Using Global Statistics to Rank Retrieval Systems without Relevance Judgments

Zhiwei Shi[1], Bin Wang[1], Peng Li[1], Zhongzhi Shi[2]

[1] Information Retrieval Group, Center for Advanced Computing Research,
Institute of Computing Technology, CAS, Beijing, 100190, China
{shizhiwei, wangbin, lipeng01}@ict.ac.cn
[2] Key Lab of Intelligent Information Processing,
Institute of Computing Technology, CAS, Beijing, 100190, China
shizz@ics.ict.ac.cn

Abstract. How to reduce the amount of relevance judgments is an important issue in retrieval evaluation. In this paper, we propose a novel method using global statistics to rank retrieval systems without relevance judgments. In our method, a series of global statistics of a system, which indicate the percentage of its documents found by k out of all the N systems (k = 1, 2, …, N), are selected, then a linear combination of the series of global statistics is utilized to fit the mean average precision (MAP) of the retrieval system. Optimal coefficients are obtained by linear regression. No human relevance judgments are required in the entire process. Compared with existing evaluation methods without relevance judgments, our method has two advantages. Firstly, it outperforms all early attempts. Secondly, it is adjustable for different effectiveness measurements, e.g. MAP, precision at n, and so forth.

Keywords: Information retrieval, evaluation, without relevance judgments, regression.

1 Introduction

Generally, to compare the effectiveness of information retrieval systems, we need to prepare a test collection composed of a set of documents, a set of query topics, and a set of relevance judgments indicating which documents are relevant to which topics. Among these requirements, relevance judgment is the most human resource exhausting and time consuming part. It even becomes incapable when the test collection is extremely large. To address this problem, the TREC conferences used a pooling technology [10], where the top n (e.g., n=100) documents retrieved by each participating system are collected into a pool and then only the documents in the pool are judged for system comparison. Zobel [12] has shown that this pooling method leads to reliable results in term of determining the effectiveness of retrieval systems and their relative rankings. Yet, the relevance determination process is still very resource intensive especially when the test collection reaches or exceeds terabyte, or much more queries are included. More seriously, when we change to a new document collection, we have to redo the entire evaluation process.

There are two possible solutions to the problem above, evaluation with incomplete relevance judgments and evaluation without relevance judgments. The former is well studied. Many well designed ranking methods with incomplete judgments were carried out. Two of them, Minimal Test Collection (MTC) method [4] and Statistical evaluation (statMAP) method [2], even got practical application in the Million Query (1MQ) track in TREC 2007 [1], and achieved satisfactory evaluation performance. The latter is comparatively less studied. Only a few papers concentrate on the issue of evaluating retrieval systems without relevance judgments. In Section 2 of this paper, we will briefly review some representative methods. We will see what they are and how they work.

In this paper, we focus our effort on the retrieval evaluation without relevance judgments. Although 'blind' evaluation is really a hard problem and its evaluation performance is far less than that of methods with incomplete judgments, it is undeniable that non-judgment evaluation has its own advantages. In some cases, relevance judgments are non-attainable. For example, when researchers compare their novel retrieval algorithms to existing methods, or search for optimal parameters of their algorithms, or conduct data fusion in a dynamic environment, relevance judgment usually seems impossible. Besides, to construct a good evaluation method without relevance judgments, researchers need to mine the retrieval results thoroughly, and try to find laws that indicate the correlation between the effectiveness of a system and features of its retrieval result. These laws are not only useful for 'blind' evaluation methods but also valuable for evaluation methods with incomplete judgments.

The main contribution of this paper is that we propose a non-judgment retrieval evaluation method using global statistics of retrieval results, where a linear combination of a series of global statistics of a retrieval system is utilized as an indicator of its retrieval performance. Details of this method will be presented in Section 3. Experimental results, which are reported in Section 4, demonstrate that the proposed method outperforms all the existing methods without relevance judgments. Finally, we conclude our work in Section 5.

