Reusability Assessment of Test Collections with Multi-levels of Judgments

Maryam Khodabakhsh
Department of Computer
Ferdowsi University of Mashhad
Mashhad, Iran
Maryam.khodabakhsh@stu-mail.um.ac.ir

Saeed Araban
Department of Computer
Ferdowsi University of Mashhad
Mashhad, Iran
Saeed.araban@um.ac.ir

***Abstract*—Constructing good test collection is an expensive and time-consuming process. Traditionally, test collections contain binary judgments. In recent years, however, there has been increasingly interest in test collections with Multi-levels judgments and of certain qualities. Such collections are even more expensive to construct. Therefore, ability to reuse test collections can not only save construction costs, but also boosts our confidence in their quality. This paper proposes a method for assessing reusability of a test collection with multi-level judgments. The proposed method can help IR researchers to determine whether an existing test collection with a set of multi-level judgments is suitable for evaluating a new IR system or not. Results of our experiments (on MAHAK test collection) suggest that this method can help assessing reusability of a test collection.**

Keywords—test collection, evaluation, confidence interval, reusability, multi-level judgments, information retrieval (IR) system

I. INTRODUCTION

An IR system deals with incomplete and underspecified information in the form of the queries issued by users. In respond to a user's query, an IR system returns a ranked list of documents that hopefully have some degree of relevance to the query with the most relevant document at the top of the list [1].

In recent years, IR research has been focused on measuring effectiveness of IR systems using *test collections* (justification for importance of the test collection selection). Therefore it is important to select a suitable test collection for evaluating an IR system.

Three main components of a test collection are [1]:

- A collection of documents; each document has a unique identifier (Docid);
- A set of queries; each query has a query id (Qid);
- A set of relevance judgments (often referred to as Qrels — query relevance set) that is a set of triples. Each triple composed of Qid, Docid and relevance level of a document to a query.

Test collections traditionally contain binary judgments of the relevance of documents to queries. However, in recent years, there has been interest in generalizing judgments to non-binary, that includes Multi-levels (aka graded scales) judgments, aspect relevance, and preferences judgments [2]. In test collections with multi-levels of judgments, relevance level of a document to a query is usually represented using numerical values. Relevance level is determined base on how much a document is relevant to a query according to its contents. For example, TREC 2005 used three levels of relevance: irrelevant, relevant and highly relevant, which are represented as 0, 1, and 2 respectively [3].

The development and evaluation of modern IR systems should be based on their ability to retrieve highly relevant documents. This is desirable from the users' perspective, because in such environments, the users tend to look at first few documents of a list that is better to be highly relevant [4].

When developing a large test collection, having a searchable collection (e.g. a music collection) and a set of queries are not very difficult. However, having enough judgments in the test collection, so that they give high confidence to the search results is difficult [5].

On one hand, judging the relevance level of documents to queries require a great deal of human efforts, which makes Qrels construction a very time consuming and expensive process. On the other hand, Qrels plays an important role in evaluation of test collections. A test collection without enough judgments would not be a good reference for assessing IR systems. Therefore, there must be a tradeoff between two of these issues. In other words, test collection developers need to make sure that their collection has enough judgments and viable to make. Furthermore, making test collections more reusable for different IR systems can amortize the high costs of developing those collections.

Whilst for a small test collection, Qrels may contain all qid/docid/judgment triples; however, it is not feasible (or may not even possible) to judge the relevance of all documents to all queries in a large test collection. Therefore, one of the most common methods is to judge each query with respect to a subset of documents. Such subsets must contain all or most documents which are relevant to a query. One method for constructing this subset is the *pooling method*. The pooling method provides a way to focus judging effort on those documents least likely to be irrelevant [6]. In the first step of this method, several IR systems are used for loading queries and retrieving and ranking documents, relevant to those queries. Then a document pool is formed by using the top *k* documents submitted by each of the participated IR systems (See Fig. 1). The assessor judges the relevance of documents to queries in the document pool, which become Qrels [7].

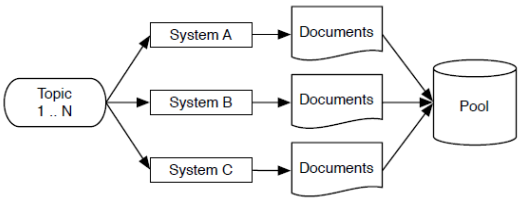

Figure 1. Pooling of documents [7].

