# Effort-based Information Retrieval Evaluation with Varied Evaluation Depth and Topic Sizes

Prabha Rajagopal
Department of Information Systems
Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
prabz13@yahoo.com

Sri Devi Ravana
Department of Information Systems
Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
sdevi@um.edu.my

## ABSTRACT

The information retrieval accessed globally is a vital productivity boost for most organization. However, the outcome of information retrieval system evaluation does not agree with the real user's satisfaction. Information retrieval systems retrieving low effort documents are preferred by users as effort factor influences user satisfaction. However, there lacks evaluation on deeper depth and reduced topic sizes with the incorporation of effort in evaluating the information retrieval systems. Therefore, this study aims to measure the effectiveness of evaluating information retrieval systems using effort-based relevance judgment with varied evaluation depth, and reduced topic sizes using effort-based relevance judgment. The evaluation depth is varied from 10, 100 and 1000, and the correlation coefficient between the baseline and those generated using effort-based relevance judgment is evaluated. Topic sizes smaller than 50 were also considered and measured to observe changes in correlation coefficient due to effort. As a result, standard evaluation depth, and reduced topic sizes are sufficient for evaluation using effort-based relevance judgment.

## CCS Concepts

**• Information system→Evaluation of retrieval results→Relevance Assessment**

## Keywords

Effort; information retrieval evaluation; relevance judgment; topic size

## 1. INTRODUCTION

An Information Retrieval (IR) system is designed to capture, process, store and retrieve information is necessary to hold a business together [1]. Online document storage enables retrieval globally and throughout the days. Global access to document retrieval is a major productivity boost for most organizations [2]. One of the main approaches in Information Retrieval (IR) evaluation is the system-oriented evaluation focusing on measuring the retrieval systems using a test collection consisting of a document corpus, topics or queries, and relevance judgment. The test collection makes it possible for scoring the IR systems' performance using effectiveness metrics. With the scores, the IR systems can be ranked and compared among each other to determine good or bad retrieval systems.

However, the evaluation of the IR systems is a challenge with increased information in the Web [3]. The difficulties in measuring the IR systems depend largely on the two components of the test collection, which are number of topics and relevance judgments. It has been much interest to many involved in the IR field to achieve good evaluation outcomes with reduced topics or better topics, and relevance judgments [5]–[6].

As for the relevance judgment, the relevance of a document had always been a priority in measuring the performance of the systems using effectiveness metrics. Yet, the outcome of IR evaluation does not agree with real user satisfaction [7], [8], and there exist noncorrelation between the effectiveness metrics and the real user experience [9]. It is known that factors such as effort, system and user effectiveness, and user characteristics [10] influence user satisfaction.

Recent studies indicated real users of the Web prefer low effort documents [11], [12]. Real users give up easily and do not put in as much effort as expert judges while identifying relevancy in a document. Therefore, an IR system incorporating retrieval of low effort documents is preferable by the user due to lesser work in identifying the relevancy of the documents compared to an IR system retrieving high effort documents [11], [12].

Previously, the importance of effort was measured in various ways [11], [13] but limited depth of evaluation and retrieval systems (top 10 only) were used to show the differences in system rankings due to low effort relevance judgments. The differences in system rankings using original and low effort relevance judgments beyond evaluation depth 10, and all retrieval systems within a test collection are unknown.

Hence, this study aims to measure the effectiveness of evaluating IR systems using low effort relevance judgment with variation in the depth of evaluation for a specific common evaluation metric. Additionally, this study also incorporates the evaluation of reduced topic sizes with the usage of effort-based relevance judgment.

The next sections touch on literature about the importance of effort and the effects of topic sizes in IR evaluation. Next, the approach for generating and evaluating retrieval systems using low effort relevance judgments is presented. Then, the evaluation

*ICBIM 2019*, September 12–14, 2019, Paris, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7232-9/19/09…$15.00
DOI: https://doi.org/10.1145/3361785.3361794

of the retrieval systems is presented and the results are discussed. Finally, conclusions are drawn.

