Selective Labeling and Incomplete Label Mitigation for Low-Cost Evaluation

Kai Hui(✉) and Klaus Berberich

Max Planck Institute for Informatics, Saarbrücken, Germany
{khui,kberberi}@mpi-inf.mpg.de

Abstract. Information retrieval evaluation heavily relies on human effort to assess the relevance of result documents. Recent years have seen efforts and good progress to reduce the human effort and thus lower the cost of evaluation. Selective labeling strategies carefully choose a subset of result documents to label, for instance, based on their aggregate rank in results; strategies to mitigate incomplete labels seek to make up for missing labels, for instance, predicting them using machine learning methods. How different strategies interact, though, is unknown.

In this work, we study the interaction of several state-of-the-art strategies for selective labeling and incomplete label mitigation on four years of TREC Web Track data (2011–2014). Moreover, we propose and evaluate MaxRep as a novel selective labeling strategy, which has been designed so as to select effective training data for missing label prediction.

1 Introduction

Evaluation in information retrieval often relies on the Cranfield paradigm [10]. To establish the relative performance of several information retrieval systems, one agrees on a set of information needs (called *topics*), which are representative of the target workload. Each of these information needs is then formulated as a keyword query, and results are obtained from each of the information retrieval systems under comparison. Following that, human assessors label retrieved result documents with regard to their relevance. Finally, based on the collected labels, a retrieval effectiveness measure such as mean-average precision (MAP) or normalized discounted cumulative gain (nDCG) is computed to establish a relative order of the compared information retrieval systems according to their retrieval performance.

Manual labeling is laborious and costly, in particular when the number of topics and/or the number of compared systems is large. As a reaction, recent years have seen a fair amount of research that seeks to reduce the cost of information retrieval evaluation. *Selective labeling*, as a first direction, chooses a subset of returned result documents to label. Among the simplest strategies, depth-k pooling [16,17] only collects labels for documents returned in the top-k result of any of the compared systems. More sophisticated strategies leverage knowledge about the retrieval effectiveness measure used, for instance, Carterette and

C. Iliopoulos et al. (Eds.): SPIRE 2015, LNCS 9309, pp. 137–148, 2015.
DOI: 10.1007/978-3-319-23826-5_14

Allan [7] who label only documents with a potential effect on the relative order of any two systems. While cutting costs, selective labeling leads to result documents whose relevance label is not known. Such incomplete labels can also arise for other reasons, for example, when evaluating a novel information retrieval system that did not contribute to the original pool of result documents. *Mitigating incomplete labels*, as a second direction, seeks principled ways to make up for missing relevance assessments. The default of dealing with them is to assume that result documents are irrelevant if they have not been labeled. While this may appear pessimistic at first glance, it is not unreasonable given that most documents will be irrelevant to any specific information need. Alternative approaches have come up with novel effectiveness measures [3], removed documents without known label from consideration [15], and made use of machine learning methods to predict missing labels [4].

Contributions. What has received some prior attention but has not been fully explored, though, is how the different strategies for selective labeling and incomplete label mitigation interact with each other. As a *first contribution* of this paper we thus examine the interaction of state-of-the-art selective labeling and incomplete label mitigation strategies on four years of TREC Web Track data (2011–2014). The performance of different combinations is studied both in terms of approximating MAP scores (in terms of root mean square error) as well as system rankings (in terms of Kendall's τ). Also, strategies for selective labeling have typically been designed with no consideration of how incomplete labels are dealt with later on. Hence, as a *second contribution*, inspired by recent work in machine learning [19] and the cluster hypothesis [14], we propose MaxRep as a novel selective labeling strategy. MaxRep selects documents to label so as to maximize their representativeness of the pool of result documents, thus yielding effective training data for label prediction. MaxRep is formulated as an optimization problem, which permits efficient approximation.

Organization. The rest of this paper is organized as follows. Section 2 recaps existing strategies for selective labeling and incomplete label mitigation and puts our work in context. Section 3 puts forward our novel selective labeling strategy MaxRep. Our extensive experimental study is the subject of Section 4. Finally, in Section 5 we draw conclusions.

2 Technical Background and Related Work

In this section, we provide the technical background for our work by reviewing existing strategies for selective labeling and incomplete label mitigation. Moreover, we put our proposed MaxRep method in context with existing work.