In test collections of realistic size, it is unlikely that pooling method finds all the highly relevant or relevant documents in the collection. When new IR systems are subsequently evaluated using the same test collection, we face with documents that not exist in Qrels. In such condition, there are two options. One option is to collect judgments for the documents retrieved that were not previously judged. This can be costly and time consuming, especially when new IR systems must be tested over a large test collection. The other option is to only use existing judgments and effectively ignore newly retrieved documents that have not been previously judged which this approach may lead to a highly inaccurate measure of the system's true performance [6].

By considering that pooling method is used to construct Qrels and this method makes lose some judgments; the aim of this paper is to present a method to assess the reusability of Qrels of a test collection. In other words, the presented method helps us to assure sufficient judgments in Qrels. In this method, we use confidence intervals which are calculated for a retrieve measure. Confidence intervals represent reusability of test collection.

The main contributions of this paper are;

- Using the reusability concept is useful for test collection developers because they can improve Qrels in the test collection if it is necessary. One of the main contribution is viewed in this paper is to develop confidence interval estimate method which is used to assess the reusability of test collection with multi-levels judgments.
- The number of relevance levels is different from a test collection to another one. This variation in the number of levels must be considered in confidence interval estimate method. For this purpose, NDCG is used because it is independent of the number of relevance levels.
- Because of using pooling method, in the list of documents retrieved by the IR system, there are documents which do not exist in Qrels. Multinomial logistic regression model is used to predict the relevance level of these documents to queries.

The remainder of the paper is organized as fallows. Section 2 describes previous works related to reusability, evaluation measures and the methods of the prediction of unjudged documents base on relevance. The method of assessment of reusability is introduced completely in section 3. Finally this method is executed using MAHAK test collection and the results of the experiments are present in section 4. Section 5 concludes the paper.

II. RELATE WORK

The Text REtrieval Conference (TREC) has addressed the need for a web test collection. The documents and queries in the collections built in the track were taken from the web and the relevance judgments were produced using pooling as in other TREC collections. Unlike other TREC collections, the most recent web track used a three point relevance scale: irrelevant, relevant and highly relevant [8].

There are many measures to evaluate the performance of IR systems, for example precision, recall, average precision and etc. These measures are applied when the judgment about the relevance of each document to query is binary. However there is only one commonly used measure for graded relevance, namely the Discounted Cumulative Gain (DCG) [9]. For definition DCG, it is necessary to introduce CG measure.

Cumulative Gain (CG) is the sum of relevance level values (rel) measured in the top n retrieved documents.

$$CG(n) = \sum_{i=1}^{n} rel\,(i) \tag{1}$$

rel(i) demonstrates the relevance level of document *i* to query.

CG ignores the rank of documents. The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the relevance level value is reduced logarithmically proportional to the position of the result. The DCG accumulated at a particular rank position *p* is defined as:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel(i)}{\log_2(i)} \tag{2}$$

DCG is normalized against an ideal ordering of the relevant documents, IDCG:

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{3}$$

Järvelin and Kekäläinen generalized recall and precision which used to evaluate judgments with relevance level [10].

$$gP = \sum_{d \in R} r(d)/n \tag{4}$$

$$gR = \sum_{d \in R} r(d) \Big/ \sum_{d \in D} r(d) \tag{5}$$

R is the set of n documents retrieved by IR system and *r(d)* is relevance scores which are real numbers ranging [0,1].

A new measure is introduced which named Graded Average Precision (GAP). It inherences average precision

feature and it is applied to evaluate judgments with relevance level [11].

These measures are applied when Qrels is complete. In other words, there is a judgment for each document which is retrieved to each query. Pooling could miss up to 50% of the relevant documents in the collection [6]. In this case, the relevance level of the unjudged documents to the queries can be predicted. BÜttcher et al. used an SVM to predict the relevance of unjudged documents to find likely new relevant documents [12]. They trained support vector machines (SVM) text classifier using existing Qrels. Then this classifier is used to predict for any unjudged document whether the document is relevant for the given query or not [12].

Carterette's definition of the reusability is as follows: whether two new systems could be reliably ranked relative to each other using relevance predictions based on very small sets of judgments. The predictions are used to calculate a probability that two systems are likely to swap after additional judgments. He believed that reusability must evaluate as IR system's ability to produce results with high reliably [13].

