Exploring Topic Difficulty in Information Retrieval Systems Evaluation

Wei Ting Pang, Prabha Rajagopal, Mengjia Wang, Shuxiang Zhang and Sri Devi Ravana*

1Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.
2Accenture Solutions, Kuala Lumpur, Malaysia.
3Department of Computer Science, University of Auckland, New Zealand.
*sdevi@um.edu.my

Abstract. Experimental or relevance assessment cost as well as reliability of an information retrieval (IR) evaluation is highly correlated to the number of topics used. The need of many assessors to produce equivalent large relevance judgments often incurs high cost and time. So, large number of topics in retrieval experiment is not practical and economical. This experiment proposes an approach to identify most effective topics in evaluating IR systems with regards to topic difficulty. The proposed approach is capable of identifying which topics and topic set size are reliable when evaluating system effectiveness. Easy topics appeared to be most suitable for effectively evaluating IR systems.

1. Introduction

As the Internet continues to develop, the amount of information in the network is growing rapidly and dramatically. A way to retrieve information which satisfies the user’s needs from such a huge information pool using traditional cataloguing techniques is no longer working. Therefore, search engines had become the most popular retrieval tools used to locate the most relevant information from the Internet or any digital database to meet users’ information needs (expressed in a query form). The most commonly used information retrieval (IR) systems are Web search engines, such as Google, Baidu, Yahoo and Bing.

In a typical information retrieval, the user will submit a query representing their information needs in a search box, and the search engine retrieves relevant documents from the database. Then, a list of links to the documents that are considered relevant by the search engine will be returned to the users. Then the retrieved documents will be matched with the relevance judgments to determine the relevancy of the document found with some sort of numeric scale. The numeric scale is either in a form of binary or graded. In binary relevancy, each document will be classified as either “relevant” or “not relevant” to each query. Hence, when a system is evaluated based on its retrieved relevant documents in the binary approach its ranks will be determined first, followed by how many of the relevant documents were found. While in graded relevance, there are more degrees of relevance with a categorical scale of “marginally relevant”, “somewhat relevant” or “highly relevant”. With this approach, the smaller the
rank of the highly relevant documents as compared to that of the marginally relevant document, the better it is because the order of the relevant documents can be measured.

Existing researches have raised concerns on the computational time and large number of topic combination needed for IR evaluation [1][2]. Carterette et al. [1] and Berto, et al. [3], stated that it is possible to obtain good or similar results when evaluating system with reduced topics. Also, not all topics in a test collection can measure system performance effectively because topic difficulty may affect system performance or ranking. To further explore in the area of topic reduction, this study experiments on topic difficulties and reduced number of topics for IR evaluation. The objective of this paper is to identify the most effective topics in evaluating IR system in terms of topic difficulties.

The next section includes past literature on topic difficulties and topic selection. Next, Section 3 includes the calculation for topic difficulty score and Section 4 states the topic selection. Then, the results and discussion section follows. Finally, conclusions are drawn and future works are mentioned.

2. Literature

This section contains two sub-sections discussing on the suitable topic sizes for experimentation and determining topic difficulty.

2.1. Topic Size

Topic size reduction in IR experiments has become a concern in recent research because the number of topics used is generally correlated with relevance assessment cost as well as the power and reliability of a retrieval test collection [4][5]. It is known that; each test collection is costly to be produced. Therefore, fewer topics in retrieval experiment is more economical [5][6]. Number of topics to be used has often been raised in discussions and researchers are trying to investigate and determine which topic size is able to give reliable, robust and good result.

In the early stages, researchers stated that 75 topics were too little to be used while 250 was acceptable and sometimes 1000 were needed [7]. As the years passed, it was stated that 50 topics can produce a reliable result [6] but 25 topics were also sufficient and able to perform reasonably [8][9][10]. In another study conducted, system rankings based on five or ten topics were considered not safe, while rankings based on 25 topics were much safer and more stable [11]. In contrast, topic ranges between five to twenty was considered reasonable [3]. In another research, the sample size required to provide 95% confidence on a declared significance may be as small as ten [12]. Therefore, it is possible to obtain good or similar results when evaluating system effectiveness even if the number of topics used is decreasing [2][3].

Nonetheless, error is inevitable even when the topic set size is 50, which is consider as a standard size in IR experimentation. Besides, evaluation can become unstable if measured using simple effectiveness metrics such as precision at ten (P@10) [4]. [13] stated different evaluation measures may require different topic set sizes under the same set of statistical requirements. Therefore, choosing the right effectiveness metrics is an important choice because some metrics may perform better than others on certain topic set size.