2.1 Selective Labeling

Several efforts have looked into how, to reduce human effort and hence cost, only a subset of returned documents can be labeled, while still producing a reliable relative ranking of multiple information retrieval systems:

Pooling strategies merge the results returned by different systems to form a pool of result documents to be labeled by human assessors. The most common strategy, *depth-k pooling* as used by TREC, considers only documents that are returned within the top-k of any system. Cormack et al. [11], as an alternative, propose *move-to-front pooling* (MTF) as an iterative pooling procedure, requiring continued human effort, which systematically prioritizes documents returned by systems that have already returned relevant documents. Vu and Gallinari [17] make use of machine learning for pooling. Using documents from the top-5 pool as training data, they employ *learning-to-rank methods* to estimate the relevance of yet-unlabeled documents. Documents more likely to be relevant are then labeled with higher priority. Features, in their case, encode the rank at which the document was returned by different systems. Their approach thus requires two rounds of human interaction to label (i) documents in the top-5 pool as training data and (ii) a number of the remaining documents.

Aslam et al. [2] devise a biased sampling strategy that yields an unbiased estimator of MAP. A more practical sampling strategy with good empirical performance is described by Aslam and Pavlu [1]. The key idea here is to introduce a sampling distribution, so that documents ranked highly by many system, which are therefore expected to be relevant, are selected more often. The probability of selecting the document at rank r from a result list of length n is defined as

$$P[r] \approx \frac{1}{2n} \log \frac{n}{r} \ .$$

These per-system probabilities are aggregated, corresponding to choosing a system at uniform random, and documents are selected using stratified sampling.

Carterette et al. [7] propose the *minimal test collection* (MTC) method. For a specific retrieval effectiveness measure (e.g., MAP or nDCG), MTC iteratively selects discriminative documents to label until the relative order of systems has been determined. Requiring continued human interaction at every step, like MTF pooling described above, it is an active procedure.

Unlike all of the aforementioned strategies, which only take ranking information into account, our novel method MaxRep also considers document contents. Inspired by Yu et al. [19] and designed with label prediction in mind, MaxRep aims at selecting a representative set of documents from the pool of result documents to yield effective training data.

2.2 Incomplete Label Mitigation

Labels can be incomplete for different reasons, for instance, since they were collected only selectively or because the evaluated information retrieval system is novel and did not contribute to the initial result pool. Different strategies have been proposed as remedies:

As already mentioned above, a common way to deal with missing relevance labels, which is also used in TREC, is to *assume* that those documents are *irrelevant*. Given that most documents are irrelevant anyway for any specific information need, this can also be interpreted as label prediction with a simple

majority classifier. More elaborate label prediction methods will be discussed below. Sakai [15], as an alternative, proposes to remove documents without known labels from consideration yielding *condensed result lists*. Both aforementioned incomplete label mitigation strategies are agnostic to the retrieval effectiveness measure used.

In contrast, Buckley and Voorhees [3] propose bpref as an *alternative retrieval effectiveness measure* mimicking mean-average precision (MAP). With R as the number of labeled relevant documents, it is defined as

$$\text{bpref} = \frac{1}{R} \sum_{r} \left(1 - \frac{|\,\text{labeled irrelevant above rank } r\,|}{R} \right) ,$$

and the term in parenthesis can be interpreted as an estimator of precision at rank r. In their experiments, bpref proved robust and exhibited high rank correlation with MAP. However, in terms of numerical value, bpref may deviate from MAP if many labels are missing. Yilmaz and Aslman [18] describe two alternatives, based on sampling theory, that are closer to MAP. The first, induced average precision (indAP), removes documents with unknown label from consideration and can be seen as an application of the condensed list approach [15] to MAP. The second, inferred average precision (infAP), relies on the following improved estimator of precision at rank r

$$E[\text{precision at } r] = \frac{1}{r} + \frac{(r-1)}{r} \left(\frac{|\,\text{labeled above rank } r\,|}{r-1} \cdot \frac{|\,\text{labeled relevant}\,|}{|\,\text{labeled}\,|} \right) ,$$

which also takes into account what fraction of documents has been labeled.

Another family of strategies uses machine learning methods to *predict missing relevance labels*. Carterette and Allan [6] use regularized logistic regression to predict the relevance of documents. Building on the cluster hypothesis [14], document features encode *tf.idf*-based cosine similarity with documents whose labels are known. Büttcher et al. [4], to the same end, explore two approaches, namely a simple classifier based on statistical language models and a support vector machine (SVM). For the latter, document features are *tf.idf*-weights for the 10^6 most common terms in the document collection. Given the good performance of the SVM-based label prediction in their experiments, we use this as one of the incomplete label mitigation strategies in our experiments.

