X)$ values increase with the relevance information completeness. For example, using PL2 on TREC 6, when there is only 1% relevance information availale, the $1/\sigma_{diff.}(bpref)$ and $1/\sigma_{diff.}(MAP)$ values are 75.49 and 59.39, respectively. When relevance information completeness reaches 90%, the $1/\sigma_{diff.}(bpref)$ and $1/\sigma_{diff.}(MAP)$ values become 862.0 and 686.3, respectively. We do not present all the $1/\sigma_{diff.}(X)$ values for brevity reason. Overall, Table 5 shows that bpref, infAP and nDCG, are more stable than MAP in terms of evaluating retrieval performance, with different relevance judgements completeness degrees. However, it is interesting to see that such a higher stability is not affected by the degree of relevance judgements completeness. M is consistently more stable than MAP in almost all cases. Indeed, according to the number of favourable cases, bpref, infAP and nDCG are as good as each other for training an IR system.

Table 5: The number (a/b) of cases when training with M/MAP leads to a higher $1/\sigma_{diff.}$. The p-value is given by the sign-test. The values marked with $>$ (resp. $\gg$) indicate statistical significance at 0.05 (resp. 0.01) level according to the sign-test. x is the degree of relevance judgements completeness in percentage.

Model	#bpref	#infAP	nDCG
All 16 x degrees			
BM25	57/7$\gg$	60/4$\gg$	54/10$\gg$
PL2	55/9$\gg$	59/5$\gg$	58/6$\gg$
$0 < x \leq 10$			
BM25	24/0$\gg$	22/2$\gg$	17/7
PL2	23/1$\gg$	23/1$\gg$	20/4$\gg$
$\geq x >$10			
BM25	33/7$\gg$	38/2$\gg$	37/3$\gg$
PL2	32/8$\gg$	36/4$\gg$	38/2$\gg$

In addition, Figure 2 plots the mean of the retrieval performance on the TREC 6 adhoc topics using BM25. For reasons of brevity, we only report figure for bpref and MAP using BM25 on TREC 6, since we have similar observations with other model and measures on the four TREC tasks. From the figure, we find that in most cases, training by bpref provides a better retrieval performance than by MAP, when relevance judgements are highly incomplete. This finding applies even if MAP is used as the evaluation measure for the test topics. When relevance judgements become more complete, the difference in their resulting retrieval performance becomes less. Figure 3 plots the correlation between MAP and bpref across all system trainings at each sample point, for the WT10G and disk4&5 collections. From Figure 3, it is easy to see how, as expected, MAP and bpref are very correlated under high levels of completeness, but the observed correlation decreases with diminishing completeness. This mirrors the conclusions observed by Buckley & Voorhees in [5], when they studied the correlation of system rankings by MAP and bpref under various completeness levels. Moreover, while bpref exhibited a similar amount of variation in the trained parameter values to those values trained using MAP, this did not reflect in a similar resulting retrieval performance between the use of bpref and MAP

[4]For a pair of measures M and MAP, we compare their $1/\sigma_{opt}(X)$ values on four TREC ad-hoc tasks, with 16 different levels of relevance judgements completeness. There are $4 \times 16 = 64$ comparisons in total.

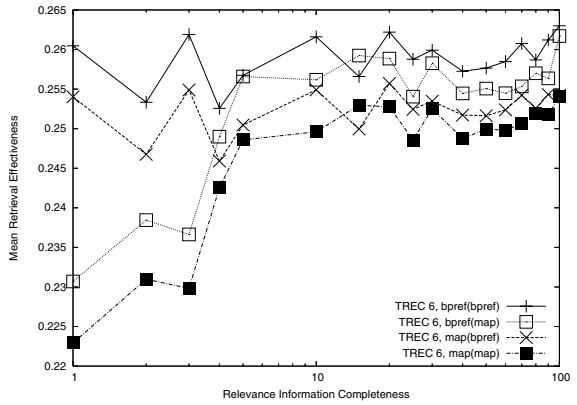


Figure 2: BM25 tested on TREC 6. The mean of the retrieval effectiveness over different degrees of relevance information completeness. X(Y) is the retrieval performance evaluated by X when the parameters are trained by Y.

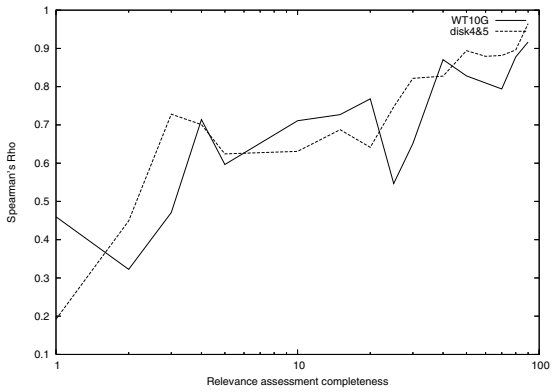


Figure 3: Correlation (Spearman's ρ) between MAP and bpref across all system trainings at each sample point, for the WT10G and disk4&5 collections.

for training, showing that training by bpref is more reliable than by MAP.