# 2. BACKGROUND

## 2.1. The Importance of Effort in Addition to Relevance

Disagreements exist between the outcome of IR evaluation and real user satisfaction [7], [8], and effort has been identified as an important aspect of user satisfaction in addition to relevance [12]. Additionally, increased document size and the degree of relevance of a document indicating 'relevant' requires more effort in judging [11]. Experimentation was conducted to measure the effort needed by users to identify and consume the information from a document with regards to time [12]. When compared using dwell time and click counts, a mismatch was found between relevance judgments from real users and expert judges. Therefore, the utility of a document to actual users is dependable in the effort needed to consume the relevant information [12].

Another study focused on measuring readability, findability and understandability effort in obtaining a relevance judgment and stated that effort should be a part of relevance judgment if user satisfaction is prioritized [13]. They even indicated user satisfaction is a function of relevance and effort. High-effort documents are harder to consume or benefit user, thus it is less likely for the users to read it [12].

The logistic regression models are used to predict the relevance assessment using initial user input on relevance and effort [12]–[14]. With the incorporation of effort into relevance judgments, Kendall's tau correlation coefficient shows the performance of the retrieval systems are different compared to when the effort is not used in the relevance judgments [13]. The result highlights the importance of effort together with relevance in satisfying user's information needs. There are factors such as findability, readability, and understandability that characterizes the effort. Based on these factors, some were highlighted as significantly affecting effort [13].

## 2.2. The Effect of Topic Size and Relevance to System Performance

Topics highly influence the comparison between retrieval systems. It is likely for a different set of topics of the same size to produce different results [15] and some topics or topic sets predict the actual effectiveness of systems better than others [3], [16]. An analysis of system ranking estimation proved that performance depends heavily on a topic subset [4]. The experiment was conducted on different estimation approaches such as data fusion, random sampling, document similarity, and document score. They also identified the random sampling approach [17] as the most stable and best-performing estimation method. With the use of eight test collections (TREC 6 – 10, and Terabyte track 4 – 6), experimental results suggest the right topic selection could improve the performance estimation by 32% on average.

Topics influence the comparison of system performance, but the document relevance also influences the evaluation of retrieval systems. An experiment was conducted to examine the effect of highly relevant documents to retrieval system evaluation [18]. The study shows the relative effectiveness of different Web track runs differ when evaluated using only highly relevant documents versus relevant documents. However, the effect of changes due to judgments in TREC test collections shows that the relative performance of the system runs was comparable [19]. Recently, a study [20] showed that the number of relevant documents is not an effect of evaluation inconsistencies. Previously, [18] mentioned the change in effectiveness could be due to small numbers of relevant documents causing the retrieval measures to be unstable.

# 3. GENERATING EFFORT-BASED RELEVANCE JUDGMENT (EQREL)

Based on the literature, some of the efforts were significantly impacting user satisfaction. Hence, in this study, the simple document features and readability features will be utilized. The simple document features contain features such as measuring the number of characters, number of words, and number of sentences in the document. The readability features contain features such as ARI (Automated Readability Index), LIX and CLI (Coleman-Liau Index). If the readability scores are low, the document is considered low effort. Similarly, low numbers in the documents' scores for the simple document features indicate low effort and vice versa. However, there isn't a standardized approach to determine what score constitutes to low or high effort.

Hence, a document will be classified as low effort using the boxplot approach, whereby documents are graded based on the scores within 0 and the upper inner fence to create binary relevance. Any suspected outliers and outliers are classified as high effort or non-relevant. The feature scores could only start from 0 and will not hold negative values, as it does not make sense to have negative scores. Each feature has separate effort-based relevance judgment to evaluate the retrieval systems in a test collection.

On the other hand, the readability effort features have standardized grades and will be classified in a possible equivalent manner. For both ARI and CLI, high effort constitutes to 'college' level. The LIX classifies 'very difficult' as high effort with the assumption this level is equivalent to 'college' level from CLI and ARI features. Therefore, all three readability effort features tend to classify the documents' effort in a somewhat equivalent manner.