3 Selecting Representative Documents to Label

We now describe MaxRep, our novel strategy for selective labeling. In contrast to existing strategies, MaxRep not only considers ranking information but also takes into account document contents. Intuitively, it aims at selecting a subset of documents that is representative, in particular of those documents expected to be relevant. MaxRep thus harvests effective training data for label prediction, since documents are representative of the overall pool of result documents, and it also makes up for the inherent bias against relevant documents.

Let $\mathcal{D}$ denote the pool of result documents for a specific topic. Our objective is to select a k-subset $\mathcal{L} \subseteq \mathcal{D}$ that best represents the pool of result documents. Intuitively, if two documents have similar contents, there is no need to label both of them, since their labels tend to be identical. We let $sim(d_i, d_j) \in [0, 1]$ denote a measure of *content similarity* between documents d_i and d_j. Further, we let $rel(d_i) \in [0, 1]$ denote a measure of *expected relevance* of document d_i

Our concrete implementation uses the cosine similarity between $tf.idf$-based document vectors as a measure of document content similarity. More precisely, with $tf(v, d)$ as the term frequency of term v in document d, $df(v)$ as its document frequency, and n as the total number of documents in the collection, the feature weight for term v in document vector $\boldsymbol{d}$ is

$$\boldsymbol{d}(v) = tf(v, d) \log \frac{n}{df(v)} ,$$

and we measure the similarity between documents d_i and d_j as

$$sim(d_i, d_j) = \frac{\boldsymbol{d}_i \cdot \boldsymbol{d}_j}{\|\boldsymbol{d}_i\| \, \|\boldsymbol{d}_j\|}$$

which ranges in $[0, 1]$ given that we only have non-negative feature weights. As in Büttcher et al. [4] our implementation only considers the 10^6 most frequent terms from the document collection. Moreover, in order to reduce noise, we ignore similarities below 0.8, setting them to zero, when choosing representative documents. As a measure of expected relevance our concrete implementation uses the probability according to the sampling distribution also used in Aslam and Pavlu [1] and described in Section 2.

We measure the representativeness of a document set $\mathcal{L}$ as

$$f(\mathcal{L}) = \sum_{d_i \in \mathcal{D}} rel(d_i) \max_{d_j \in \mathcal{L}} (sim(d_i, d_j)) \ . \tag{1}$$

This formulation rewards document sets that cover all documents from $\mathcal{D}$ that are expected to be relevant by including at least one similar document.

Building on this, we cast selecting the set of k most representative result documents into the following optimization problem

$$\operatorname*{argmax}_{\mathcal{L}} f(\mathcal{L}) \quad \text{s.t.} \quad |\mathcal{L}| = k$$

It turns out that the above optimization problem permits efficient approximation thanks to the submodularity of its objective function, which we state in the following lemma.

Lemma 1 (Submodularity). *Equation 1 defines a submodular function. Given two document sets $\mathcal{L}$ and $\mathcal{L}'$ with $\mathcal{L} \subseteq \mathcal{L}'$ and a document $d \in \mathcal{D}$, then*

$$f(\mathcal{L} \cup \{d\}) - f(\mathcal{L}) \geq f(\mathcal{L}' \cup \{d\}) - f(\mathcal{L}') \ .$$

Proof (of Lemma 1). We can rewrite for $\mathcal{X} \in \{\mathcal{L}, \mathcal{L}'\}$

$$f(\mathcal{X} \cup \{d\}) - f(\mathcal{X}) = \sum_{d_i \in \mathcal{D}} rel(d_i) \max\left(0, sim(d_i, d) - \max_{d_j \in \mathcal{X}} sim(d_i, d_j)\right) .$$

Now,

$$\begin{aligned} \mathcal{L} \subseteq \mathcal{L}' \quad &\Rightarrow \quad \forall\, d_i \in \mathcal{D} \,:\, \max_{d_j \in \mathcal{L}} sim(d_i, d_j) \leq \max_{d_j \in \mathcal{L}'} sim(d_i, d_j) \\ &\Rightarrow \quad f(\mathcal{L} \cup \{d\}) - f(\mathcal{L}) \geq f(\mathcal{L}' \cup \{d\}) - f(\mathcal{L}') \,. \end{aligned}$$