Can an IR system which did not participate in the Pool, be evaluated carefully using the pool? Answer to this question is our aim in this paper. For this purpose, the relevance level of unjudged documents must first be estimated, and then expectation and confidence intervals are calculated for NDCG using these estimates. If the confidence intervals are wide then it is concluded that the pool required more judgments.

III. Reusability Assessment

Reusability measure is important of theoretical and practical. The theory behind this new measure can be used to develop new measures and metrics. From a practical side, it can be used by test collection developers to determine whether existing Qrels in test collection is sufficient to evaluate a new IR system, or new judgments are needed [6]. Qrels is used to evaluate the performance of a new IR system.

NDCG is one of the classical information retrieval measures, not only in binary mode; but also it can be used in non binary. For using NDCG, it is assumed that the test collection is complete. In other words, there is a judgment for each document retrieved to each query, and otherwise using NDCG is not justified.

Using Qrels for evaluating a new IR system by NDCG, the problem is accursed when the unjudged documents are retrieved. So dealing with these documents is important. It can be assumed that these documents are irrelevant. In this case, our evaluation of system's performance will not be accurate. Several measures have been introduced to solve the problem of unjudged documents, but this method does not work well in determining of accuracy of assessment. But given the reusability measure, it is possible to determine confidence in the evaluation of new IR system by test collection.

For reusability assessment, confidence intervals are estimated for the metric of interest (it is NDCG here), before it is necessary to acquire the relevance level of unjudged documents by multinomial logistic regression.

A. Interval Estimation of Reusability

Reusability measure is expressed in the form of confidence intervals. The confidence intervals width is representative of reusability. If a set of judgment J exists on a set of query Q, and we want to evaluate a new IR system on Q by metric m. having J and a list of documents retrieved by system, confidence interval must be estimated for metric m. if the confidence interval is wide, it can be concluded that J is not appropriate and more judgments are required.

Confidence intervals allow users to determine the uncertainty in estimating the performance of the new IR system. Uncertainty is the consequence of unjudged documents retrieved by the system. The more uncertainty, the less reusability of test collection is [6].

B. Confidence Intervals

To compute confidence intervals for metric m, variance and expectation must be calculated for m [6]:

$$E[\bar{m}] = \frac{1}{n}\sum_{i=1}^{n} E[m(Q_i)] \tag{6}$$

$$Var[\bar{m}] = \frac{1}{n^2}\sum_{i=1}^{n} Var[m(Q_i)] \tag{7}$$

The expectation (mean) of metric m is demonstrated by $\bar{m}$. $m(Q_i)$ is metric m that evaluated on Q_i. n is the number of queries in Q. So $100(1-\alpha)$ % confidence interval for m is as follows [6]:

$$\left[E[\bar{m}] - z_{\frac{\alpha}{2}}\sqrt{\frac{Var[\bar{m}]}{n}}, E[\bar{m}] + z_{\frac{\alpha}{2}}\sqrt{\frac{Var[\bar{m}]}{n}}\right] \tag{8}$$

The confidence interval is valid if $\bar{m}$ is distributed normally. On the other hand, the distribution of each metric tends to be normal with increasing the number of queries. It is an application of Central Limit Theorem.

If the reusability is shown in the form of confidence interval calculated for metric m, then it is necessary that this metric supports the relevance levels. NDCG is independence of relevance levels [9] and used for calculating the performance of IR system.

According to (8), there are 2 ways to reduce confidence interval:

1. The value of variance of m becomes decreased. In other words, to judge more unjudged documents.
2. The number of queries is increased.

C. Example

In this session, an example is presented to understand better. Suppose an IR system retrieves four documents which the

first and third documents have the relevance level of 2 and 1 respectively. Second and fourth documents are unjudged (See Fig. 2).

Figure 2. The situation of documents in example

Depending on how unjudged documents resolve; there are different cases as shown in Table 1. The value of variance is:

$$0.1504 / 9 = 0.0167$$

Now consider that the forth document is judged and has the relevance level of 1. The number of cases decreases from 9 to 3 and the value of variance is: 0.0133.

So, the variance is decreased by clarify the situation of an unjudged document. Thereby, uncertainly which is the result of retrieval of unjudged documents is reduced.

D. Estimate the Relevance Levels of Unjudged Documents

Using pooling method in test collection construction causes to loss several documents which is relevant to query. IR system may retrieve documents that do not exist in Qrels. If they are assumed irrelevant then our estimate of expectation and variance is inaccurate. So the relevance level of them must be predicted. The multinomial logistic regression model is a proper approach to predict the relevance level of unjudged documents. This regression is used when the dependent variable includes more than two categories and is nominal, in other words, a set of categories which cannot be ordered in any meaningful way [14].