2.2. Determining topic difficulty

Determining topic difficulty is one of the important factors in IR system evaluation because it may affect system ranking to some extent. Therefore, topic difficulty distribution should not be biased, as it is neither too difficult nor too easy, in order to keep reliability of test collections [14]. Based on past studies, topic difficulty can be determined using average of average precision (AAP) for each topic across all systems. In this study, AAP method will also be used to determine the topic difficulty.

High average precision and high AAP of a system on a certain topic means that the system is performing effectively on the topic and it is an easy topic, respectively. On the other hand, low average precision would mean that system is performing badly on the topic but the topic is neither a difficult or hard topic [14][17].
Apart from AAP value, a past study measured the hardness of a topic using the best average precision that any system has achieved on the topic [18]. The study concluded that there exist hard topics that are hard for all the systems. In another study, difficult topics were classified as being those with a low median average precision score and at least one high outlier score [19]. Equation (1) was proposed to define topic difficulty where $D_t$ is difficulty of topic and $t$ is the topic [20].

$$D_t = \frac{\text{max}_t - \text{mean}_t}{\text{sd}_t}$$  

(1)

A study conducted by Mandl (2009) showed that easier topics had higher influence on the final result or in other words, it can be said that easy topics correlate well with the result. Therefore, if topic difficulty is known prior to performing an evaluation, this will decrease the amount of resource used by directly selecting easier topics to determine the evaluation results. In addition, the study also states that hard topics are good in discriminating between systems [17], however, another study states easy topics discriminate most and least effective systems [21].

3. Calculation of Topic Difficulty (TD) Score

In this experiment, Topic Difficulty (TD) score will be used to determine the difficulty of a topic as well as to select topics based on the score. The formula for TD score is shown in Equation (2) where, $\sum AP_t$ is the summation of all systems average precision value of certain topic, which is also known as AAP and $\# \text{ of systems}$ is the total number of systems. Higher TD score denotes easier topics, while low TD score denotes difficult topic.

$$TD\, score = \frac{\sum AP_t}{\# \text{ of systems}}$$  

(2)

AAP is known as average of average precision for each topic across systems. Table 1 shows the AAP formula for various topics.

| Topics | Systems |
|--------|---------|
| $t_1$  | $s_1$   | $s_2$   | ... | $s_n$ | AAP  |
| $t_2$  | $s_1$   | $s_2$   | ... | $s_n$ | AAP  |
| ...    | ...     | ...     | ... | ...   | ...  |
| $t_n$  | $s_1$   | $s_2$   | ... | $s_n$ | AAP  |

Assuming there are only two systems with average precision for $s_1 t_1$ is equal to 0.1, and $s_2 t_1$ is equal to 0.2, the TD score of 0.15 is obtained when computed using the formula in Table 1.

$$AP(s_1, t_1) = 0.1$$

$$AP(s_2, t_1) = 0.2$$

$$TD\, score \, for \, t_1 = \frac{(0.1 + 0.2)}{2} = 0.15$$

3
4. Modified Classic method

Figure 1 shows the modified classic method flow used in this research. Firstly, \( k \) topics are chosen randomly for evaluation. Next, with the topic chosen, the AP and MAP value is calculated with all systems. Then, with the MAP value obtained, comparison is made with the golden MAP to determine how close is the chosen topic MAP is to the truth. The whole process is repeated 1000 times to obtain the mean value of Kendall’s Tau. The variable \( k \) has a minimum of five and maximum of twenty topics.

![Figure 1. Modified classic method flow](image)

\*\( k = 5,6,7\ldots 20 \)

5. Results and discussion

The Kendall’s Tau correlation coefficient was computed for different types of topic from different test sets (Set249, Subset A & Subset B) and the modified classic to identify the types of topic that are most effective in evaluating IR systems. The modified classic is used as the baseline, represented by the red line and square icon in Figure 2 and Figure 3. The Subset A and Subset B contains 50 topics respectively that were randomly chosen from 249 topics of TREC 2004 Robust Track and includes 110 systems. Both the subsets contain different topics. Recall that a high Kendall’s Tau value denotes the value is nearer to the actual truth and represents best results.

![Graph showing Kendall’s Tau correlation coefficient for different topics](image)
Figure 2. Kendall’s Tau Value of Subset A Topics with Different Topic Types Based on TD

The Kendall’s Tau correlation coefficient of Subset A for different topic set size and different types of topic difficulties, and the baseline are shown in Figure 2. From the figure, a noticeable difference can be observed between the three types of topic difficulties with that of the baseline. Based on the experiment conducted on Subset A, the easy topic with any set size resulted a higher Kendall’s Tau value compared to modified classic. On the other hand, larger topic set size with moderate topic difficulties (middle topic) did not result in higher Kendall’s Tau value than the modified classic. Contrary to topic set size with middle topic hardness, hard topic with larger topic set size gave a higher Kendall’s Tau value compared to that of the baseline.