In this section, we have conducted experiments with partial relevance judgements. From Table 4, we found that the parameter values, optimised by infAP and nDCG are significantly more stable than those optimised by MAP, when the relevance judgements completeness degree is very low. On the other hand, when the relevance judgements completeness degree becomes higher, the difference in the stability between them is not necessarily significant. On the other hand, bpref was not as stable as a training measure. From the experiments for the stability of the retrieval effectiveness (Table 5), we have a different conclusion - we found that bpref, infAP and nDCG are significantly more stable than MAP regardless of the relevance judgements completeness level. Finally, from Figure 2, we found that training by M (bpref, infAP or nDCG) provides a better retrieval performance on the test topics than MAP when relevance judgements completeness is very low (smaller than 10 percent). When relevance judgements completeness increases, such an advantage of M over MAP diminishes.

6. CONCLUSIONS

In this paper, we have extensively studied (through the application of over 5,200 training runs) how the use of different evaluation measures for training affects the retrieval performance on two TREC collections, with normal and partial relevance assessments available. Four evaluation measures, bpref, infAP, nDCG and MAP, are used in our study. From our experiments with normal relevance judgements, we observe no obvious difference in the retrieval performance brought by the training process using M (i.e. bpref, infAP or nDCG) and MAP. We have also conducted experiments by simulating further relevance judgements incompleteness, following [5]. We found that training with infAP and nDCG provides significantly more stable optimised parameter values than training with MAP, when the relevance judgements completeness is extremely low, even if MAP is used as the evaluation measure for the test topics. However, training with bpref does not provide more stable optimised parameter values than MAP. Moreover, from the experiments for the stability of the retrieval effectiveness, we found that M is significantly more stable than MAP regardless of the relevance judgements completeness. Finally, we found that training by M provides a better retrieval performance than training by MAP when relevance judgements completeness is extremely low. Such an advantage of M diminishes when the completeness degree becomes higher.

Training can be seen as a core aim of IR system evaluation. Hence, the study performed here contains important conclusions relating to building robust IR systems when only sparse training data is available. Moreover, we have proposed a methodology for assessing evaluation measures for training purposes. This methodology can be applied for assessing the robustness of a trainable IR technique under sparse relevance assessments.

In the future, we plan to extend our study to the use of other evaluation measures for training. For example, we can experiment with induced AP, subcollection AP [29] or rank-biased precision [9] - alternative measures proposed for use under incomplete relevance judgements - to investigate how suitable these measures are for training compared to the four measures studied in this paper.

7. REFERENCES

[1] J. Allan, B. Carterette, J. Aslam, V. Pavlu, B. Dachev, and E. Kanoulas. Million Query TREC 2007 Overview. In *Proceedings of TREC 2007*.

[2] G. Amati. *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, Univ. of Glasgow, 2003

[3] S. Buttcher, C. Clarke and I. Soboroff. The TREC 2006 Terabyte Track. In *Proceedings of TREC 2006*.

[4] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In *Proceedings of SIGIR 2000*.

[5] C. Buckley and E. Voorhees. Retrieval evaluation with incomplete judgements. In *Proceedings of SIGIR 2004*.

[6] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of ICML 2005*.

[7] C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of SIGIR 1991*.

[8] G. V. Cormack, C. R. Palmer and C. L. A. Clarke. Efficient Construction of Large Test Collections. In *Proceedings of SIGIR 1998*.

[9] L. Gronqvist. Evaluating Latent Semantic Vector Models with Synonym Tests and Document Retrieval. In Proceedings of SIGIR 2005 ELECTRA Workshop.

[10] D. Harman, M. Braschler, M. Hess, M. Kluck, C. Peters, P. Schauble, P. Sheridan. CLIR Evaluation at TREC. In *Proceedings of CLEF 2000*.

[11] B. He and I. Ounis. Setting Per-field Normalisation Hyper-parameters for the Named-page Finding Search Task. In *Proceedings of ECIR 2007*.

[12] B. He. *Term Frequency Normalisation for Information Retrieval*. PhD thesis, University of Glasgow, 2007.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. In *ACM Transactions on Information Systems (TOIS)*, 2002.

[14] T. Joachims, H. Li, T.-Y. Liu, C. Zhai Learning to Rank for Information Retrieval (LR4IR 2007). In *SIGIR Forum*, 41:2, 2007.