In this regression, one category of the dependent variable is chosen as the reference category. If there are dependent variable categories 1, 2, …, J (0 is the reference category), the below relation is used to determine that y_i being in which categories:

$$\Pr(y_i = k) = \frac{\exp(X_i.\beta_k)}{1 + \sum_{j=1}^{J} exp(X_i.\beta_j)} \quad k = 1, 2, \dots, J \qquad (9)$$

To calculate, that the probability of y_i being in category 0 is given by the adding-up constraint that the sum of the probabilities of y_i being in the various categories equals one.

$$\Pr(y_i = 0) = \frac{1}{1 + \sum_{j=1}^{J} exp(X_i.\beta_j)} \qquad (10)$$

- β is model parameter vector which is estimated by maximum likelihood.
- X_i is independent variable vector (feature vector). In this paper, the document similarity character is used that means for each document i, the cosines of similarity of document i to other documents is calculated. The reason of using this feature is that if an IR system retrieves a document that does not exist in Qrels and is similar to one or more relevant documents, it is concluded that the document itself is relevant. In fact, the assumption is that documents which are remarkably similar tend to be relevant to the same requests [15].

IV. RESULTS AND DISCUSSIONS

A. Data

To simulate experiments, we have used MAHAK [16] test collection and several IR systems. The existing Qrels in MAHAK contains documents which their relevancy to queries is demonstrated by relevance levels. In MAHAK test collection, the numbers of 2, 1 and 0 are used for highly relevant documents, relevant and irrelevant respectively. Of course it is different the number of relevance levels and their

TABLE I. EXPECTATION FOR FIRST CASE IN EXAMPLE

The number of cases	The relevance level of document				DCG	IDCG	NDCG	$NDCG_i * p(NDCG_i)$	$(NDCG - E[NDCG])^2$
	1th	2th	3th	4th					
1	2	2	1	2	5.63	5.75	0.98	$0.98 * 0.1 = 0.098$	0.02
2	2	2	1	1	5.13	5.13	1	$1 * 0.1 = 0.1$	0.03
3	2	2	1	0	4.63	4.63	1	$1 * 0.1 = 0.1$	0.03
4	2	1	1	2	4.63	5.13	0.90	$0.9 * 0.1 = 0.09$	0.004
5	2	1	1	1	4.13	4.13	1	$1 * 0.1 = 0.1$	0.03
6	2	1	1	0	3.63	3.63	1	$1 * 0.1 = 0.1$	0.03
7	2	0	1	2	3.63	4.63	0.78	$0.78 * 0.1 = 0.078$	0.004
8	2	0	1	1	3.13	3.63	0.86	$0.86 * 0.1 = 0.086$	0.0004
9	2	0	1	0	2.63	3	0.88	$0.88 * 0.1 = 0.088$	0.002
								E[NDCG] = 0.84	0.1504

representation from a test collection to another one.

Ten open source search engines (Xapian, Sphinx, Indri, etc) are used to produce IR systems. The results of execution of IR systems on MAHAK test collection produce needed runs. The runs are ranked base on NDCG (true NDCG). These runs are used as data.

B. *Methodology*

The methodology is as follows [6]:

1. Select m runs randomly. These runs are named initial runs (remaining runs are named testing runs).
2. Construct pool from the top k documents retrieved for each query by initial runs.
3. Predict the relevance levels of unjudged documents by multinomial logistic regression in both initial and testing runs.
4. Calculate the expectation of NDCG for testing runs (expected NDCG).
5. Calculate the variance of NDCG for testing runs.
6. Calculate confidence intervals.

Confidence in the evaluation is the width of the 95% confidence interval. If the confidence intervals are wide, there is uncertainly in evaluation and more judgments are required to determine the performance of IR system. But if the confidence intervals are not wide then it is unlikely to need more judgments.

