THE PREDICTIONS OF PERFORMANCE METRICS IN INFORMATION RETRIEVAL: AN EXPERIMENTAL STUDY

Sinyinda Muwanei[1], Sri Devi Ravana[2*], Wai Lam Hoo[3], Douglas Kunda[4]

[1,2,3]Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

[1,4]School of Science Engineering and Technology, Mulungushi University, Kabwe, Zambia

Email: smuwanei@gmail.com[1], sdevi@um.edu.my[2*](corresponding author), wlhoo@um.edu.my[3], dkunda@mu.ac.zm[4]

DOI: https://doi.org/10.22452/mjcs.sp2021no2.3

ABSTRACT

Information retrieval systems are widely used by people from all walks of life to meet diverse user needs. Hence, the ability of these retrieval systems to return the relevant information in response to user queries has been a matter of concern to the information retrieval research community. To address this concern, evaluations of these retrieval systems is extremely critical and the most popular way is the approach that employs test collections. This approach has been the popular evaluation approach in information retrieval for several decades. However, one of the limitations of this evaluation approach concerns the costly creation of relevance judgments. In recent research, this limitation was addressed by predicting performance metrics at the high cut-off depths of documents by using performance metrics computed at low cut-off depths. However, the challenge the research community is faced with is how to predict the precision and the non-cumulative gain performance metrics at the high cut-off depths of documents while using other performance metrics computed at the low cut-off depths of at most 30 documents. This study addresses this challenge by investigating the predictability of performance metrics and proposing two approaches that predict the precision and the non-cumulative discounted gain performance metrics. This study has shown that there exist dataset shifts in the performance metrics computed from different test collections. Furthermore, the proposed approaches have demonstrated better results of the ranked correlations of the predictions of performance metrics than existing research.

Keywords: Performance metrics, correlation, prediction, evaluation, retrieval, high cost, low cost

1.0 INTRODUCTION

Information retrieval systems are used by people from all walks of life to meet diverse information needs. Hence, the ability of these systems to return the relevant information in response to user queries has for several decades been a matter of concern to the information retrieval (IR) research community. To address this concern, evaluations of these retrieval systems are critical and the most popular way to evaluate these systems is the test collection based approach. This approach has been the popular evaluation approach in information retrieval for several decades. A test collection comprises a corpus of documents, topics and relevance judgments [1]. Topics are representations of information needs of users, while relevance judgments are found in query relevance files and they show which documents from the corpus are either relevant or not relevant to each of the topics.

Despite many research efforts to improve the test collection based approach, there are still several limitations that require to be addressed and one such limitation is the cost to generate the relevance judgments. Therefore, over the decades, there have been several proposals from the research community to address this limitation. For instance, there have been proposals to perform evaluations of retrieval systems without the use of relevance judgments [2] and also generating relevance judgments using machine learning approaches [3]. Of particular interest to our research is the recent proposal that addresses the above-highlighted limitation through predicting the performance metrics at the high cut-off depths of documents by using other performance metrics that were computed at the low-cut-off depths [4]. The authors of this proposal referred to the performance metrics that were computed at the cut-off depths of at least 100 documents as the high-cost performance metrics. In addition, the performance metrics that were computed using the cut-off depths of less than 100 documents were referred to as low-cost performance metrics and in our study, we adopt this naming. The benefit of using predictions for the high-cost performance metrics is that the usage of relevance judgments is minimized and largely restricted to the computation of the low-cost performance metrics. Therefore, this leads to a reduction in the cost of generating relevance judgments. Also, for this proposal, the authors [4] reported that the high-cost rank biased precision (RBP) metrics were accurately predicted at the cut-off depths of at most 30 documents for the low-cost performance metrics. However, at the same low cut-off depth of 30 documents, they reported inaccurate predictions of performance metrics such as precision

and non-cumulative discounted gain (nDCG).

The challenge the research community is faced with is how to predict the precision and nDCG performance metrics at the high cut-off depths of documents while using other performance metrics that have been computed at the low cut-off depths of at most 30 documents (i.e $d \leq 30$). Also, a close inspection of this recent proposal [4] reveals that the authors used one set of test collections to compute the performance metrics employed for the training of regression models for their approach, but used a different set of test collections to compute performance metrics employed for testing purposes. The use of different sets of test collections to compute performance metrics for training and testing purposes was done without a prior investigation to ascertain whether such use was appropriate for this research. Such use of sets of test collections would only be permissible in the case where there is a high similarity in performance metrics computed from the sets of test collections.

To address the above-mentioned challenges, we first conduct a brief survey of related literature. This is followed by a discussion of the methodology followed to conduct the two sets of experiments reported in this study. The first experiment investigates whether machine learning models trained using performance metrics computed using one set of test collections may be used to predict performance metrics computed using a different set of test collections. The second experiment implements the proposed approaches for the predictions of the high-cost performance metrics. The results and discussion of the experiments are reported and lastly, the conclusion.

A more recent study [5] presented a preliminary investigation into the usage of transformed sets of topic scores of the low-cost performance metrics to predict the high-cost performance metrics. Therefore, our research extends this study [5] and provides the following contributions:

1. To seek to determine the predictability of performance metrics. That is, to determine whether machine learning models trained using performance metrics computed using one set of test collections may be used to predict high-cost performance metrics computed using a different set of test collections.

2. To suggest several ways of improving the approaches of predicting the high-cost performance metrics.

3. To propose two approaches for the prediction of the high-cost precision and nDCG performance metrics that employ topic scores and ranked correlation coefficients to select the suitable set of low-cost performance metrics for use in the training and prediction processes. To our knowledge, this is the first time in information retrieval evaluation that ranked correlation coefficients are employed to select the best set of low-cost performance metrics for use in predictive regression models.

The rest of the paper is structured as follows: related works are reported in section 2, materials and methods are presented in section 3, results are presented in section 4, the discussion is presented in section 5 and lastly, the conclusion in section 6.

2.0 RELATED WORK

This section describes the two main bodies of previous work related to our study and these are the correlations of performance metrics and the IR evaluation methods. As regards the correlations of performance metrics, [6] investigated several performance metrics including F-measure and the four-fold point coefficient. The author found that F-measure and the four-fold point coefficient have similar properties. In another study, [7] investigated the correlation of average precision with R-precision. The authors provided a geometric argument that showed that the area under the precision-recall curve could be approximated by both the average precision and R-precision. Using this approximation, the author explained the correlation between the two metrics. [8] investigated the finding of one highly relevant document and the author found that O-measure and normalized weighted reciprocal rank are highly correlated. In the same year, [9] explored the correlations between at least 20 performance metrics and the author found that Q-measure and average precision are highly correlated. In a separate study, [10] investigated several performance metrics and their reflection on user search performance. They found that metrics either focused on high precision in an answer list or captured a broad summary and they suggested that the relative performance of retrieval systems may depend on the group of measures used in the evaluation. In the following year, [11] explored the correlations between precision, recall and fallout. The author showed that for recall, precision evolves by following a concave decreasing function. In addition, the author showed that concerning fallout, recall follows a concave increasing function. In a later study, [12] investigated correlations between at least 120 metrics. Arising from this investigation, the authors grouped the performance metrics into 7 clusters depending on their correlations. In another study, [13] explored the disagreement between at least 10 performance metrics and the authors

investigated differences that need to be considered when choosing metrics to be used for IR evaluation. More recently, [4] investigated the correlations of at least 20 performance metrics and they found a high correlation of R-Precision with bpref, mean average precision (MAP) and nDCG performance metrics. In addition, the authors identified a strong correlation between reciprocal rank (RR) and RBP.

Turning now to the previous research on IR evaluation, there have been several proposed methods to reduce costs while performing evaluations. These include methods that concern the selection of subsets of topics for retrieval evaluation [14]–[21] and methods for the efficient creation of test collections [22]–[24]. In addition, other proposed methods relate to the inference of relevance judgments [25]–[28]. Recent research investigated the use of machine learning-based methods to predict high-cost performance metrics and for these methods, the predictors were the low-cost performance metrics [4]. Though the authors [4] demonstrated strong ranked correlations of the predictions of the RBP performance metrics, they reported the inaccurate predictions of other high-cost performance metrics such as the precision and nDCG especially in the case where the low-cost performance metrics were computed at the cut-off depths $d \leq 30$ of documents [4]. Also, a close inspection of this recent proposal [4] reveals that the authors used one set of test collections to compute the performance metrics employed for the training of regression models for their approach, but used a different set of test collections to compute performance metrics employed for testing purposes. The use of different sets of test collections to compute performance metrics for training and testing purposes was done without a prior investigation to ascertain whether such use was appropriate for this research. Such use of sets of test collections would only be permissible in the case where there is a high similarity in performance metrics computed from the sets of test collections. Our study bridges this gap by seeking to determine whether machine learning models trained using performance metrics computed using one set of test collections may be used to predict high-cost performance metrics computed using a different set of test collections. In addition, this study also provides suggestions to improve existing approaches to predict the high-cost performance metrics. Lastly, this study also proposes two approaches that predict the precision and nDCG high-cost performance metrics, particularly at the cut-off depths $d \leq 30$ documents for the low-cost performance metrics.