2 Related Work

In 2001, Soboroff et al. [6] firstly proposed the concept of evaluating retrieval systems in the absence of relevance judgments. They generated a set of pseudo-relevance judgments by randomly selecting and declaring some documents from the pool of top 100 documents as relevant. This set of pseudo-relevance judgments (instead of a set of human relevance judgments) was then used to determine the effectiveness of the retrieval systems. Four versions of this random pseudo-relevance method were designed and tested on data from the ad hoc track in TREC 3, 5, 6, 7 and 8. They were simple random pseudo-relevance method, the variant with duplicate documents, the variant with Shallow pools and the variant with Exact-fraction sampling. All their resulting system assessments and rankings were well correlated with actual TREC rankings, and the variant with duplicate documents in pools got the best performance, with an average Kendall's tau value 0.50 over the data of TREC 3, 5, 6, 7 and 8.

Soboroff et al.'s idea came from two results in retrieval evaluation. One is that incomplete judgments do not harm evaluation results greatly. Zobel's research [12] had showed that the results obtained using pooling technology were quite reliable given a pool depth of 100. He also found that even though the pool depth was limited to 10, the relative performance among systems changed little, although actual precision scores did change for some systems. The other is that partially incorrect relevance judgments do not harm evaluation results greatly. Voorhees [9] ascertained that despite a low average overlap between assessment sets, and wide variation in overlap among particular topics, the relative rankings of systems remained largely unchanged across the different sets of relevance judgments. These two points are bases of Soboroff et al.'s random pseudo-relevance method, and give explanation to the result that their rankings were positively related to that of the actual TRECs. As a matter of fact, the two points are bases of all the retrieval evaluation methods without or with incomplete relevance judgments.

Aslam and Savell [3] devised a method to measure the relative retrieval effectiveness of systems through system similarity computation. In their work, the similarity between two retrieval systems was the ratio of the number of documents in their intersection and union. Each system was scored by the average similarity between it and all other systems. This measurement produced results that were highly correlated with the random pseudo-relevance method. Aslam and Savell hypothesized that this was caused by 'tyranny of the masses' effect, and these two related methods were assessing the systems based on 'popularity' instead of 'performance'. The analysis by Spoerri [7] suggested that the 'popularity' effect was caused by considering all the runs submitted by a retrieval system, instead of only selecting one run per system. Our later experimental results will show that this point of view is partially correct. The 'popularity' effect could not be avoided completely by only selecting one run per system. This is indeed a hard problem for all the evaluation methods without relevance judgments.

Wu and Crestani [11] developed multiple 'reference count' based methods to rank retrieval systems. They made the distinction between an 'original' document and its duplicates in all other lists, called the 'reference' documents, when computing a document's score. A system's score is the (weighted) sum of the scores of its 'original' documents. Several versions of reference count method were carried out and tested. The basic method (Basic) scored each 'original' document by the number of its 'reference' documents. The first variant (V1) assigned different weights to 'reference' documents based on their ranking positions. The second variant (V2) assigned different weights to the 'original' document based on its ranking position. The third variant (V3) assigned different weights to both the 'original' documents and the 'reference' documents based on their ranking positions. The fourth variant (V4) was similar to V3, except that it normalized the weights to 'reference' documents. Wu and Crestani's method output similar evaluation performance to that of the random pseudo-relevance method. Their work also showed that the similarity between the multiple runs submitted by the same retrieval system affected the ranking process. If only one run was selected for any of the participant system for any query, for 3-9 systems, V3 outperformed random pseudo-relevance method by 45.6%; for 10-15 systems, random pseudo-relevance method outperformed V3 by 6.5%.

Nuray and Can [5] introduced a method to rank retrieval systems automatically using data fusion. Their method consists of two parts. One is selecting systems for data fusion, and the other is selecting documents as pseudo relevant documents as the fusion result. In the former part, they hypothesized that systems returning documents different from the majority could provide better discrimination among the documents and systems. In return, this could lead to a more accurate pseudo relevant documents and more accurate rankings. To find proper systems, they introduced the 'bias' concept for system selection. In their work, bias was 1 minus the similarity between a system and the majority, where the similarity is a normalized dot product of two vectors. In the latter part, Nuray and Can tested three criterions, namely Rank position, Borda count and Condorcet. Experimental results on data from TREC 3, 5, 6 and 7 showed that bias plus Condorcet got the best evaluation results and it outperformed the reference count method and random pseudo relevance method greatly.