C. *Evaluation Method*

1. Testing runs must be ranked base on NDCG (expected NDCG).
2. The quality of rankings of testing runs (ranking base on true NDCG and expected NDCG) are evaluated by Kendall's τ rank correlation. τ is proportional to the number of pairs that have swapped between two rankings. If τ = 1, no pairs have swapped; and if τ = 0, half the pairs are swapped.

$$\tau = 1 - \frac{2 \times |d_{\Delta}(\rho_1, \rho_2)|}{N(N-1)} \qquad (11)$$

   - N is the number of objects which exist in list.
   - $d_{\Delta}(\rho_1, \rho_2)$: difference distance between two rankings.
3. For IR, $\tau \geq 0.9$ is proper [6].

D. *Examples*

In this session, some examples are presented in detail. We first select 1 run randomly (m=1) and construct pool from the top 5 documents retrieved for each queries, then judge them by Qrels in MAHAK. 118 of these documents have the levels of high relevant and relevant.

This pool is used to evaluate remaining runs and the relevance level of unjudged documents of these runs must be predicted by multinomial logistic regression and using document similarity. Then expected NDCG, variance and confidence interval is calculated. Fig. 3 shows expected

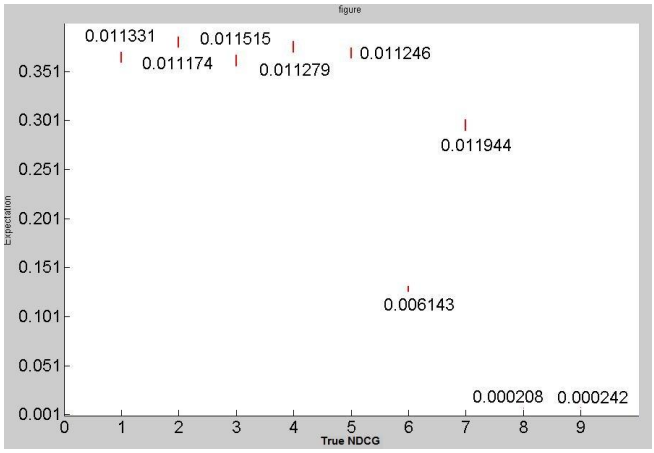

Figure 3. Example 1: Expected NDCG and confidence intervals for 9 remaining runs that ranked base on true NDCG.

NDCG and 95% confidence intervals for remaining runs which ranked base on true NDCG.

Ranking of remaining runs base on expected NDCG has a Kendall's τ rank correlation of 0.6111 with ranking base on true NDCG. Initial run is ranked 6th base on true NDCG.

Another example shows the effect of more judgments which are obtained from another runs. The initial run which is selected in this example is ranked 1th base on true NDCG. After selecting the top 5 documents retrieved for each queries to construct pool and jugging them, 375 (instead of 181) number of these documents has levels of high relevant and relevant. The value of Kendall's τ rank correlation between true NDCG and expected NDCG is 0.7888 and the width of confidence interval for another runs are narrower as shown in Fig.4.

E. *Experimental Results*

We have executed all steps which are described in session Methodology many times. In each execution, initial

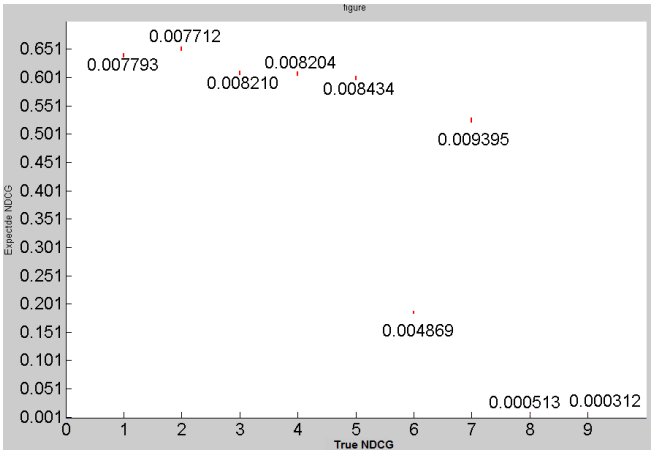

Figure 4. Example 2: Expected NDCG and confidence intervals for 9 remaining runs that ranked base on true NDCG.

runs are selected randomly. Table 2 represents average values which are obtained for m and k in experiments.

m and k are experimental parameters. The number of documents which is retrieved for 216 (the number of queries in MAHAK) queries exports in the next column of the table. The results of experiment are confidence intervals and Kendall's τ rank correlation between true NDCG and expected NDCG which is calculated for testing runs. The effect of increasing pool depth and the number of runs contributed in constructing pool can be seen in the table.