□

Having established the submodularity of our objective function, we can make use of the result by Nemhauser et al. [12] and greedily build up the set of representative documents $\mathcal{L}$. More precisely, starting from $\mathcal{L}_0 = \emptyset$, in the i-th iteration we include the document from $\mathcal{D} \setminus \mathcal{L}_{i-1}$ that maximizes $f(\mathcal{L}_i)$, and finally report $\mathcal{L}_k$ as a result. This greedy algorithm gives a $(1 - \frac{1}{e})$-approximation [12], guaranteeing the performance of the proposed greedy algorithm.

4 Experimental Evaluation

In this section, we describe our experimental evaluation. We report on the performance of different combinations of strategies for selective labeling, including MaxRep as the one proposed in this work, and incomplete label mitigation. This is done on four years' worth of participant data from the TREC Web Track (2011–2014), and we investigate how well combinations can approximate the system ranking, in terms of Kendall's τ, but also how well they can approximate MAP scores, in terms of root mean square error (RMSE).

4.1 Datasets

Our experiments are based on the ClueWeb09[1] and ClueWeb12[2] document collections. Queries and relevance labels are taken from the adhoc task of the TREC Web Track (2011–2014). This leaves us with a total of 200 queries (50 per year) and their corresponding relevance labels. We also obtained the runs submitted by participants of the TREC Web Track. There are 62 runs for 2011, 48 runs for 2012, 61 runs for 2013, and 42 runs for 2014. For each submitted run we consider the top-20 search results returned. In 2013 a subset of 21 queries was only labeled up to depth 10. For those queries we apply the condensed list approach, that is, for each system we consider the 20 highest-ranked labeled documents as its result.

[1] http://www.lemurproject.org/clueweb09.php/
[2] http://www.lemurproject.org/clueweb12.php/

4.2 Methods

We consider the following non-active strategies for *selective labeling*:

- **uniform random sampling**, as described by Buckley and Voorhees [3], we give the method an advantage by sampling retrospectively from relevant and irrelevant documents (we report averages based on 30 repetitions);
- **incremental pooling**, as described by Carterette [5,7], we select documents according to the best rank assigned by any system and break ties according to the average rank across all systems;
- **statAP**, as described by Aslam and Pavlu [1], with additional judgments obtained from pooling (we report averages based on 30 repetitions);
- **our method** MaxRep as described in Section 3.

To *mitigate incomplete labels*, we consider the following strategies:

- **trec-map** treats documents with unknown label as irrelevant;
- **bpref** [3] separates the labeled non-relevant documents from unlabeled documents;
- **indAP** [18] regards missing labels as non-existing
- **infAP** [18] relies on an improved estimator of precision at rank r
- **statAP** [1] computes AP with adjustments by inclusion probability from the document sampling phase
- **predict-map**, SVM-based label prediction approach [4], which we implemented using the scikit-learn [13] toolkit.

This gives us a total of 21 combinations to investigate. Given that statAP as a strategy for mitigating incomplete labels requiring inclusion probabilities as an input from selective labeling, we only compute statAP when labels have been selected with statAP itself.

4.3 Approximation of System Ranking and MAP Scores

Our first experiment studies how well different strategies can approximate the system ranking in terms of Kendall's τ and how well they can approximate the MAP scores of individual systems. To this end, we select a varying percentage, from 1% up to 95%, to label using the different strategies. Figure 1 shows the Kendall's τ value obtained for different selective labeling strategies on each of the four years (2011–2014) considered. Comparing the different incomplete label mitigation strategies, we observe that predict-map, the SVM-based label prediction approach, consistently achieves good performance, regardless of how documents to label are selected. In most plots, with as little as 20% of labeled documents, predict-map thus achieves a Kendall's τ value above 0.9, which indicates that the obtained system ranking is practically indistinguishable from the ground truth. Using trec-map and assuming that documents without known labels are irrelevant, totally mixing the labeled non-relevant and unlabeled documents, at the other extreme, performs worst in most plots. Not surprisingly, this is most