3.0 MATERIALS AND METHODS

This section describes the data collection, the performance metrics, the baseline method used for our study as well as the metrics used for evaluating our results. In addition, this section describes the experimental methodology used to investigate the predictability of performance metrics and the proposed approaches for predicting the precision and nDCG high-cost performance metrics while using the low-cost performance metrics computed at the cut-off depths $d \leq 30$. The experiments presented in this paper employ topic scores of performance metrics.

3.1 Data collection and performance metrics

Similar to recent research [4], this study employed relevance judgments and runs of TREC 2000-2001, 2013-2014 Web Tracks (WT) [29]–[32] and TREC 2004 Robust Track (RT) [33]. Also, similar to the previous study [4], the selected performance metrics were precision [36], nDCG [34], RBP [35], binary preference [38], inferred average precision (infAp) [39] and expected reciprocal rank (ERR) [37]. Furthermore, fmeasure and recall [36] were also utilized in this study due to their high correlation with both precision and nDCG. Each of the low-cost performance metrics were computed at the cut-off depths of between 10 and 30 documents (i.e $10 \leq d \leq 30$). The precision and nDCG high-cost performance metrics were computed at the cut-off depths $d = 100$ and $d = 1000$ documents.

3.2 Baseline method

The baseline method for our study is reported in [4]. This method uses linear regression and system scores of the low and high-cost performance metrics. The suitable subset of predictors was chosen from the power set of the low-cost performance metrics by using the least root mean square error metric during the training of the regression models at several cut-off depths of documents.

3.3 Performance analysis

In this this study, several metrics have been used to evaluate the results. To start with, since the investigation of the predictability of performance metrics concern machine learning classification, recall, precision, accuracy and area under the curve (AUC) metrics were employed. The results were evaluated at various cut-off depths $d \leq 30$ of low-cost performance metrics. The recall, precision and accuracy evaluation metrics are represented by equations (1) to (3) below.

$$Accuracy@k = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision@k = \frac{TP}{TP + FP} \quad (2)$$

$$Recall@k = \frac{TP}{TP + FN} \quad (3)$$

Where TP means correctly classified as belonging to a particular class of performance metrics, TN means correctly classified as not belonging to a particular class of performance metrics, FP means wrongly classified as belonging to a particular class of performance metrics and FN means wrongly classified as not belonging to a particular class of performance metrics.

Lastly, we turn to AUC. This metric is an important measure of performance for classification models. It represents the measure of the separability of the classes of the low-cost performance metrics.

Regarding the proposed approaches, like previous research [4], the focus is the ranked correlations of the predictions of the high-cost performance metrics. Therefore to evaluate the performance of these approaches, the ranked correlation coefficient should be used. Hence, the use of the Kendall's tau (see expression 4) in this study and this correlation coefficient is defined by:

$$Kendall'sTau = \frac{C - D}{\frac{1}{2}n(n-1)} \quad (4)$$

Where C is the number of concordant pairs while D is the number of discordant pairs and n is the total number of samples in the dataset.

3.4 Experimental methodology to investigate the predictability of performance metrics

This section describes the methodology for the experiment that seeks to determine whether machine learning models trained using performance metrics computed by employing one set of test collections may be used to predict the high-cost performance metrics computed using other sets of test collections. Before describing the details of the experimental methodology, let's first look at this scenario, suppose there are three test collections A, B and C, and assume the low and high-cost performance metrics are computed using these test collections. The question is: if a regression model is trained with the performance metrics computed from test collections A and B and later used to predict the high-cost performance metrics computed using test collection C, how predictable will these high-cost performance metrics be? This scenario depicts the existing research [4] where a regression model was trained using performance metrics computed using one set of test collections and predictions were made for the high-cost performance metrics computed using other test collections. The predictability of high-cost performance metrics is measured by how linearly separable the class of one set of low-cost performance metrics is from another. If the linear separability is high, then the predictability of the high-cost performance metrics is low because high linear separability signifies low similarity between the performance metrics in the different classes. For a thorough discussion on the identification and resolution of dissimilarity of classes of data in machine learning research, refer to [40].

In our investigation, there are three classes of the low-cost performance metrics, namely class 1, class 2 and class 3. Turning now to the descriptions of these classes, class 1 represents the training set and comprises low-cost performance metrics computed using relevance judgments and runs of TREC 2000-2001 WT and TREC 2004 RT. Secondly, class 2 represents the first test set and comprises low-cost performance metrics computed using relevance judgments and runs from TREC 2012 WT. Lastly, class 3 represents the second test set and comprises low-cost performance metrics computed using relevance judgments and runs from TREC 2013 WT.

The basic idea is to iteratively merge the training set with each test set and train and use a classifier to check for linear separability of the two classes. For instance, merge the performance metrics in class 1 with those from class 2,

train and use a classifier to check for the linear separability of the low-cost performance metrics of the two classes by reporting the results of the classification evaluation metrics.

As regards the experimental methodology, outlined below are the series of steps:

1. Step 1: Using relevance judgments and runs of test collections, compute low-cost performance metrics at the cut-off depths d such that $10 \leq d \leq 30$ and the high-cost performance metrics at the cut-off depths j such that j = 100 and j = 1000 respectively. The expression 5 below shows a set D_s comprised of these computed performance metrics.
$$D_s = \{infAp@d, precision@d, Recall@d, ERR@d, RBP@d, nDCG@d, Bpref@d, fmeasure@d, P@j, nDCG@j\} \quad (5)$$
2. Step 2: Create three data sets. The first data set is the training set which comprises performance metrics computed using TREC Tracks 2000-2001 WT and TREC 2004 RT. The next two data sets are test sets comprising performance metrics computed using TREC 2012-2013 WT respectively. The expression 6 shows the training set as a subset comprised of low-cost performance metrics computed using TREC 2000-2001 WT and TREC 2004 RT. Also, two test sets are shown as subsets comprising low-cost performance metrics computed using TREC 2012-2013 WT respectively.
$$trainingset \leftarrow \sigma_{trec \in 2000WT, 2001WT, 2004WT}(D_s) \quad (6)$$
$$testset_{trec2012} \leftarrow \sigma_{trec \in 2012WT}(D_s)$$
$$testset_{trec2013} \leftarrow \sigma_{trec \in 2013WT}(D_s)$$
3. Step 3: Assign the low-cost performance metrics in the three data sets described in the previous step to 3 distinct classes. The expression 7 shows the assignment of classes to training and test sets.
$$trainingset \leftarrow \sigma_{trec \in 2000WT, 2001WT, 2004RT}(D_s) append\ class1 \quad (7)$$
$$testset_{trec2012} \leftarrow \sigma_{trec \in 2012WT}(D_s) append\ class2$$
$$testset_{trec2013} \leftarrow \sigma_{trec \in 2013WT}(D_s) append\ class3$$
4. Step 4: Iteratively, merge the low-cost performance metrics for the training set with those for each test set and build a classifier for each represented cut-off depth d such that $10 \leq d \leq 30$ of the low-cost performance metrics. Use the 10-fold cross validation during the build of the classification models and to address the problem of class imbalance, employ the under-sampling technique at this step.
5. Step 5: Using the classifier developed in the previous step, measure how linearly separable the low-cost performance metrics for each test set are from those of the training set. Evaluate the performance of the classifiers using the metrics for evaluating machine learning classification models described above.

3.5 The proposed approaches to predict the high-cost performance metrics

This section presents the two proposed approaches to predict the high-cost precision and nDCG performance metrics. The first approach is also known as proposed-LR employs linear regression at all the cut-off depths of the low-cost performance metrics while the second approach also known as the proposed-RF employs the random forest models. Both approaches employ performance metrics and the transformed set of performance metrics. However, the difference between them lies in the procedures for selecting the most suitable performance metrics to be used in the machine learning models. Since both approaches employ the transformed set of performance metrics, the method of generating the transformed set of low and high-cost performance metrics is described first. This is followed by detailed descriptions of the two proposed approaches.

3.5.1 Generation of new sets of performance metrics

A recent study [4] reported that concerning correlations with other performance metrics, precision is least correlated. In addition, [4] proposed a method that demonstrated the inaccurate predictions of the high-cost performance metrics such as precision and nDCG especially in the case where the low-cost performance metrics were computed at the cut-off depths of at most 30 documents. A close look at the performance metrics used in this study has revealed that they have skewed distributions and extreme scores. Therefore, in this research, these two issues are addressed by generating new sets of topic scores. According to [41], the usage of appropriate mathematical functions addresses the skew in predictors of regression models. By addressing the skewed distributions and extreme scores for the performance metrics represented in this study, there is an observed increase in correlations and the gain in information between high and low-cost performance metrics. To generate the new set of topic scores, the mathematical functions used are the yeo-johnson, exponential, cube root and logarithmic functions [41].