# 4. EVALUATING RETRIEVAL SYSTEMS USING LOW EFFORT RELEVANCE JUDGMENTS AND REDUCED TOPIC SIZE

The first experimentation is the evaluation of retrieval systems using low effort relevance judgments conducted for different depth of evaluation. The evaluation depths attempted are 10, 100 and 1000. The purpose of different evaluation depth is to determine the variation in system rankings due to eqrel with deeper evaluation. For each of this evaluation depth, the system effectiveness is calculated using qrel, and eqrel for each feature. The correlation coefficient of system ranks is performed between the same evaluation depth from both qrel and eqrel.

The next experimentation measures the effectiveness of evaluating IR systems using low effort relevance judgment with reduced topic sizes. A topic size reduction should reduce the work in generating relevance judgment without jeopardizing the quality of evaluation.

Figure 1 shows the evaluation method for reduced topic size. Firstly, obtain the system ranks from the qrel for metrics AP@k (k = 10, 100, 1000). The original system ranks consider all 50 topics for each system. These will be used as the baseline system ranks for comparison. Next, randomly select a reduced topic size, for example, 30 topics. Compute the effectiveness scores for the systems using these 30 topics with the qrel. Then, perform

Kendall's tau correlation coefficient between the system ranks from baseline and reduced topic size from the qrel.

Then, use the same topics from the qrel, compute the system effectiveness scores using eqrel instead. Rank the systems and perform Kendall's tau correlation coefficient between the system ranks from baseline and the reduced topic size from eqrel. For each eqrel, a different set of random topics is selected. However, the same topics are used in the qrel to obtain the effectiveness, system rank and correlation coefficient for the specific topic size evaluation. Since a different set of topics evaluate the retrieval systems differently [5], the same topics are selected for evaluating with qrel and eqrel to measure only the changes due to effort features. Whereas, the random selection of topics for each feature is to allow different combinations of topics to measure the retrieval systems.

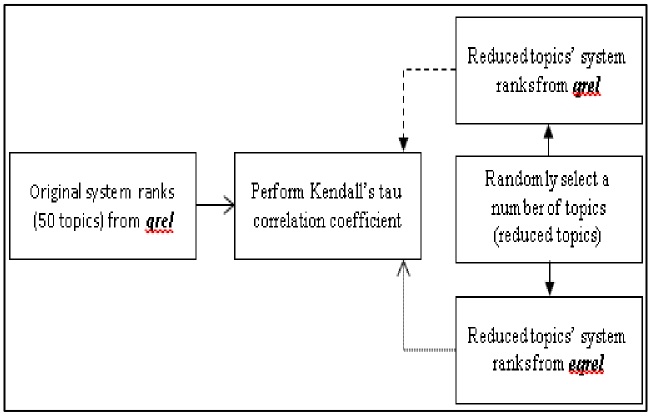
Reduced topics' system ranks from qrel
Original system ranks (50 topics) from qrel
Perform Kendall's tau correlation coefficient
Randomly select a number of topics (reduced topics)
Reduced topics' system ranks from eqrel

**Figure 1. Evaluation approach for reduced topic size.**

The Kendall's tau values from reduced topic size using qrel and eqrel will be evaluated to determine the effectiveness of using eqrel for reduced topic size. If Kendall's tau value from eqrel is equally good or better than Kendall's tau value from qrel, the reduced topic could be a good alternative in measuring the system effectiveness using low effort relevance judgments.

## 5. RESULTS AND DISCUSSIONS

The results for the first objective of this study is to measure the variation due to the depth of evaluation using the effort-based relevance judgments. In Figure 2, evaluation depths are 10, 100 and 1000, against the topic sizes 20, 30 and 40. The plots are shown for simple document features and readability features. The lines represent the evaluation using qrel and eqrel, although the qrel line is plotted for reference purposes.

From Figure 2, it can be observed that the trend for both qrel and eqrel are very similar despite the different features. When the evaluation depth is increased from 10 to 100, Kendall's tau values increase for 78% (14 out of 18) of the plots using qrel. As for the plots using eqrel, Kendall's tau values increase for 72% (13 out of 18) of the plots when the evaluation depth increases from 10 to 100.