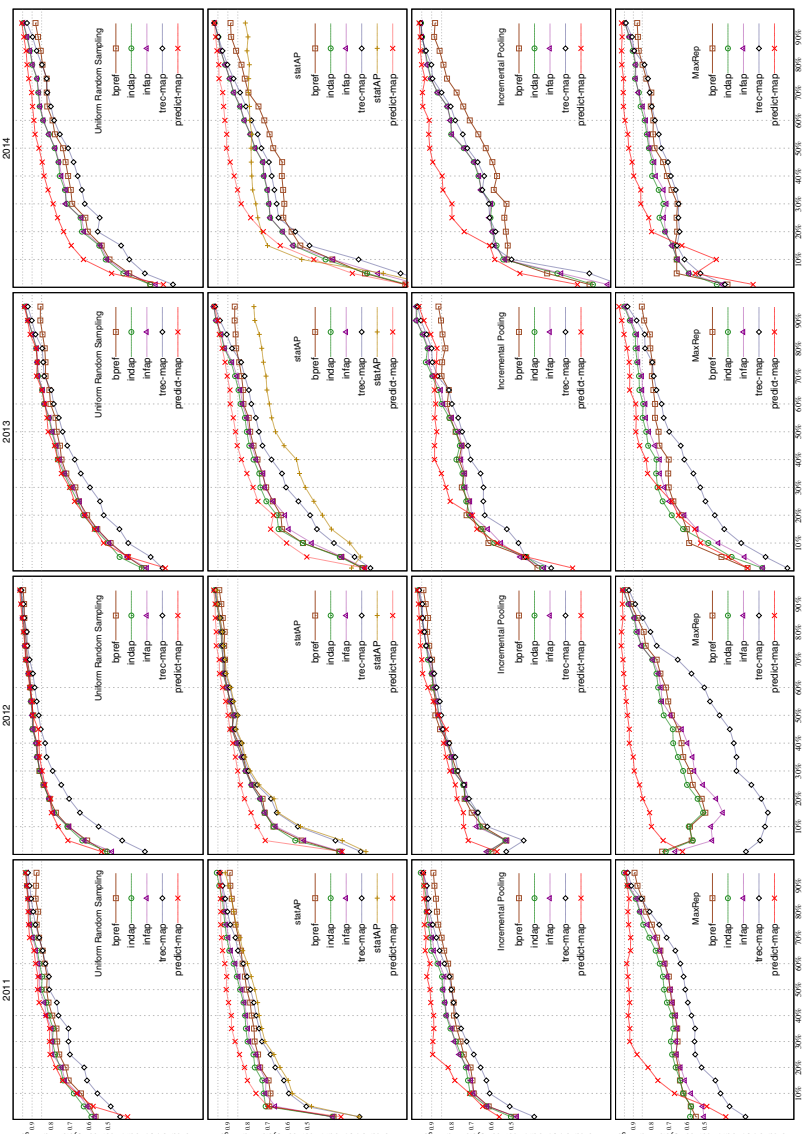

Fig. 1. Approximation of system rankings: Columns correspond to different years of the TREC Web Track. Rows correspond to different selective labeling strategies. X-axes indicate percentages of labeled documents. Y-axes indicate Kendall's τ correlation.