A demonstration of how the new sets of topic scores of low and high-cost performance metrics are generated is provided with the aid of expressions (8) to (11). Regarding expression 9, P_20all represents a vector of topic scores of the precision performance metrics computed at the cut-off depth d=20 of documents. Now P_20all is used as input into the logarithmic function and the output is a vector P_20all_log containing the transformed set of topic scores. Expressions 8 and 10 show the application of the cube root and exponential functions on the nDCG and infAP performance metrics respectively.

$$\text{ndcg\_10}all \rightarrow \sqrt[3]{\text{ndcg\_10all}} \rightarrow \text{ndcg\_10all\_cbrt} \tag{8}$$

$$\text{P\_20all} \rightarrow log(\text{P\_20all} + 1) \rightarrow \text{P\_20all\_log} \tag{9}$$

$$\text{infap\_30all} \rightarrow exp\left(\text{infap\_30all}, \frac{1}{2}\right) \rightarrow \text{infap\_30all\_exp} \tag{10}$$

$$\text{rbp\_25\_all} \rightarrow \frac{\Lambda}{1-\Lambda}(\text{rbp\_25all}) \rightarrow \text{rbp\_25all\_yj} \tag{11}$$

By employing these transformations on the topic scores computed using the function definitions of the performance metrics, new sets of topic scores of low-cost performance metrics are generated that are more informative and better correlated to the high-cost performance metrics. The expression $\Lambda/(1-\Lambda)$ is referred to as the yeo-johnson technique [41]. The next section presents a description of the first proposed approach.

3.5.2 The linear regression based proposed approach (proposed LR)

This section presents the linear regression based proposed approach also called proposed LR to predict the high-cost precision and nDCG performance metrics. The predictors are low-cost performance metrics computed using cut-off depths of up to 30 documents. The detailed descriptions of each step for this approach are presented in Table 1 below.

Table 1: The steps for the linear regression based approach to predict the high-cost performance metrics

Step #	Description
1	Using relevance judgments and runs of test collections presented in section 3.1, compute low and high-cost performance metrics to form a matrix M_s
2	Create new matrices by applying the procedure presented in section 3.5.1 to matrix Ms
3	Create a new matrix M_low_cost by combining the low-cost performance metrics in the matrices from step1 and step 2.
4	Create a new matrix M_high_cost by combining the high-cost performance metrics in the matrices from step1 and step 2.
5	Generate pairs of the matrix M_low_cost with every vector in the matrix M_high_cost e.g < P_100all, M_low_cost >
6	Using pearson correlation, information gain, linear regression model training, prediction on validation, tau correlation computation and comparison, identify the most suitable pair from similar pairs from step 5 at every cut-off depth of low-cost performance metrics. The vectors of high-cost performance metrics form similar pairs if they are generated from the same vector of high-cost performance metric from matrix M_s computed in step 1.
7	At each cut-off depth d of the low-cost performance metrics, train regression models using the chosen suitable pairs from the previous step to produce approximation regression functions. One example of the regression approximation function is Equation (12) below: $\overline{P@1000} = 0.131 * \text{P\_10all} - 0.264 * \text{fmeasure\_10all\_log}$ (12)
8	Using approximation functions, make predictions of the high-cost performance metrics and ranked correlations are computed. Lastly, the results of the ranked correlations for the proposed approaches are compared with the baseline

3.5.3 The tree regression based proposed approach (proposed RF)

This section presents the tree regression based proposed approach also called proposed RF to predict the high-cost precision and nDCG performance metrics. Also, the predictors are low-cost performance metrics computed using

cut-off depths of up to 30 documents. The detailed descriptions of each step for this approach are presented in Table 2 below.

Table 2: The steps for the tree regression based approach to predict the high-cost performance metrics

Step #	Description
1	Using relevance judgments and runs of test collections presented in section 3.1, compute low and high cost performance metrics to form a matrix M_s
2	Create new matrices by applying the procedure presented in section 3.5.1 to matrix Ms
3	Create a new matrix M_low_cost by combining the low-cost performance metrics in the matrices from step1 and step 2.
4	Create a new matrix M_high_cost by combining the high-cost performance metrics in the matrices from step1 and step 2.
5	Generate pairs of the matrix M_low_cost with every vector in the matrix M_high_cost e.g < P_100all, M_low_cost >
6	Using each pair from previous step, train the classification and regression tree models, perform predictions on the validation set and compute and compare tau correlations. Select the suitable pair from similar pairs from step 5 at every cut-off depth of low-cost performance metrics.
7	At each cut-off depth k of the low-cost performance metrics, train the random forest regression tree models using the chosen pairs from the previous step.
8	Using the random forest regression tree models, make predictions of the high-cost performance metrics and compute ranked correlations. Lastly, the results of the ranked correlations for the proposed approaches are compared with the baseline.

These proposed approaches were implemented in python 3.7 and the anaconda environment 4.10.3 and their results for the ranked correlations of predictions of the high-cost performance metrics are presented in the following section.

4.0 RESULTS

This section presents two sets of results. The first set concerns the experiment that sought to determine whether machine learning models trained using performance metrics computed using one set of test collections may be used to predict high-cost performance metrics computed by employing a different set of test collections. The second set of results relates to the performance of the proposed approaches and how they compare with the baseline method.

4.1 Results for the investigation of the predictability of performance metrics

This section presents the results for the experiment that sought to determine whether machine learning models trained using performance metrics computed using one set of test collections may be used to predict high-cost performance metrics computed by employing a different set of test collections. Recall that the low-cost performance metrics were computed at the cut-off depths of between 10 and 30 documents while the high-cost performance metrics at the cut-off depths of 100 and 1000 documents. As earlier mentioned, the training set comprised performance metrics computed using the relevance judgments and runs from test collections TREC 2000-2001 WT and TREC 2004 RT. The test sets constituted performance metrics computed using the relevance judgments and runs of the test collections TREC 2012 and 2013 WT. This investigation aimed to find the similarity between the performance metrics of the training and each of the test sets. The high values of the classification evaluation metrics meant low similarity.

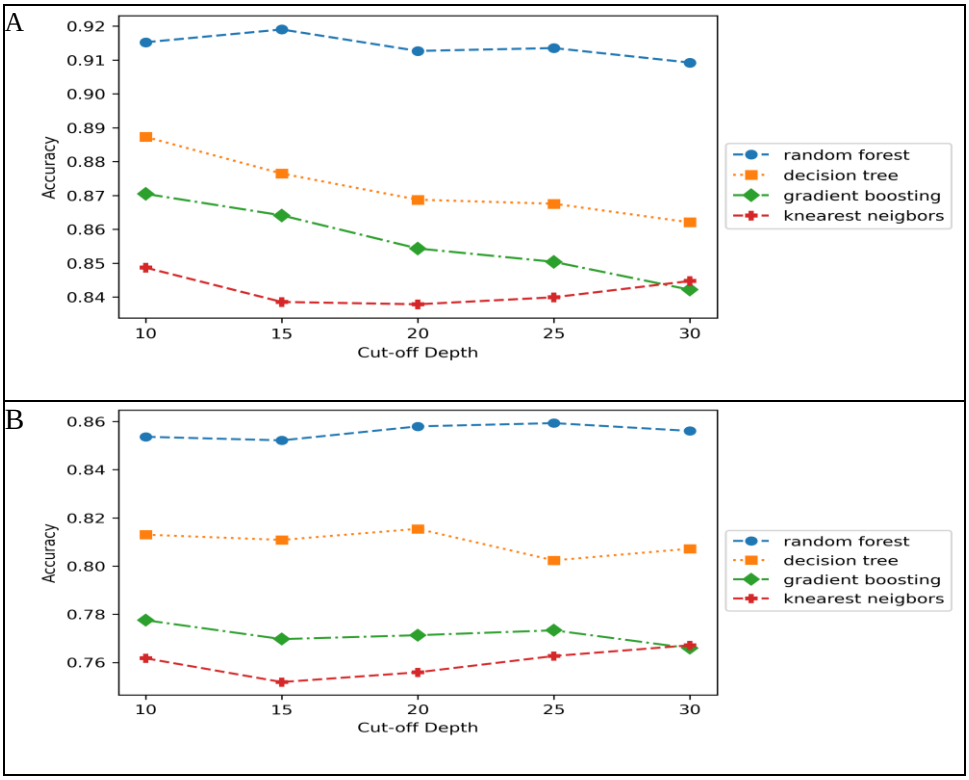

Fig. 1: The Accuracy results for the classification between sets of low-cost performance metrics representing training and test sets. The training set comprises of the low-cost performance metrics computed using test collections TREC 2000-2001 WT and TREC 2004 RT.

Fig.1 presents the results of the accuracy for the classification of the sets of the low-cost performance metrics representing training and test sets using several classifiers. Fig.1(A) shows the accuracy of the different classifiers when the test set constituted performance metrics computed using the relevance judgments and runs from TREC 2012 WT while Fig.1(B) highlights the accuracy of the different classifiers when the test set comprised performance metrics computed using the relevance judgments and runs from TREC 2013 WT. A close inspection of Fig.1(A) shows that at the cut-off depth d = 10, the accuracy of the predictions of the classifiers was in the range from 84 to 92 percent. The predictions of the random forest (0.9152), decision tree (0.8873), gradient boosting (0.8705) and knearest neighbor (0.8487) classifiers were at least 84 percent. As regards Fig.1(B), the results show that at the cut-off depth d = 10, the accuracy of the predictions of the classifiers was in the range from 76 to 86 percent. The predictions of the random forest (0.8537),decision tree(0.8130),gradient boosting(0.7775) and knearest neighbor(0.7617) classifiers were at least 76 percent. Concerning the cut-off depth d = 20, the accuracy of the predictions of the classifiers was in the range of 75 to 86 percent. The predictions of the random forest(0.8580),decision tree(0.8155),gradient boosting(0.7713) and knearest neighbor(0.7560) classifiers were at least 75 percent. The next result concerns the precision evaluation metric.