The variation in evaluation depth from 100 to 1000 using qrel increases Kendall's tau values for 61% (11 out of 18) of the plots. Similarly, Kendall's tau values increase for 72% (13 out of 18) of the plots using eqrel when evaluation depth is increased from 100 to 1000.

The changes that occur between the evaluation depth are not necessarily similar for qrel and eqrel. However, the difference in the trend using eqrel is very similar to using qrel for evaluation with increased depth. From the observations, the usage of evaluation depth 100 or 1000 could produce better evaluation compared to depth 10 using the effort-based relevance judgments.

Figure 3 shows the plots for evaluating the IR systems us-ing qrel and eqrel with reduced topic sizes. The various features are shown at the top of the graph and their respective topic sizes (40, 30, 20) are shown at the bottom of the graph. The plots are shown for AP@100 and AP@1000 for the evaluation done using qrel and eqrel. Based on the observations, topic size 40 produces better evaluation compared to the other smaller topic sizes. However, some instances the topic size 30 yields better evaluation compared to topic size 40. The mentioned scenario occurred for evaluation depth 1000 for readability features regardless of the usage of qrel or eqrel for the evaluation.

However, Kendall's tau values are mostly moderate for topic sizes 40 and 30 except for LIX feature. In fact, simple document features for AP@1000 have strong Kendall's tau values for topic sizes 40 and 30. Despite the variation in topic sizes, the results produced are comparable, as stated by [3]. Recall that different topics of the same size could produce different results while some topics may be more suitable for evaluating the retrieval systems [3],[5].

It is important to highlight at this point that for each feature, different combinations of topics for each topic set were used. It is unlikely that any two topic sets were the same. The difference in topics could have been a major contributor to variations in the system ranking evaluations.

## 6. CONCLUSION

This study has shown that the evaluation of IR systems can be measured effectively due to changes in evaluation depth and topic size using low effort relevance judgment. A standard evaluation depth is recommended for evaluating IR systems using the low effort relevance judgment, although shallow depth of evaluation could produce sufficiently good evaluation outcome. However, increased evaluation depth produces a better evaluation outcome compared to a shallow depth. The reduced topic sizes also appear to be suitable for the evaluation of IR systems using low effort relevance judgments.

Further experimentations could be conducted to include additional metrics to observe changes in correlation coefficient due to evaluation depth or topic sizes. Additionally, other test collections could also be added to determine the similarity in the outcome.