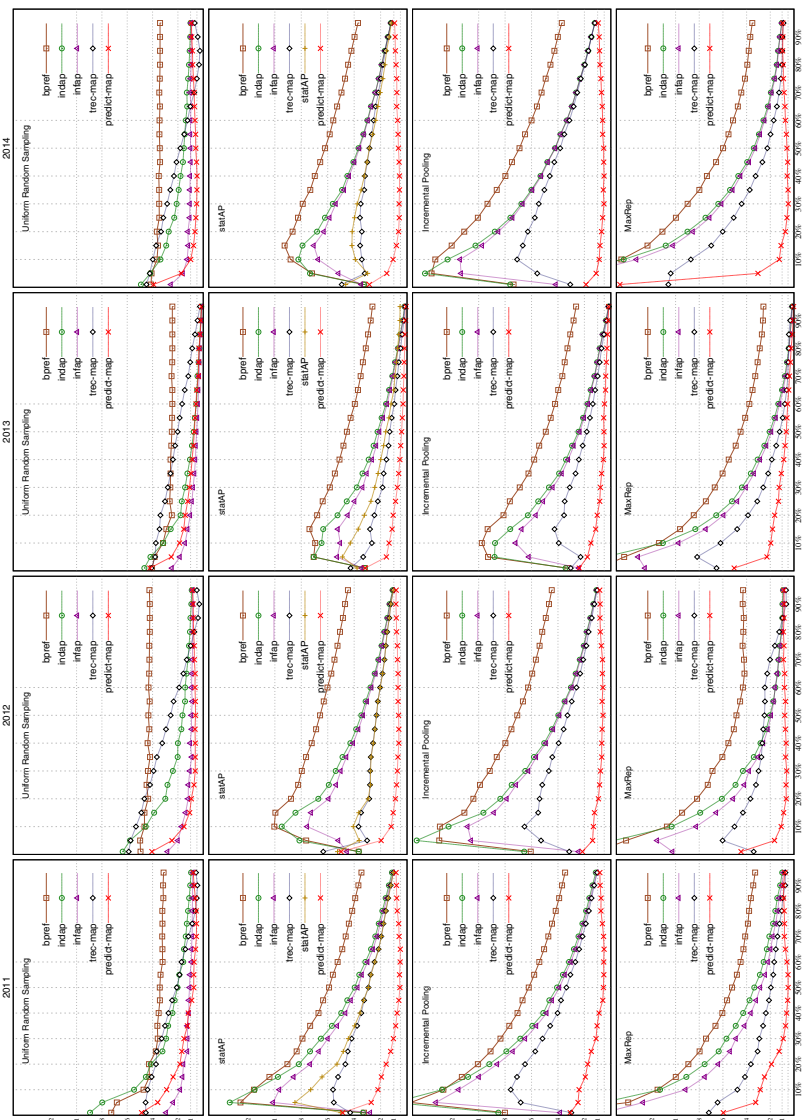

Fig. 2. Approximation of MAP scores: Columns correspond to different years of the TREC Web Track. Rows correspond to different selective labeling strategies. X-axes indicate percentages of labeled documents. Y-axes indicate root mean square error (RMSE).

pronounced when using our selective labeling strategy MaxRep. Figure 2 plots the corresponding root mean square error (RMSE), measuring how well the different combinations approximate MAP scores of individual systems. Predicting missing labels using predict-map again achieves the best result by yielding lowest approximation errors. The highest approximation errors are almost consistently seen for bpref, which is not surprising given that, as described in Section 2, it is different from MAP.

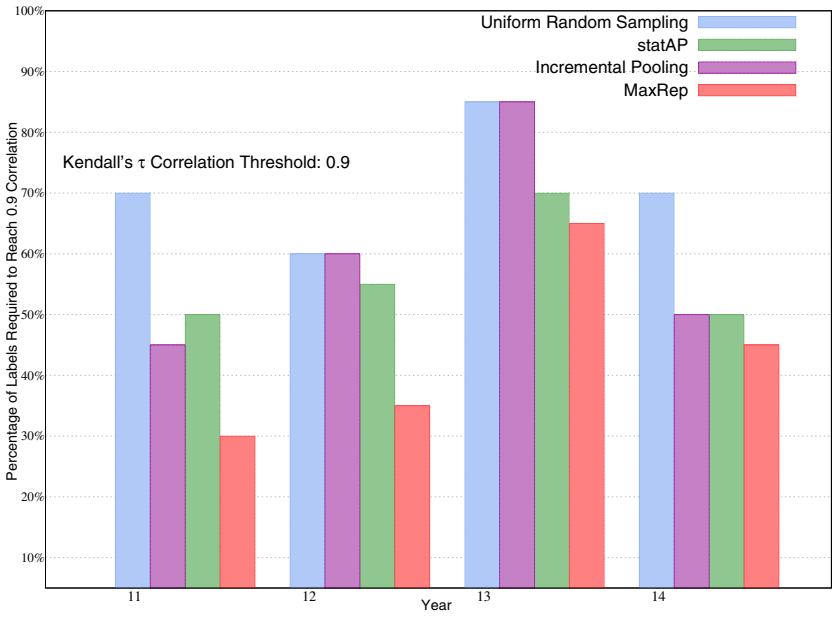

Fig. 3. Percentage of labeled documents required to achieve a Kendall's τ correlation above 0.9 when using label prediction.

4.4 Selective Labeling under Label Prediction

Given the good performance of label prediction in the previous experiment, we now investigate which selective labeling strategy performs best with it. To this end, in Figure 3, we plot the percentages of documents that need to be labeled, with different selective labeling strategies, when using predict-map for label prediction to achieve a Kendall's τ score above 0.9.