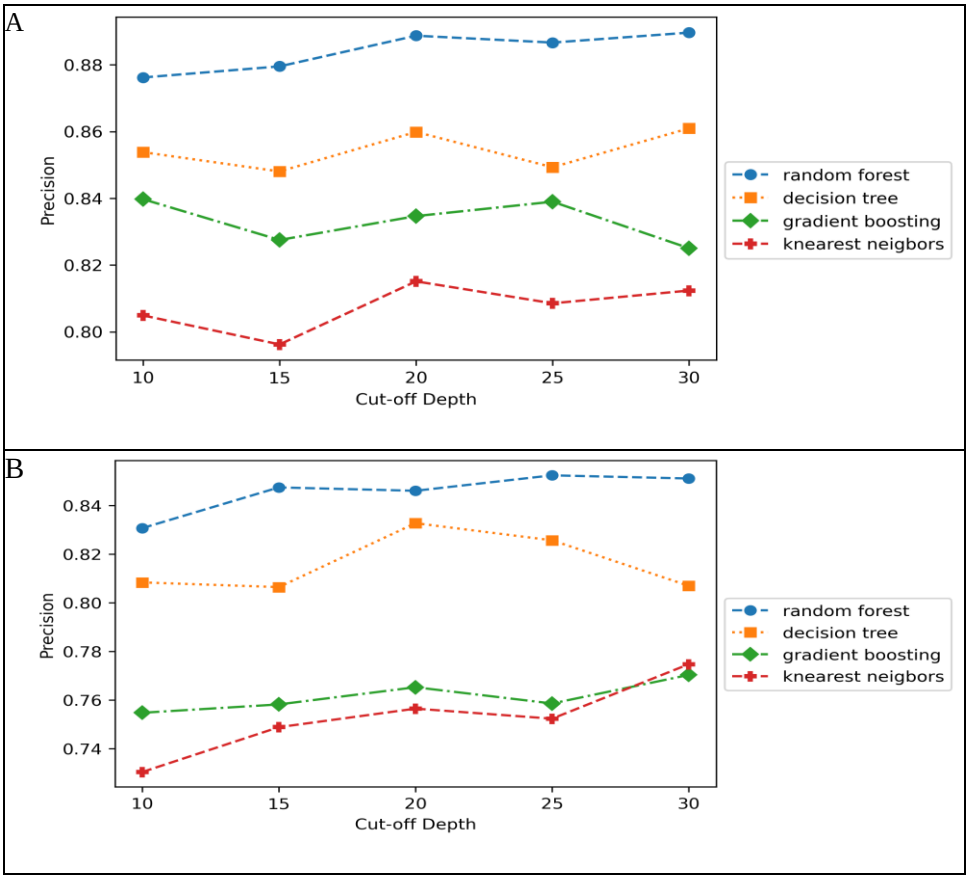

Fig. 2: The precision results for classification between sets of low-cost performance metrics representing training and test sets. The training set comprises of low-cost performance metrics computed using test collections TREC 2000-2001 WT and TREC 2004 RT.

Fig.2 presents the results of the precision for the classification of the sets of low-cost performance metrics representing training and test sets using several classifiers. Fig.2(A) shows the precision of the different classifiers when the test set constituted performance metrics that were computed using the relevance judgments and runs of TREC 2012 WT while Fig.2(B) shows the precision of the different classifiers when the test set comprised performance metrics that were computed using the relevance judgments and runs of TREC 2013 WT. As regards Fig.2(A), at the cut-off depth d = 20, the precision of the predictions of the classifiers was in the range of 81 to 89 percent. The predictions of the random forest (0.8887), decision tree(0.8599),gradient boosting(0.8347) and knearest neighbor(0.8151) classifiers were at least 81 percent. Looking at Fig.2(B), it is apparent that at the cut-off depth d = 10, the precision of the predictions of the classifiers was in the range of 73 to 84 percent. The predictions of the random forest(0.8307),decision tree(0.8084),gradient boosting(0.7547) and knearest neighbor(0.7303) classifiers were at least 73 percent.

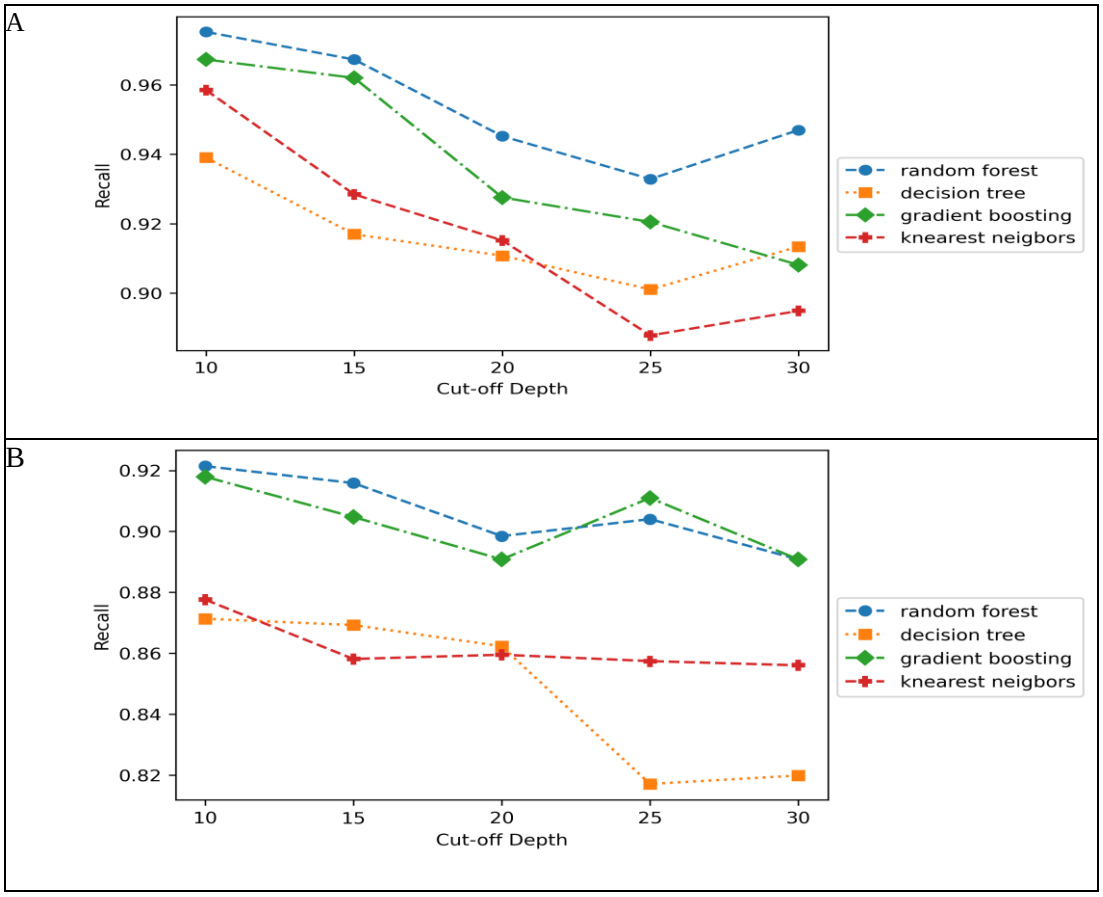

Fig. 3: The recall results for the classification of the sets of low-cost performance metrics representing the training and test sets. The training set comprises of low-cost performance metrics computed using test collections TREC 2000-2001 WT and TREC 2004 RT.

Fig.3 presents the results of the recall for the classification of the sets of low-cost performance metrics representing training and test sets using several classifiers. Fig.3(A) shows the recall of the different classifiers when the test set constituted performance metrics computed using the relevance judgments and runs from TREC 2012 WT while Fig.3(B) displays the recall of the different classifiers when the test set comprised performance metrics computed using the relevance judgments and runs from TREC 2013 WT. It is clear from Fig.3(A) that at the cut-off depth d = 10, the recall of the predictions of the classifiers was in the range of 93 to 98 percent. The predictions of the random forest (0.9753), decision tree (0.9391), gradient boosting (0.9673) and knearest neighbor(0.9585) classifiers were at least 93 percent. Regarding Fig.3(B), it is apparent that at the cut-off depth d = 15, the recall of the predictions of the classifiers was in the range from 85 to 92 percent. The predictions of the random forest (0.9159), decision tree (0.8693), gradient boosting (0.9047) and knearest neighbor (0.8581) classifiers were at least 85 percent.