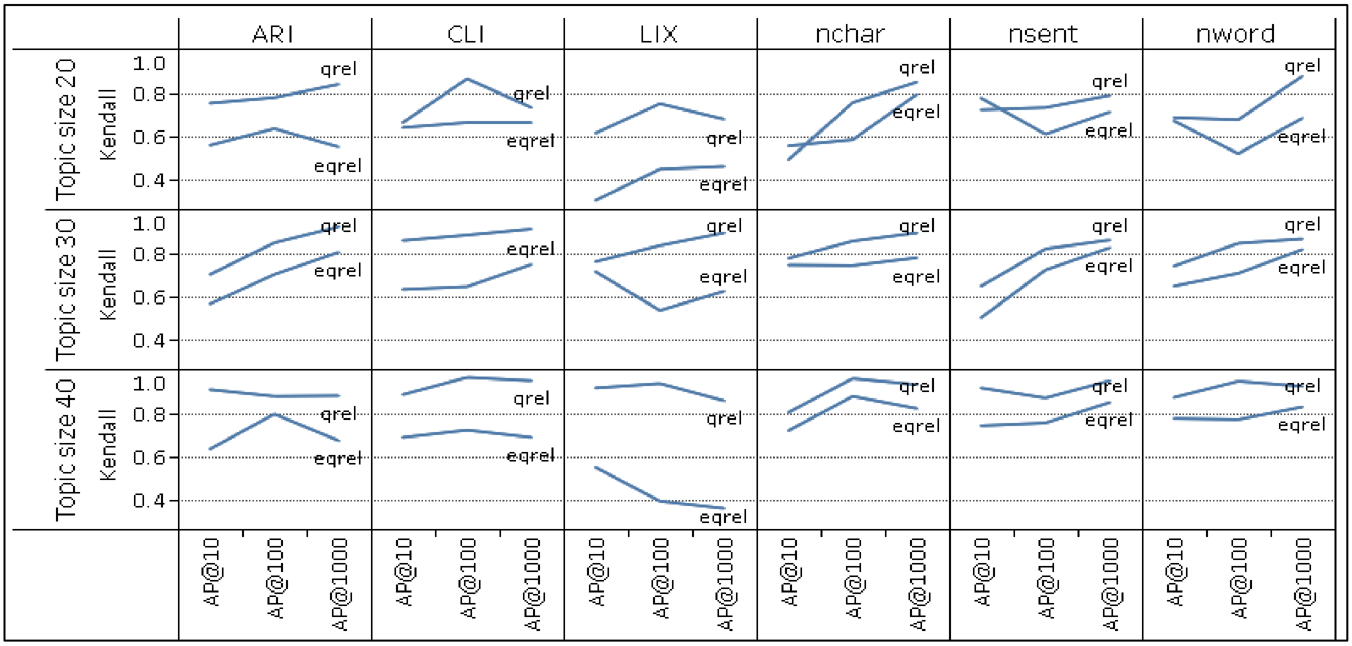

**Figure 2. Variation in correlation coefficient due to changes in evaluation depth.**

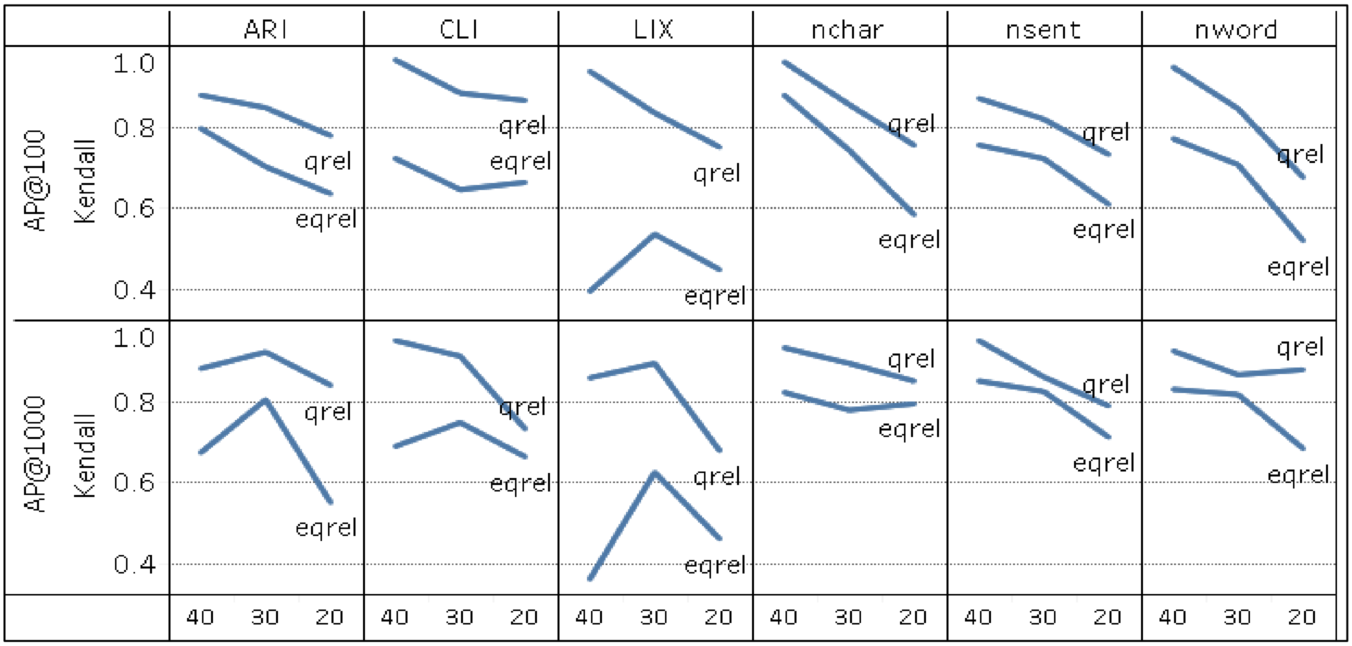

**Figure 3. Variation in correlation coefficient due to reduced topic size.**

## 7. ACKNOWLEDGMENTS

This research was supported by UMRG Programme RP059A-17SBS from University of Malaya.