[15] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[16] K. Kuriyama, N, Kando, T. Nozue and K. Oyama. Pooling for a large scale test collection : Analysis of the search results for the pre-test of the NTCIR-1 Workshop. In *Proceedings of NTCIR-1*, 1999.

[17] I. Mateevaa, C. Burges, T. Burkard, A. Laucius and L. Wong. High Accuracy Retrieval with Multiple Nested Ranker. In *Proceedings of SIGIR 2006*.

[18] D. Metzler. Direct maximization of rank-based metrics. Technical report, Univ. of Massachusetts, 2005.

[19] D. Metzler and W. B. Croft. A Markov Random Field Model for Term Dependencies. In *Proceedings of SIGIR 2005*.

[20] D. Metzler. Estimation, Sensitivity, and Generalization in Parameterized Retrieval Models. In *Proceedings of CIKM 2006*.

[21] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable judgements retrieval platform. In *Proceedings of the OSIR Workshop 2006*.

[22] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC 4. In *Proceedings of TREC 4*, 1995.

[23] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *Proceedings of TREC-1*, 1992.

[24] T. Sakai. Alternatives to Bpref. In *Proceedings of SIGIR 2007*.

[25] I. Soboroff, E. Voorhees and N. Craswell. Summary of the SIGIR 2003 Workshop on Defining Evaluation Methodologies for Terabyte Scale Test Collections. *SIGIR Forum* 37(2), 2003.

[26] K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an "ideal" judgements retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[27] M. Taylor, H. Zaragoza, N. Craswell, S. Robertson, and C. Burges. Optimisation methods for ranking functions with multiple parameters. In *Proceedings of CIKM 2006*.

[28] E. Voorhees. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

[29] E. Yilmaz and J. A. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgements. In *Proceedings of CIKM 2006*.

[30] Y. Yue, T Finley, F. Radlinski, T. Joachims. A Support Vector Method for Optimizing Average Precision. In *Proceedings of SIGIR 2007*.

[31] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR 2001*.

APPENDIX

We use the BM25 and PL2 models in this work. In BM25 [23], the relevance of a document d to a query Q is:

$$score(d,Q) = \sum_{t \in Q} w^{(1)} \cdot \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \qquad (1)$$

where k_1 and k_3 are parameters, for which the default setting is $k_1 = 1.2$ and $k_3 = 1000$ [22]. qtf is the number of occurrences of a given term in the query; K is the length normalisation component, which is given by $K = k_1((1-b)+b\frac{l}{avg\_l})$. l and avg_l are the length of document d and the average length of documents in the collection respectively. b is the term frequency normalisation hyper-parameter, for which the default setting is $b = 0.75$ [22]. In this work, we optimise b by maximising the retrieval effectiveness.

For the PL2 Divergence from Randomness (DFR) model [2], the relevance score of a document d for a query Q is:

$$\begin{aligned} score(d,Q) &= \sum_{t \in Q} qtw \cdot \frac{1}{tfn+1}\big(tfn \cdot \log_2 \frac{tfn}{\lambda} \qquad (2) \\ &\quad +(\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn)\big) \end{aligned}$$

where λ is the mean and variance of a Poisson distribution. It is given by $\lambda = F/N$. F is the frequency of the query term in the collection and N is the number of documents in the whole collection. The query term weight qtw is given by qtf/qtf_{max}, where qtf is the query term frequency. qtf_{max} is the maximum query term frequency among the query terms. The normalised term frequency tfn is given by the so-called Normalisation 2 from the DFR framework [2]:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg\_l}{l}), (c > 0) \qquad (3)$$

where tf is the term frequency of the term t in document d and l is the length of the document. avg_l is the average document length in the whole collection. c is the hyper-parameter that controls the normalisation applied to the term frequency with respect to the document length. The default value is $c = 1.0$ [2]. We optimise the parameter c to by maximising its retrieval effectiveness. For query expansion, we use the Bo1 term weighting model [2]. In Bo1, the informativeness $w(t)$ of a term t is given by:

$$w(t) = tf_x \cdot log_2 \frac{1+P_n}{P_n} + log_2(1+P_n) \qquad (4)$$

where tf_x is the frequency of the term in the pseudo-relevant set, and P_n is given by $\frac{F}{N}$. F is the term frequency of the query term in the whole collection and N is the number of documents in the collection. The top exp_term informative terms are identified from the top exp_doc ranked documents, and these are added to the query ($exp\_term \geq 1$, $exp\_doc \geq 2$). Finally, the query term frequency qtw of an expanded query term is given by $qtw = qtw + \frac{w(t)}{w_{max}(t)}$, where $w_{max}(t)$ is the maximum $w(t)$ of the expanded query terms. qtw is initially 0 if the query term was not in the original query.