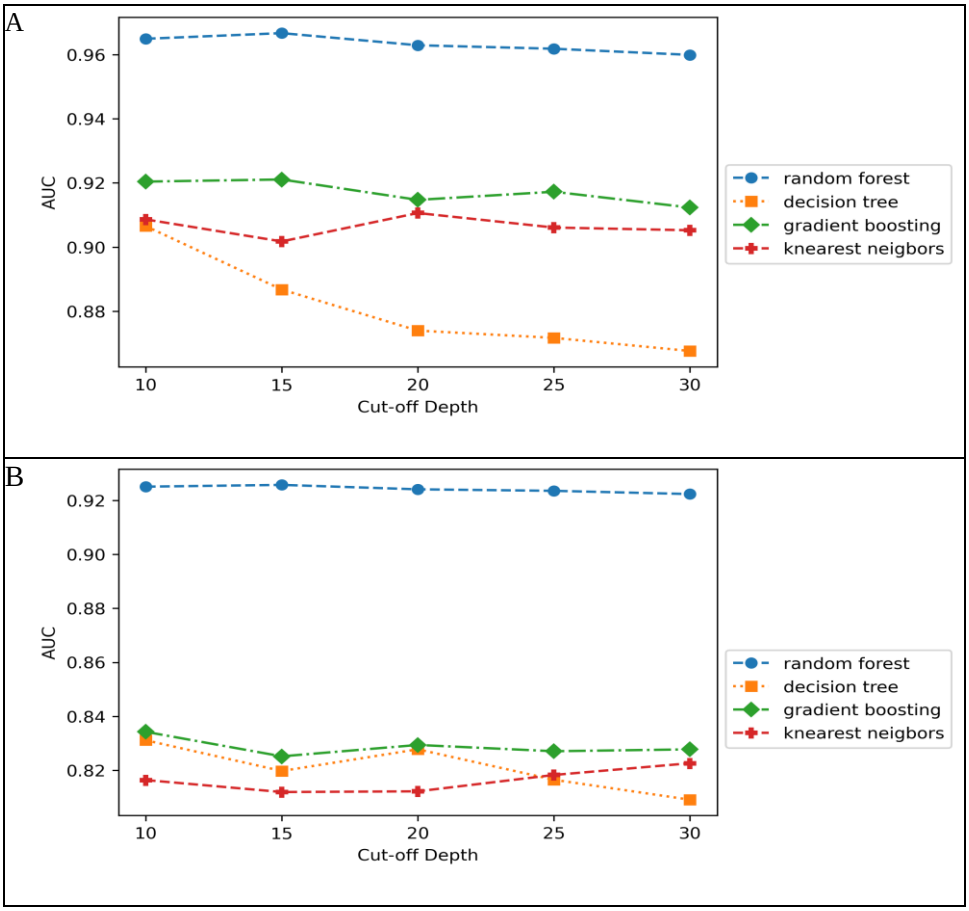

Fig. 4: The area under the curve (AUC) results for classification between sets of the low-cost performance metrics representing training and test sets. The training set comprises of low-cost performance metrics computed using test collections TREC 2000-2001 WT and TREC 2004 RT.

Fig.4 highlights the results of the AUC for the classification of the sets of low-cost performance metrics representing training and test sets using several classifiers. Fig.4(A) displays the AUC of the different classifiers when the test set constituted performance metrics computed using the relevance judgments and runs from TREC 2012 WT while Fig.4(B) shows the AUC of the different classifiers when the test set comprised performance metrics computed using the relevance judgments and runs from TREC 2013 WT. As regards Fig.4(A), the result shows that at the cut-off depth d = 10, the AUC of the predictions of the classifiers was in the range of 90 to 97 percent. The predictions of the random forest (0.9649), decision tree (0.9065), gradient boosting (0.9204) and knearest neighbor (0.9086) classifiers were at least 90 percent. Turning now to Fig.4 (B), the result shows that at the cut-off depth d = 10, the AUC of the predictions of the classifiers was in the range of 81 to 93 percent. The predictions of the random forest (0.9251), decision tree (0.8311), gradient boosting (0.8342) and knearest neighbor (0.8164) classifiers were at least 81 percent.

4.2 The results of the proposed and baseline approaches

This section presents the results of the ranked correlations of the predictions of the high-cost performance metrics of the proposed approaches. The results also include comparisons of the ranked correlations for the proposed approaches with the baseline. The presentation of results begins with the ranked correlations for the predictions of the nDCG high-cost performance metrics at cut-off depths d=100 (i.e nDCG@100) and d=1000 (nDCG@1000). What follows are the results of the ranked correlations for the predictions of the precision high-cost performance metrics at cut-off depths d=100 (i.e P@100) and d=1000 (i.e P@1000). For all the presented results, the training set comprises performance metrics computed using test collections TREC 2000-2001 WT and TREC 2004 RT and the predictors are low-cost performance metrics computed at cut-off depths $d \leq 30$ documents.

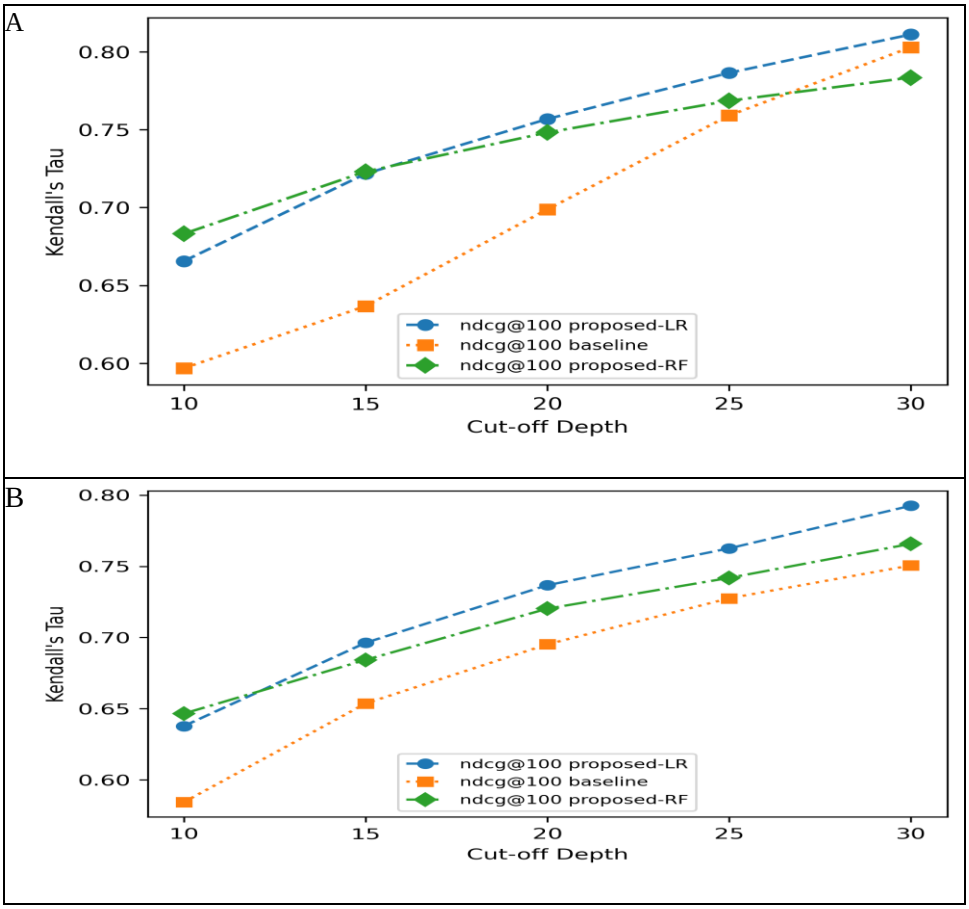

Fig.5: The Kendall's tau correlation for the prediction of nDCG@100 high-cost performance metrics while using low-cost performance metrics computed at cut-off depths $10 \leq d \leq 30$ documents.

Fig.5 shows Kendall's tau correlations for the prediction of nDCG@100 using low-cost performance metrics (i.e predictors) computed at cut-off depths $d \leq 30$ documents.Fig.5(A) presents Kendall's tau correlations for the predictions of nDCG@100 using the test set comprised performance metrics that were computed using relevance judgments and runs from TREC 2013 WT while Fig.5(B) shows Kendall's tau correlations for the predictions of nDCG@100 using the test set constituted of performance metrics that were computed by utilizing relevance judgments and runs from TREC 2014 WT. It is clear from Fig.5(A), that the proposed approaches had better Kendall's tau correlations than the baseline for the predictions of nDCG@100 at various cut-off depths of the predictors. For example, at the cut-off depth $d = 15$, Kendall's tau correlation for the predictions of proposed-LR(0.7217) and proposed-RF(0.7231) were higher than the baseline(0.6367) by at least 13.35 percent. Looking at Fig. 5B, the proposed approaches had better Kendall's tau correlations than the baseline for predictions of nDCG@100 at all the cut-off depths of the predictors. For example, at the cut-off depth $d = 10$, Kendall's tau correlation for the predictions of proposed-LR(0.6378) and proposed-RF(0.6465) exceeded the baseline(0.5843) by at least 9.16 percent.