As can be seen, our selective labeling strategy MaxRep performs best across all four years under consideration. It thus consistently requires the lowest percentage of documents to be labeled to achieve a system ranking that is practically indistinguishable from the ground truth. Its relative advantage is clearest for the years 2011 and 2012 for which MaxRep requires as little as $30 - 35\%$ of labeled documents. Also in this experiment, uniform random sampling performs worst,

typically requiring more than 60% of labeled documents to achieve a Kendall's τ value above the threshold. Additionally, we conduct paired two-tailed t-test between different baselines w.r.t. our method for these least percentage of labels required to get over 0.9 correlation, and our method outperform the uniform random sampling and incremental pooling at 95% significant level (p-value=.008 and .032), meanwhile outperform the statAP at 90% level (p-value=.063).

As for comparison on RMSE, from Figure 2, we can see that our method is comparable to other methods in terms of approximating MAP scores. However, no clear winner is observed among different selective labeling methods when combined with mitigation through label prediction.

5 Conclusion

Low-cost evaluation has been an active area of research within information retrieval for the past decade. In this work, we have investigated how different strategies for selective labeling and mitigating incomplete labels interact. To this end, we conducted a large-scale experimental evaluation on ClueWeb09/12 with participant data from the adhoc task of TREC Web Track 2011–2014. We found that label prediction is a robust and viable strategy to mitigate incomplete labels, as long as at least 20% of documents have been labeled as training data. Moreover, with label prediction in mind, we proposed a novel strategy MaxRep for selective labeling. In contrast to existing strategies, it considers both ranking information and document contents and seeks to select a representative subset of documents to label. Our experiments confirmed that MaxRep is beneficial and outperforms other strategies when label prediction is used.

As part of our ongoing research, we investigate how strategies for selective labeling and incomplete label mitigation can be adapted for retrieval effectiveness measures such as α-nDCG [9] and ERR-IA [8] that capture novelty & diversity. Moreover, we study the reusability of this semi-automatically generated labeled collection, examining the reliability in evaluating systems without contributing to the initial document collection.

References

1. Aslam, J.A., Pavlu, V.: A practical sampling strategy for efficient retrieval evaluation. Report (May 2007)
2. Aslam, J.A., Pavlu, V., Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In: SIGIR, pp. 541–548 (2006)
3. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: SIGIR, pp. 25–32 (2004)
4. Büttcher, S., Clarke, C.L.A., Yeung, P.C.K., Soboroff, I.: Reliable information retrieval evaluation with incomplete and biased judgements. In: SIGIR, pp. 63–70 (2007)
5. Carterette, B.: Robust test collections for retrieval evaluation. In: SIGIR, pp. 55–62 (2007)

6. Carterette, B., Allan, J.: Semiautomatic evaluation of retrieval systems using document similarities. In: CIKM, pp. 873–876 (2007)
7. Carterette, B., Allan, J., Sitaraman, R.: Minimal test collections for retrieval evaluation. In: SIGIR, pp. 268–275 (2006)
8. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: CIKM, pp. 621–630 (2009)
9. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR, pp. 659–666 (2008)
10. Cleverdon, C.: The cranfield tests on index language devices. In: Aslib proceedings, vol. 19, pp. 173–194. MCB UP Ltd (1967)
11. Cormack, G.V., Palmer, C.R., Clarke, C.L.A.: Efficient construction of large test collections. In: SIGIR, pp. 282–289 (1998)
12. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of approximations for maximizing submodular set functions–i. Mathematical Programming **14**, 265–294 (1978)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
14. Rijsbergen, C.J.V.: Information Retrieval, 2nd edn. Butterworth-Heinemann, Newton (1979)
15. Sakai, T.: Alternatives to bpref. In: SIGIR, pp. 71–78 (2007)
16. Spärck Jones, K., Van Rijsbergen, K.: Information retrieval test collections. Journal of Documentation **32**(1), 59–75 (1976)
17. Vu, H.-T., Gallinari, P.: A machine learning based approach to evaluating retrieval systems. In: HLT-NAACL, pp. 399–406 (2006)
18. Yilmaz, E., Aslam, J.A.: Estimating average precision with incomplete and imperfect judgments. In: CIKM, pp. 102–111 (2006)
19. Yu, K., Bi, J., Tresp, V.: Active learning via transductive experimental design. In: ICML, pp. 1081–1088 (2006)