In Subset B, easy topic also performed well compared to the other types of topic as shown in Figure 3. As for the middle topic with larger topic set, a higher Kendall’s Tau value compared to the modified classic is observed. Moreover, only a few topic sets with hard topic were higher than the modified classic while the remaining were lower than the baseline. Both subsets have similar outcomes when compared to the baseline.

Figure 3. Kendall’s Tau Value of Subset B Topics with Different Topic Types Based on TD Score

6. Conclusion

In system-based IR evaluation experimentation, large number of topic combination is a problem as it indirectly translates to high computational cost. In this study, the research objective was achieved using Topic Difficulty (TD) Score to determine the difficulty of a topic and identifying suitable topic types for effective IR system evaluation. Easy topics are most suited for evaluating IR systems effectively.

There is limitation in this study such that some numbers of topics only were selected for the experimentation although there could be many possible combinations of topics. In future, further experimentations on using easy topics to evaluate IR systems could be conducted. Variations to the usage of easy topics could provide further insights to the effectiveness in evaluating IR systems.
7. Acknowledgment

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

References

[1] Carterette, B., Allan, J., and Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 268-275. ACM, August.

[2] Guiver, J., Mizzaro, S., and Robertson, S. (2009). A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, 27(4), 21.

[3] Berto, A., Mizzaro, S., and Robertson, S. (2013). On using fewer topics in information retrieval evaluations. *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, p. 9. ACM, September.

Voorhees, E. M. (2009). Topic set size redux. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 806-817. ACM, July.

[4] Zhang, S., and Ravana, S. D. (2016). Estimating reliability of the retrieval systems effectiveness rank based on performance in multiple experiments. *Cluster Computing*, 1-16.

[5] Sakai, T. (2015). Topic set size design. *Information Retrieval Journal*, 128.

[6] Voorhees, E.M. and Buckley, C. (2002). The effect of topic set size on retrieval experiment error. *Proceedings of 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 316–323. ACM Press, New York.

[7] Sparck Jones, K., and van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of documentation*, 32(1), 59-75.

[8] Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments?. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 307-314. ACM, August.

[9] Sanderson, M. and Zobel, J. (2005). Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. *Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval (SIGIR)* Salvador, Brazil. 162-169.

[10] Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. *SIGIR 2006*, pages 525–532.

[11] Voorhees, E.M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697 716.

[12] Sakai, T., 2016. Topic set size design. *Information Retrieval Journal*, 19(3), pp.256-283.

[13] Dinçer Bekir Taner, 2013. Design of information retrieval experiments the sufficient topic set size for providing an adequate level of confidence. Turkish Journal of Electrical Engineering & Computer Sciences

[14] Eguchi, K., Kando, N. and Kuriyama, K. (2002). Sensitivity of IR Systems Evaluation to Topic Difficulty. *Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002)* 585-589.

[15] Kwok, K. (2005). An Attempt to Identify Weakest and Strongest Queries. *SIGIR Workshop Predicting Query Difficulty*. http://www.haifa.il.ibm.com/sigir05-qp.

[16] Mizzaro, S. (2008). The good, the bad, the difficult, and the easy: Something wrong with information retrieval evaluation? *Proc. 30th European Conf. on Information Retrieval*, Glasgow, Scotland, March-April.

[17] Mandl, T. (2009). Easy Tasks Dominate Information Retrieval Evaluation Results. *BTW* (pp. 107-116).
[18] Hu, X., Bandhakavi, S. and Zhai, C. (2003). Error analysis of difficult TREC topics. *Proceedings of 26th ACM International Conference on Research and Development in Information Retrieval (SIGIR ’03)*, Toronto, Canada, pp. 407-8.

[19] Voorhees, E. M. (2003). Overview of the TREC 2003 Robust Retrieval Track. *Proc. 12th Text REtrieval Conference (TREC 2003)*, Gaithersburg, Maryland, November.

[20] Ravana, S. D., and Moffat, A. (2008). Exploring evaluation metrics: GMAP versus MAP. Proceeding SIGIR ’08 Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval Pages 687-688 Singapore, July 20 – 24.

[21] Roitero K., Maddalena E., Mizzaro S. (2017). Do Easy Topics Predict Effectiveness Better Than Difficult Topics?. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2017. Lecture Notes in Computer Science, vol 10193. Springer, Cham.