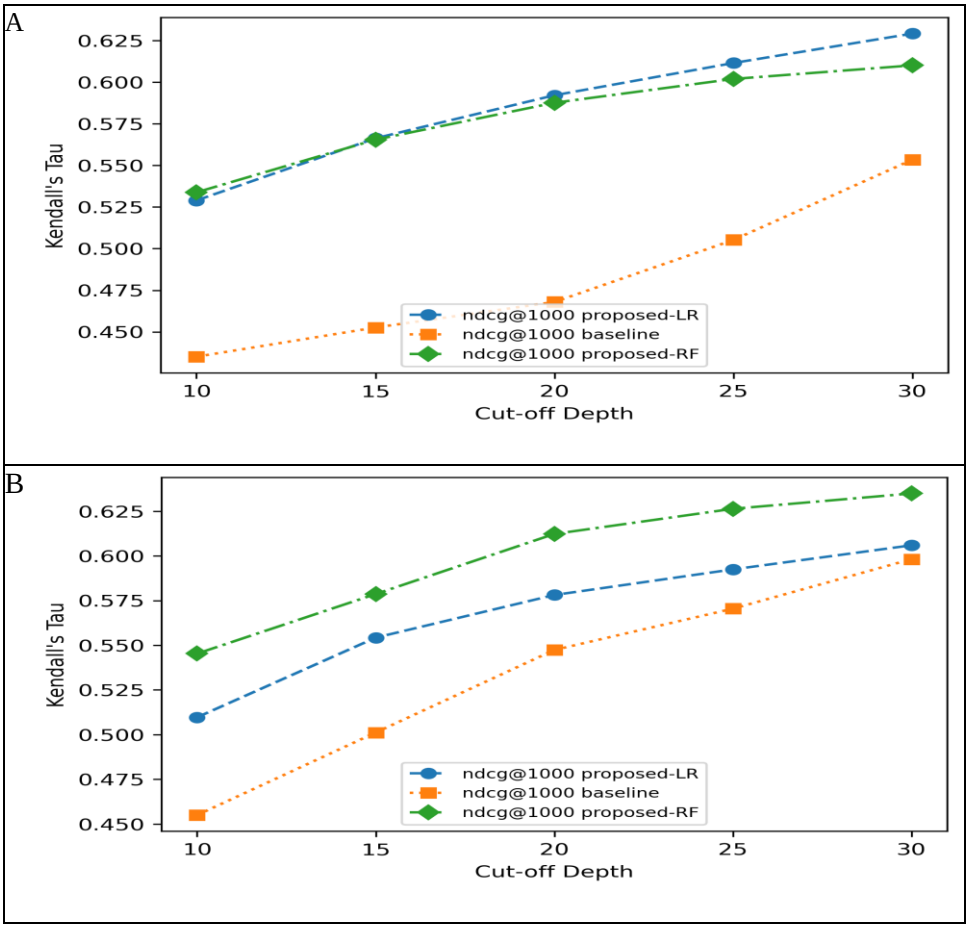

Fig. 6: The Kendall's tau correlation for the prediction of nDCG@1000 high-cost performance metrics while using low-cost performance metrics computed at cut-off depths $10 \leq d \leq 30$ documents.

Fig.6 shows Kendall's tau correlations for the prediction of nDCG@1000 using predictors computed at cut-off depths $d \leq 30$ documents. Fig.6(A) presents Kendall's tau correlations for the predictions of nDCG@1000 using the test set comprised of performance metrics that were computed using relevance judgments and runs of TREC 2013 WT while Fig.6(B) shows Kendall's tau correlations for the predictions of nDCG@1000 using the test set constituted of performance metrics that were computed using relevance judgments and runs of TREC 2014 WT. Regarding Fig.6(A), the proposed approaches had better Kendall's tau correlations than the baseline for predictions of nDCG@1000 at all the cut-off depths of the predictors. For example, at the cut-off depth $d = 15$, Kendall's tau correlation for the predictions of proposed-LR(0.5663) and proposed-RF(0.5655) were more than the baseline(0.4526) by at least 24.94 percent. Turning now to Fig.6(B), the result shows that the proposed approaches had better Kendall's tau correlations than the baseline for predictions of nDCG@1000 at all the cut-off depths of predictors. For example, at the cut-off depth $d = 10$, Kendall's tau correlation for the predictions of proposed-LR(0.5097) and proposed-RF(0.5453) exceeded the baseline(0.4550) by at least 12.02 percent.

Fig.7 shows Kendall's tau correlations for the prediction of P@100 using predictors computed at cut-off depths $d \leq 30$ documents. Fig.7(A) presents Kendall's tau correlations for the predictions of P@100 using the test set comprised of performance metrics that were computed by employing relevance judgments and runs of TREC 2013 WT while Fig.7(B) shows Kendall's tau correlations for the predictions of P@100 using the test set constituted of performance metrics that were computed using relevance judgments and runs of TREC 2014 WT. As regards Fig.7(A), the proposed approaches had better Kendall's tau correlations than the baseline for predictions of precision@100 at all the cut-off depths of the predictors. For example, at the cut-off depth $d = 10$, Kendall's tau correlation for the predictions of proposed-LR(0.6591) and proposed-RF(0.6599) exceeded the baseline(0.4520) by at least 45.82 percent. Turning now to Fig. 7(B), the result shows that the proposed approaches had better Kendall's tau correlations than the baseline for predictions of precision@100 at all the cut-off depths of the predictors. For example, at the cut-off depth $d = 20$, the Kendall's tau correlation for the predictions of proposed-LR(0.7102) and proposed-RF(0.7343) were higher than the baseline(0.5372) by at least 32.20 percent.

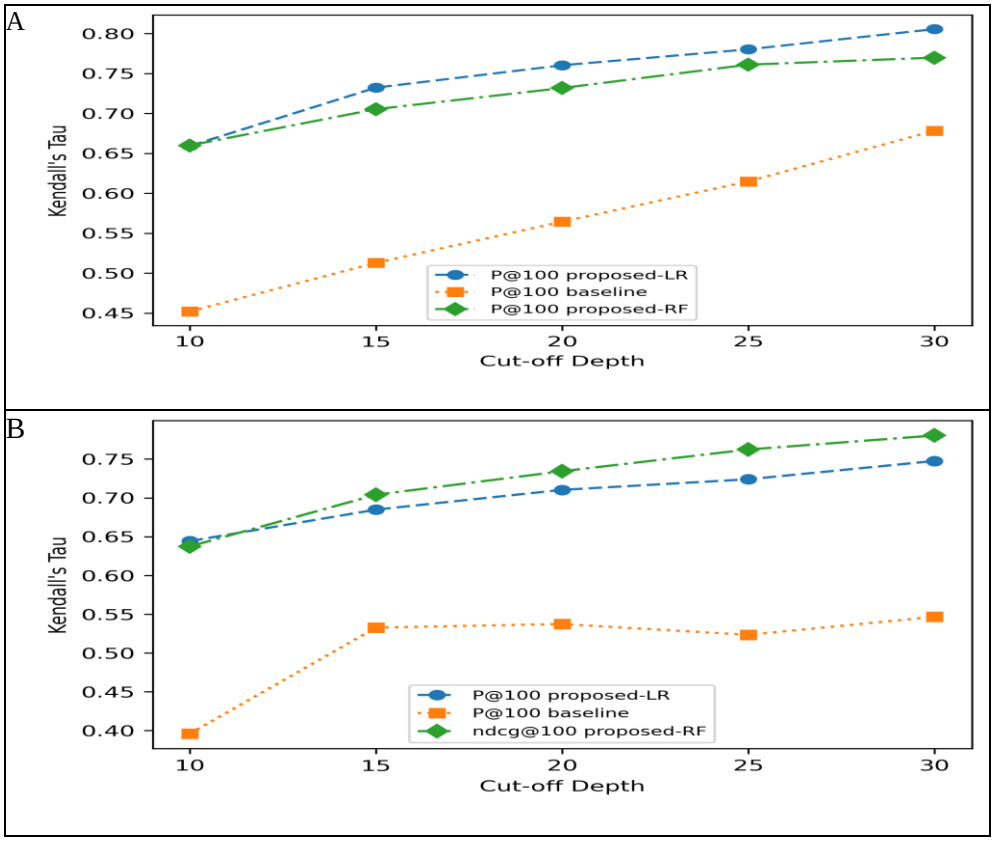

Fig. 7: The Kendall's tau correlation for the prediction of P@100 high-cost performance while using low-cost performance metrics computed at cut-off depths $10 \leq d \leq 30$ documents.

Fig.8 shows Kendall's tau correlations for the prediction of P@1000 using predictors computed at cut-off depths $d \leq 30$ documents. Fig.8(A) presents Kendall's tau correlations for the predictions of P@1000 using the test set comprised of performance metrics that were computed by utilizing relevance judgments and runs of TREC 2013 WT while Fig.8(B) shows Kendall's tau correlations for the predictions of P@1000 using the test set constituted of performance metrics that were computed by utilizing relevance judgments and runs of TREC 2014 WT. Regarding Fig.8(A) below, the results show that the proposed approaches had better Kendall's tau correlations than the baseline for predictions of precision@1000 at all the cut-off depths of the predictors. For example, at the cut-off depth $d = 25$, Kendall's tau correlation for the predictions of proposed-LR(0.6352) and proposed-RF(0.6173) exceeded the baseline(0.3730) by at least 65.5 percent. Also, a close inspection of Fig. 8(B) shows that the proposed approaches had better Kendall's tau correlations than the baseline for predictions of precision@1000 at all the cut-off depths of the predictors. For example, at the cut-off depth $d = 15$, Kendall's tau correlation for the predictions of proposed-LR(0.6245) and proposed-RF(0.5921) were higher than the baseline(0.5334) by at least 10.61 percent.