More recently, Spoerri proposed a method using the structure of overlap between search results to rank retrieval systems. This method provides us a new view on how to rank retrieval systems without relevance judgments. He used local statistics of retrieval results as indicators of relative effectiveness of retrieval systems. Concretely, if there are N systems to be ranked, N groups are constructed randomly with the constraint that each group contains five systems and each system will appear in five groups; then the percentages of a system's documents not found by other systems (Single%) as well as the difference between the percentages of documents found by a single system and all five systems (Single%-AllFive%) are calculated as indicators of relative effectiveness respectively. Spoerri found that these two local statistics were highly and negatively correlated with the mean average precision and precision at 1000 scores of the systems. By utilizing the two statistics to rank systems from subsets of TREC 3, 6, 7 and 8, Spoerri obtained appealing evaluation results. The overlap structure of the top 50 documents were sufficient to rank retrieval systems and produced the best results, which outperformed previous attempts to rank retrieval systems without relevance judgments significantly.

So far, we have reviewed 5 representatives of non-judgment evaluation methods. Among these methods, Single% method proposed by Spoerri [8] is the most appealing one. Its average Spearman's rank correlation coefficient achieves 0.80 over data of TREC 3, 6, 7 and 8. More meaningfully, Spoerri's method provides us a new view of what information in retrieval results is more valuable for system ranking. Only the random grouping is a little bit confusing. Following study will show that more explicit information can be used in non-judgment retrieval evaluation.

3 Methodology

In this section, we will introduce our method for ranking retrieval systems using global statistics. Basically, our idea comes from the careful study of Spoerri's work in 2007 [8]. We find that the expectation of local statistics utilized in Spoerri's research, e.g. Single%, is actually a linear combination of a series of global statistics. So, why don't we seek for a series of optimal coefficients to make the combination better fit the measurement of systems' retrieval effectiveness, e.g. MAP or some measurement else? Here comes our idea of ranking retrieval system with global statistics. Before we go into more details of our method, let us check the statistics in Spoerri's work first.

We have just described Spoerri's method in the previous section. The statistics 'Single%' is the percentage of documents found by a single system and not by other four in a random group. Apparently, 'Single%' is a local statistics, for it involves five systems in a random group. In Spoerri's work, the value of this local statistics is obtained experimentally. More concretely, for a given system, 'Single%' is calculated on each of the five random groups containing this system and each of the 50 topics, then these 'Single%' values are averaged. Obviously, if we replace the average value of 'Single%' with its expectation, the result will remain the same, or become statistically more accurate. Now we check the expectation of 'Single%'.

Suppose that we have N systems, each of which is a document list. Consider a given system and a random group containing it. This means that we have a certain system and four other random systems in the group. For any document that is found by the given system, if it has ever appeared in k out of N systems (k = 1, 2, …, N), the probability that it appears in the group as 'single' is:

$$p_k^{(1)} = \frac{C_{k-1}^{0} \cdot C_{N-k}^{4}}{C_{N-1}^{4}} \tag{1}$$

This can be interpreted as the probability that we pick 0 out of k-1 systems that contain the document, 4 out of N-k systems that do not contain the document and the given system to form the group.

Thus, by the law of total probability, we have the expectation of 'Single%' as follows:

$$E(Single\%) = \sum_{k=1}^{N} p_k^{(1)}(N_k\%) \tag{2}$$

where $p_k^{(1)}$ is described in formula (1), and $N_k\%$ is the percentage of the given system's documents found by k (including the given system) out of all the N systems ($k = 1, 2, \ldots, N$). Notice that N_k % is a global statistics opposite to local statistics.

Similarly, we can write the expectations of 'AllFive%' and other local statistics in the form like formula (2):

$$E(AllFive\%) = \sum_{k=1}^{N} p_k^{(5)}(N_k\%) \tag{3}$$

and

$$E(Single\% - AllFive\%) = \sum_{k=1}^{N} (p_k^{(1)} - p_k^{(5)})(N_k\%) \tag{4}$$

where

$$p_k^{(5)} = \frac{C_{k-1}^{4} \cdot C_{N-k}^{0}}{C_{N-1}^{4}} \tag{5}$$

Now we get that the expectation of local statistics used in Spoerri's method is actually a linear combination of a series of global statistics. With formula (2) and (4), we do not need to consider the random grouping any more. If we have these global statistics from the retrieval document lists, we can obtain the expectations of statistics used in Spoerri's method.