The effect of increasing the number of runs contributed in constructing pool can be seen in rows of the table which nearly 637 relevance documents are retrieved (4, 7, and 11). Increasing the number of runs which contributes in constructing pool while keeping the number of relevance documents constant causes both decreasing the width of confidence intervals (increasing the reusability) and increasing the quality of ranking of runs.

V. CONCLUSIONS

In the past, the approach of estimation of performance used for assessing reusability is not proper because the confidence is not determined. The confidence interval method is used to assist the reusability of binary test collection. In this paper, we have extended this method to assist the reusability of test collection with multi-levels judgments. If confidence intervals are wide, more judgments must need to evaluate new IR system accuracy.

Expectation and variance must be calculated for NDCG to estimate confidence intervals and multinomial logistic regression model is used to predict the relevance levels of unjudged documents.

The results of experiments show that confidence intervals which estimated are accurate because these intervals contain true NDCG value. Also Kendall's τ rank correlation between true NDCG (with prefect set of Qrels) and expected NDCG (with small set of Qrels) is 0.9.

TABLE II. THE RESULTS OF EXPERIMENTS FOR MAHAK TEST COLLECTION

m	k	The number of relevance documents	Confidence interval	τ
1	1	121	0.128020155	0.29365
	5	375	0.0069893855	0.74605
	10	534	0.0065785105	0.78575
	20	663	0.0059941193	0.8889
2	1	170	0.007941276	0.7858
	2	424	0.00572081	0.85716
	10	616	0.005418785	0.9286
	20	731	0.005170732	0.9286
3	1	187	0.006889205	0.7143
	5	470	0.005039548	0.85715
	10	633	0.004924617	0.9048
	20	766	0.004829821	0.9048

REFERENCES

[1] M. Sanderson, "Test Collection Based Evaluation of Information Retrieval Systems," in Foundations and Trends in Information Retrieval. vol. 4, 2010, pp. 247–375.

[2] B. Carterette, P. N. Bennett, and O. Chapelle, "A Test Collection of Preference Judgments," in in SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval New York, NY, USA: ACM, 2008.

[3] B. Carterette and P. N. Bennett, "Evaluation measures for preference judgments," in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval Singapore, Singapore: ACM, 2008.

[4] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval Athens, Greece: ACM, 2000.

[5] B. A. Carterette, "Low-Cost and Robust Evaluation of Information Retrieval," University of Massachusetts Amherst, 2008, p. 255.

[6] B. Carterette, E. Gabrilovich, V. Josifovski, and D. Metzler, "Measuring the reusability of test collections," in Proceedings of the third ACM international conference on Web search and data mining New York, New York, USA: ACM, 2010, pp. 231-240.

[7] H. Joho, R. Villa, and J. M. Jose, " Interaction Pool: Towards a User-centered Test Collection," in In SIGIR (ACM) Amsterdam, The Netherlands, 2007, pp. 17-20.

[8] E. M. Voorhees, "Evaluation by highly relevant documents," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval New Orleans, Louisiana, United States: ACM, 2001.

[9] 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 Hong Kong, China: ACM, 2009.

[10] J. Kekäläinen and K. Järvelin, "Using graded relevance assessments in IR evaluation," J. Am. Soc. Inf. Sci. Technol., vol. 53, pp. 1120-1129, 2002.

[11] S. E. Robertson, E. Kanoulas, and E. Yilmaz, "Extending average precision to graded relevance judgments," in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval Geneva, Switzerland: ACM, 2010.

[12] . B¨uttcher, C. L. A. Clarke, and P. C. K. Yeung, "Reliable information retrieval evaluation with incomplete and biased judgements," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval Amsterdam, The Netherlands: ACM, 2007.

[13] B. Carterette, "Robust test collections for retrieval evaluation," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval Amsterdam, The Netherlands: ACM, 2007.

[14] in http://sites.stat.psu.edu/~jls/stat544/lectures/lec19.pdf.

[15] B. Carterette and J. Allan, "Semiautomatic evaluation of retrieval systems using document similarities," in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management Lisbon, Portugal: ACM, 2007.

[16] K. S. Esmaili, H. Abolhassani, M. Neshati, E. Behrangi, A. Rostami, and M. M. Nasiri, "Mahak: A Test Collection for Evaluation of Farsi Information Retrieval Systems," in Amman, Jordan: ACS/IEEE International Conference on Computer Systems and Applications, 2007, pp. 639 - 644.