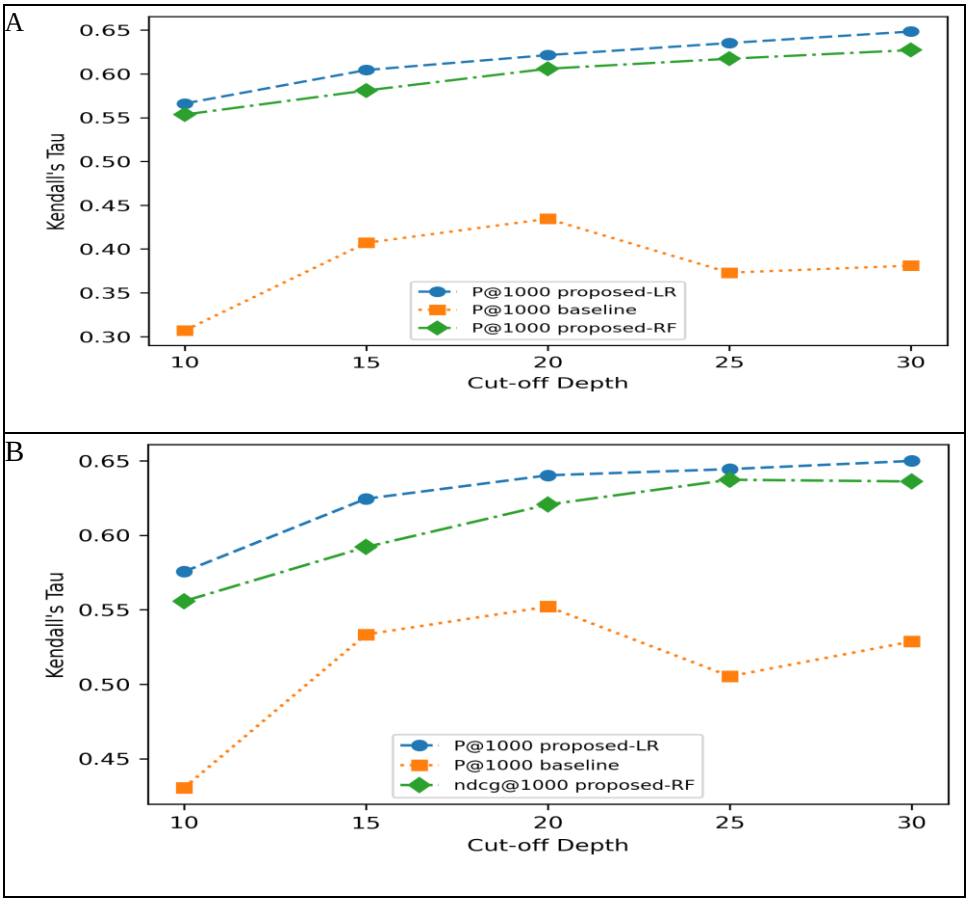

Fig. 8: The Kendall's tau correlation for the prediction of P@1000 high-cost performance metrics while using low-cost performance metrics computed at cut-off depths $10 \leq d \leq 30$ documents.

5.0 DISCUSSION

In this section, the discussion is presented for the results reported in the previous section. Section 5.1 presents the discussion for the results reported in section 4.1 and then follows section 5.2 that presents the discussion for the results reported in section 4.2.

5.1 The discussion for the investigation of the predictability of performance metrics

Recall that the investigation on the predictability of performance metrics used classifiers of several machine learning models whose results were presented using the accuracy, precision, recall, and the area under the curve (AUC) classification metrics. Regarding the accuracy evaluation metric, all the classifiers distinguished the performance metrics of training sets from test sets with accuracies ranging between 75.19 and 91.90 percent at the various cut-off depths of documents. Also, concerning the precision evaluation metric, all the classifiers distinguished the performance metrics of training sets from those of test sets with precision between 73.03 and 88.96 percent at the various cut-off depths of documents. Furthermore, as regards the recall evaluation metric, all the classifiers distinguished the performance metrics of training sets from test sets with recall between 81.47 and 97.53 percent at the various cut-off depths of documents. Lastly, concerning the AUC evaluation metric, all the classifiers distinguished the performance metrics of training sets from those in test sets with the AUC ranging between 77.67 and 96.67 percent at the various cut-off depths of documents. It is clear from these results that the linear separability between the low-cost performance metrics of training and test sets is very high. This means that the distributions of the performance metrics computed using one set of test collections differs from the distributions of performance metrics computed using other sets of test collections. The implication of this is that if machine learning predictive models are trained using performance metrics computed using relevance judgments and runs from one set of test collections and later used to predict performance metrics relating to other test collections, the predictability of these performance metrics are likely to be low. This may partly explain why the results in existing research [4] are inaccurate. Therefore, to ensure improved results of the predictions of the high-cost performance

metrics, the following suggestions are provided:

1. Use the solutions from machine learning research to address the difference in distributions that exist in performance metrics computed using the relevance judgments and runs from different test collections. Machine learning research has proposed several solutions to address this issue [40]. First, they propose to employ the KL-divergence technique to make adjustments to the values of the performance metrics in the training and test sets to ensure high similarity. Also, machine learning research proposed the use of logistic regression to adjust the parameters of the regression models.

2. A test collection could be nominated as the one about which the distributions of performance metrics computed using other test collections would be adjusted with respect to the data set shift before model training and prediction of the high-cost performance metrics.

3. Generate new sets of topic scores of performance metrics to be used in predictive models. These new sets of topic scores could be generated via the appropriate mathematical transformations that address extreme scores and skew of the existing topic scores computed from the functions representing the performance metrics. These new set of topic scores should be more informative and have higher correlations with respect to the high-cost performance metrics and this should in turn improve the predictions of the high-cost performance metrics. It is this suggestion that we employed in the proposed approaches presented in this paper.

5.2 The discussion for the results obtained by proposed approaches and compared to baseline

This section discusses the results of the ranked correlations of the predictions of the precision and nDCG performance metrics for the proposed approaches. The discussion also includes a comparison of the predictions of proposed approaches with the baseline [4].

The presented results in section 4.2 have shown that the proposed approaches produced better results than the baseline. These improved results of the proposed approaches are not surprising and are attributed to some differences between the proposed and baseline approaches. To begin with, the proposed approaches employ the topic scores of the low-cost performance metrics as opposed to the system scores used by the baseline method. System scores are computed by averaging the topic scores of performance metrics and the averaging operation leads to some error in the computed values. In addition, since the high-cost performance metrics are predicted, naturally, there is also some error during the process of prediction. Hence, when the proposed approaches are compared to the baseline, more error is incurred for the latter. Also, the methods used to select predictors of the low-cost performance metrics are worth some detailed discussion. The proposed approaches each employ their methods of selecting predictors and it is clear that the proposed predictor selection methods are better. Since the focus of this study is the predictions of the ranked correlations of performance metrics, the predictor selection methods for the proposed approaches also employ ranked correlations in the selection process of predictors of the low-cost performance metrics. Therefore, this leads to the selection of better predictors of the low-cost performance metrics. In contrast, the baseline method only employs the power set coupled with the root mean square to select the better predictors of the performance metrics.

Lastly, the proposed approaches use two sets of topic scores of the low-cost performance metrics. The first set comprises topic scores computed using equation definitions of the performance metrics. The second set constitutes topic scores generated by applying the cube root, yeo-Johnson, exponential and logarithmic functions to the first set of topic scores. Generating topic scores in this way addresses the extreme values and skew in distributions of topic scores. Also, the generated set of topic scores of low-cost performance metrics may correlate better with the target high-cost performance metrics and may be more informative of the target high-cost performance metrics.

5.3 Threat to validity

Validity concerns the scientific and conceptual soundness of a research study and its purpose is to produce valid conclusions [42]. In this study, the internal, external and construct validity threats have been addressed. Internal validity refers to the ability of the research design to rule out alternative explanations of the obtained results for the experiments. To avoid threats to internal validity, our experiments follow well-explained procedures. External validity refers to the generalizability of the results. To ensure the generalizability results, both experiments used several test sets with performance metrics computed using different test collections. In addition, the experiment on the predictability of performance metrics employed several machine learning models. Construct validity refers to the congruency between the results of the study and the theoretical underpinning's. For this study, the term high-

cost performance metric was clearly defined at the inception of the study and to ensure sufficiency of data, several test collections have been used to compute the high and low-cost performance metrics.