Here comes the question. A linear combination of these global statistics with fixed coefficients can be a good indicator of system's relative effectiveness, what if we replace the fixed coefficients with the optimal coefficients? It will definitely produce better system rankings. Besides, the optimal coefficients can be tunable. Different coefficients could be optimized corresponding to different effectiveness measurements, e.g. MAP, precision at n, or any sound measurements. This is our idea.

To make our method experimentally comparable to early methods without relevance judgments, we will use the MAP measurement as the target of our optimization in this work. That is, we are seeking for a series of coefficients $a_1, a_2, \ldots, a_M$, so that we can minimize the sum of squares of errors with the true MAP:

$$\sum_{i=1}^{N} (y_i - MAP_i)^2 \tag{6}$$

where MAP_i is the MAP value of the ith system and y_i is defined as:

$$y_i = \sum_{k=1}^{M} a_k (N_k^{(i)}\%) \tag{7}$$

where a_k is the coefficient to be optimized, and $N_k^{(i)}\%$ is the percentage of the ith system's documents found by k out of all the N systems ($k = 1, 2, \cdots, M, M \leqslant N$). By using linear regression, we can easily get these optimal coefficients. In turn, we calculate y_i for the ith system and obtain their rankings eventually.

Typically, when devising methods for retrieval evaluation without relevance judgments, researchers often seek for some law(s) inside a small part of data and apply the law(s) on the entire data set to see whether it works well. Accordingly, we will generate 5 series of coefficients optimized based on the data from TREC 3, 5, 6, 7 and 8 respectively, and examine their ranking performances on all the 5 data sets. Each of the 5 series of coefficients is in fact an implementation of our evaluation method. R3, R5, R6, R7 and R8 are short for these 5 series of coefficients as well as their corresponding ranking methods, where Rx means the method comes from TREC x (x=3, 5, 6, 7, 8).

4 Experimental Results

4.1 Some Clarification

Before we come to the experimental results, we would like to make some details clear first.

Firstly, in our experiments, the value of M in formula (7) is set to 30. The number of systems (runs), N, varies in different TREC data (see Table 1 for details). To make our method optimized based on one TREC data capable for being applied to others, we need a fixed number of M, which fits for all N of TREC 3, 5, 6, 7, 8. We also noticed that as parameter k goes from 1 to N, the statistics $N_k^{(i)}\%$ decreases rapidly to zero. A fixed number of M, if not too small, will not make the model lose too much information. 30 is a good but not only choice for parameter M. It fits all TREC data, and is not too small.

Table 1. Number of TREC runs.

TREC	Number of Runs
3	40
5	61
6	74
7	103
8	129

Secondly, different from Spoerri's work, we plan to rank all the systems for each TREC opposite to a subset of them. Without any limitation, we will definitely encounter the problem of 'popularity' effect mentioned previously in Section 2. To avoid this situation, when we calculate statistics $N_k^{(i)}\%$, different runs from same system will be counted only once.

Besides, to make a fair comparison, we need to repeat Spoerri's method over all systems for each TREC. The repetition is not exactly the same as Spoerri's original one. Based on the analysis in previous subsection, we replace the average of 'Single%' with the expectation of 'Single%', so that we can eliminate random turbulence in the original method. We process the statistics 'Single%-AllFive%' in the same way.

At last, the correlation between the rankings from our proposed methods, as well as other methods to be compared with, and the TREC official rankings (based on MAP) is measured using the Spearman's rank correlation coefficient. One reason is that it suits better for evaluating correlation between ratio sequences, e.g. MAP, than Kendall's tau. The other reason is that we can directly compare our results with those of previous attempts reviewed in Section 2, since most of them provided Spearman's rank correlation coefficient results.