## 8. REFERENCES

[1] Lohrey, J. (2019). The Importance of Information Storage & Retrieval Systems in an Organization. Retrieved April 13, 2019, from https://smallbusiness.chron.com/importance-information-storage-retrieval-systems-organization-75891.html

[2] What is Document Retrieval and How Does it Improve Your Organization? (n.d.). Retrieved April 13, 2019, from http://www.docuvantage.com/document-retrieval

[3] J. Guiver, S. Mizzaro, and S. Robertson, "A few good topics: Experiments in Topic Set Reduction for Retrieval Evaluation," *ACM Trans. Inf. Syst.*, vol. 27, no. 4, pp. 1–26, 2009

[4] C. Hauff, D. Hiemstra, F. Jong, and L. Azzopardi, "Relying on topic subsets for system ranking estimation," in

*Proceeding of the 18th ACM conference on Information and knowledge management CIKM 09*, 2009, pp. 1859–1862.

[5] E. M. Voorhees and C. Buckley, “The effect of topic set size on retrieval experimental error,” in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR’02*, 2002, pp. 316–323.

[6] S. Culpepper, S. Mizzaro, M. Sanderson, and F. Scholer, “TREC: Topic Engineering Exercise,” in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2014, pp. 1147–1150.

[7] W. Hersh *et al.*, “Do batch and user evaluations give the same results?,” in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR’00*, 2000, pp. 17–24.

[8] A. H. Turpin and W. Hersh, “Why batch and user evaluations do not give the same results,” in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR’01*, 2001, pp. 225–231.

[9] A. Moffat, F. Scholer, and P. Thomas, “Models and metrics:IR Evaluation as a User Process,” in *Proceedings of the Seventeenth Australasian Document Computing Symposium on - ADCS ’12*, 2012, pp. 47–54.

[10] A. Al-Maskari and M. Sanderson, “A Review of Factors Influencing User Satisfaction in Information Retrieval,” *J. Assoc. Inf. Sci. Technol.*, vol. 61, no. 5, pp. 859–868, 2010.

[11] R. Villa and M. Halvey, “Is relevance hard work?: Evaluating the effort of making relevant assessments,” in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR’13*, 2013, pp. 765–768.

[12] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey, “Relevance and Effort: An Analysis of Document Utility,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM ’14*, 2014, pp. 91–100.

[13] M. Verma, E. Yilmaz, and N. Craswell, “On Obtaining Effort Based Judgements for Information Retrieval,” in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM ’16*, 2016, pp. 277–286.

[14] P. Chandar, W. Webber, and B. Carterette, “Document Features Predicting Assessor Disagreement,” in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 745–748.

[15] E. M. Voorhees, “The Philosophy of Information Retrieval Evaluation,” in *Evaluation of Cross-Language Information Retrieval Systems. CLEF 2001. Lecture Notes in Computer Science, vol 2406*, 2002, pp. 355–370.

[16] A. Berto, S. Mizzaro, and S. Robertson, “On Using Fewer Topics in Information Retrieval Evaluations,” in *Proceedings of the 2013 Conference on the Theory of Information Retrieval - ICTIR ’13*, 2013, pp. 30–37.

[17] I. Soboroff, C. Nicholas, and P. Cahan, “Ranking retrieval systems without relevance judgments,” in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR’01*, 2001, pp. 66–73.

[18] 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-SIGIR’01*, 2001, pp. 74–82.

[19] E. M. Voorhees, “Variations in relevance judgments and the measurement of retrieval effectiveness,” *Inf. Process. Manag.*, vol. 36, no. 5, pp. 697–716, 2000.

[20] T. Jones, A. Turpin, S. Mizzaro, F. Scholer, and M. Sanderson, “Size and Source Matter: Understanding Inconsistencies in Test Collection-Based Evaluation,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM’14*, 2014, pp. 1843–1846.