6.0 CONCLUSION

In this study, the predictability of performance metrics computed using several test collections has been investigated and two approaches that predict the high-cost precision and nDCG performance metrics have also been proposed. This study has shown that the similarity is low between low-cost performance metrics computed using relevance judgments and runs of different test collections. Also, this study has shown that the proposed approaches performed better than the baseline on most of the cut-off depths of documents of the low-cost performance metrics. Despite these achievements, there is still room for future work. To start with, a question to pose for future investigations is: to what extent are the individual low-cost performance metrics affected by the low similarity of performance metrics computed using different test collections? The answer to this question will help identify the best corrective strategy to employ to address the data set shift for each performance metric for better predictive results. In this study, an attempt has been made to ensure the generalizability of results by using several test collections produced over several years. However, only TREC test collections have been used in the experiments. Hence, in future, similar studies could be replicated using other test collections from initiatives such as the cross-language evaluation forum (CLEF) and the national institute of informatics test collection for information resources (NCTIR). In as far as the predictive approaches are concerned, there is a need to explore approaches that incorporate more complex models such as deep learning models.

7.0 ACKNOWLEDGMENT

This research work has been supported by Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/1/2020/ICT06/UM/02/1)

REFERENCES

[1] S. I. Moghadasi, S. D. Ravana, and S. N. Raman, “Low-cost evaluation techniques for information retrieval systems: A review,” *Journal of. Informetrics*, vol. 7, no. 2, pp. 301–312, 2013.

[2] K. Roitero, A. Brunello, G. Serra, and S. Mizzaro, “Effectiveness evaluation without human relevance judgments: A systematic analysis of existing methods and of their combinations,” *Information. Processing and Management*, vol. 57, no. 2, 2020, doi: 10.1016/j.ipm.2019.102149.

[3] M. Makary, M. Oakes, R. Mitkov, and F. Yammout, “Using supervised machine learning to automatically build relevance judgments for a test collection,” in *Proceedings – International Workshop on Database and Expert Systems Applications, DEXA*, 2017, vol. 2017-Augus, pp. 108–112, doi: 10.1109/DEXA.2017.38.

[4] S. Gupta, M. Kutlu, V. Khetan, and M. Lease, “Correlation, Prediction and Ranking of Evaluation Metrics in Information Retrieval,” in *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*, 2019, pp. 636–651.

[5] S. Muwanei, S. D. Ravana, W. L. Hoo, and D. Kunda, “The Prediction of the High-Cost Non-Cumulative Discounted Gain and Precision Performance Metrics in Information Retrieval Evaluation,” in *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2021, pp. 25–30.

[6] T. Ishioka, “Evaluation of criteria for information retrieval,” in *Proceedings - IEEE/WIC International Conference on Web Intelligence, WI 2003*, 2003, pp. 425–431, doi: 10.1109/WI.2003.1241232.

[7] J. A. Aslam, E. Yilmaz, and V. Pavlu, “A geometric interpretation of r-precision and its correlation with average precision,” in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’05*, 2005, p. 573, doi: 10.1145/1076034.1076134.

[8] T. Sakai, “On the Properties of Evaluation Metrics for Finding One Highly Relevant Document,” *IPSJ Digital. Courier*, vol. 3, pp. 643–660, 2007, doi: 10.2197/ipsjdc.3.643.

[9] T. Sakai, "On the reliability of information retrieval metrics based on graded relevance," *Information. Processing and. Management*, vol. 43, no. 2, pp. 531–548, 2007, doi: 10.1016/j.ipm.2006.07.020.

[10] J. A. Thom and F. Scholer, "A comparison of evaluation measures given how users perform on search tasks," *ADCS 2007 –in Proceedings of the Twelfth Australasian. Doctoral. Computing. Symposium.*, pp. 100–103, 2007.

[11] L. Egghe, "The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations," *Information Processing. And Management.*, vol. 44, no. 2, pp. 856–876, Mar. 2008, doi: 10.1016/j.ipm.2007.03.014.

[12] A. Baccini, S. Déjean, L. Lafage, and J. Mothe, "How many performance measures to evaluate information retrieval systems?," *Knowledge and Information. Systems.*, vol. 30, no. 3, pp. 693–713, Mar. 2012, doi: 10.1007/s10115-011-0391-7.

[13] T. Jones, F. Scholer, P. Thomas, and M. Sanderson, "Features of disagreement between retrieval effectiveness measures," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 847–850, doi: 10.1145/2766462.2767824.

[14] C. Hauff, D. Hiemstra, L. Azzopardi, and F. De Jong, "A case for automatic system evaluation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5993 LNCS, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. van Rijsbergen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 153–165.

[15] M. Hosseini, I. J. Cox, N. Milic-Frayling, V. Vinay, and T. Sweeting, "Selecting a subset of queries for acquisition of further relevance judgements," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6931 LNCS, pp. 113–124, doi: 10.1007/978-3-642-23318-0_12.

[16] B. Carterette, J. Allan, and R. Sitaraman, "Minimal test collections for retrieval evaluation," in *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, vol. 2006, pp. 268–275, doi: 10.1145/1148170.1148219.

[17] M. Hosseini, I. J. Cox, N. Milić-Frayling, M. Shokouhi, and E. Yilmaz, "An uncertainty-aware query selection model for evaluation of IR systems," in *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012, pp. 901–910, doi: 10.1145/2348283.2348403.

[18] J. A. Aslam and R. Savell, "On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments," in *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 2003, no. SPEC. ISS., pp. 361–362, doi: 10.1145/860500.860501.

[19] M. Kutlu, T. Elsayed, and M. Lease, "Intelligent topic selection for low-cost information retrieval evaluation: A New perspective on deep vs. shallow judging," *Information Processing and. Management.*, vol. 54, no. 1, pp. 37–59, 2018, doi: 10.1016/j.ipm.2017.09.002.

[20] P. Rajagopal and S. D. Ravana, "Effort-based Information Retrieval Evaluation with Varied Evaluation Depth and Topic Sizes," in *ACM International Conference Proceeding Series*, 2019, pp. 143–147, doi: 10.1145/3361785.3361794.

[21] K. Roitero, M. Soprano, and S. Mizzaro, "Effectiveness evaluation with a subset of topics: A practical approach," in *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 2018, pp. 1145–1148, doi: 10.1145/3209978.3210108.

[22] M. M. Rahman, M. Kutlu, T. Elsayed, and M. Lease, "Efficient Test Collection Construction via Active Learning," in *ICTIR 2020 - Proceedings of the 2020 ACM SIGIR International Conference on Theory of Information Retrieval*, 2020, pp. 177–184, doi: 10.1145/3409256.3409837.

[23] M. M. Rahman, M. Kutlu, and M. Lease, "Constructing test collections using multi-armed bandits and active learning," in *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2019, pp. 3158–3164, doi: 10.1145/3308558.3313675.

[24] D. E. Losada, J. Parapar, and A. Barreiro, "When to stop making relevance judgments? A study of stopping methods for building information retrieval test collections," *Journal of the Association of Information. Science and. Technology*, vol. 70, no. 1, pp. 49–60, 2019, doi: 10.1002/asi.24077.

[25] M. Makary, M. Oakes, and F. Yamout, "Towards automatic generation of relevance judgments for a test collection," in *2016 11th International Conference on Digital Information Management, ICDIM 2016*, 2016, pp. 121–126, doi: 10.1109/ICDIM.2016.7829763.

[26] David E Losada, Javier Parapar, and Alvaro Barreiro, "Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems," *Information. Processing. And Management.*, vol. 53, no. 5, pp. 1005–1025, 2017.

[27] Prabha Rajagopal, Sri Devi Ravana, Yun Sing Koh, and Vimala Balakrishnan, "Evaluating the effectiveness of information retrieval systems using effort-based relevance judgment," *Aslib Journal of. Information. Management*, 2019.

[28] S.D. Ravana, P. Rajagopal, and V. Balakrishnan, "Ranking retrieval systems using pseudo relevance judgments," *Aslib Journal of. Information. Management*, vol. 67, no. 6, pp. 700–714, 2015, doi: 10.1108/AJIM-03-2015-0046.

[29] C. Buckley, "Overview of the TREC-9 web track," in *Text REtrieval Conference*, 2001, pp. 81–85.

[30] D. Hawking and N. Craswell, "Overview of the TREC-2001 web track," *Nist Spec. Publ. Sp*, no. 250, pp. 61–67, 2002.

[31] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. A. Clarke, and E. M. Voorhees, "Overview of the TREC 2013 web track," 2013.

[32] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees, "TREC 2014 Web Track Overview," 2015.

[33] J. Allan, "Overview of the trec 2004 robust retrieval track," in *Proceedings of TREC*, 2004, vol. 13.

[34] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transaction. Information Systems.*, vol. 20, no. 4, pp. 422–446, 2002, doi: 10.1145/582415.582418.

[35] Alistair Moffat and Justin Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Transaction .Information. Systems*, vol. 27, no. 1, pp. 1–27, 2008.

[36] C. Zhai and S. Massung, *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan and Claypool, 2016.

[37] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 621–630.

[38] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, 2004, p. 25, doi: 10.1145/1008992.1009000.

[39] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proceedings of the 15th ACM international conference on Information and knowledge management*, 2006, pp. 102–111.

[40] J. Quinonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer, *Dataset shift in machine learning*. Mit Press, 2009.

[41] S. Galli, *Python feature engineering cookbook : over 70 recipes for creating, engineering, and transforming features to build machine learning models*. Packt Publishing, 2020.

[42] G. R. Marczyk, D. DeMatteo, and D. Festinger, *Essentials of research design and methodology*. John Wiley and Sons, Inc, 2021.