4.2 Model Selection

Spoerri had stated that the overlap structure of the top 50 documents were sufficient to rank retrieval systems and produced the best results [8]. We just test our methods and Spoerri's methods with three pool sizes, namely 20, 30 and 50.

Among all our methods, R5 with pool size 20 produces the best result. It also works very stable. So we take R5 as the representative of our method. For Spoerri's methods, Single% with pool size 20 is selected as a representative, for it works slightly better than Single%-AllFive% on all systems.

4.3 Comparison with All Previous Attempts

We make a comparison between our method and all previous attempts. The comparison result is given in Table 2.

Table 2. Spearman's correlation coefficients for best results from different methods.

	RC	**RS**	**BC**	**SS**	**Single%**	**R5**
Trec3	0.587	0.627	**0.867**	0.751	0.824	0.716
Trec5	0.421	0.429	0.657*	0.488	0.563	**0.912**
Trec6	0.384	0.436	**0.717**	0.609	0.618	0.601
Trec7	0.382	0.411	0.453	0.551	0.550	**0.603**
Avg.3-7	0.444	0.476	0.674	0.600	0.639	**0.708**
Std.3-7	0.097	0.101	**0.210**	0.112	0.127	0.146
Trec8	-	-	-	**0.613**	0.569	0.514
Avg.3-8	-	-	-	0.602	0.625	0.**669**

In Table 2, RC is the best result produced by reference count method; RS represents the result of random pseudo relevance method, where relevance ratio is set to 10% rather than the actual ratio in its original version; BC accounts for the result of Bias plus Condorcet method, a data fusion based method. Results of these three methods are cited from Nuray and Can's paper [5]. They did not provide results on TREC 8, so we just have their results on TREC 3, 5, 6 and 7. For the number with a '*' (BC on TREC 5), in their original paper, same result in different tables conflict, and we pick the bigger number presenting in Table 5. SS is short for method based on system similarity. Since there is no Spearman's rank correlation coefficient result available in Aslam and Savell's work [3], we make an implementation of this method. In the implementation, we have tested several pool depths, where pool depth 100 produces the best result, thus is presented in Table 2. Single% is the representative of Spoerri's overlap structure based method, and R5 is the representative of our method.

Each line of Table 2 presents results from different methods on same TREC data. The bold number indicates the best result on a TREC data. We can see that over all five TREC data, BC method achieves best twice, SS method wins best once, and R5 method gets best twice. Especially, R5 method gets the best result on the two average evaluation performances. When averaging on TREC 3-7, R5 method outperforms the second best result (from BC with underline in Table 5) 5%. When averaging on TREC 3-8, R5 method outperforms the second best result (from Single% with underline in Table 5) 7%. In a word, regarding the average Spearman's correlation coefficients on TREC data 3, 5, 6, 7 and 8, R5 method outperforms all the existing retrieval evaluation methods that do not use human relevance judgments.

Besides, we find that all methods work quite unstably. The stand deviation of Spearman's correlation coefficients for all methods on TREC data 3, 5, 6 and 7 runs from 0.097 to 0.210. The better average evaluation result a method gets, the more instable it is. Our R5 method is an exception. It gets the best average result but the second large deviation.

5 Conclusion and Future Work

We end this paper with a conclusion reemphasizing the main points of our work.

In this work, we propose a retrieval evaluation method using global statistics of retrieval systems, where a linear combination of a series of global statistics of a retrieval system is utilized as an indicator of its retrieval performance. Compared with existing evaluation methods without relevance judgments, our method has two advantages. Firstly, it outperforms all early attempts regarding data on TREC 3, 5, 6, 7 and 8. Secondly, the method is adjustable for different effectiveness measurements, e.g. MAP, precision at n, and so forth. In contrast, some early attempts, e.g. reference account method, system similarity method and Single% method, can not change their scoring strategy to fit different effectiveness measurements.

The proposed method has its weakness as well. It works unstably on different data set, and mixes best systems with ordinary ones. This is also the common problem for all non-judgment evaluation methods. With meticulous analysis, we have found the fundamental factor that depresses the performance of non-judgment evaluation. How to tackle this problem is our future work.

Acknowledgments. This work is supported by the National Science Foundation of China under Grant No. 60776797, the Major State Basic Research Project of China (973 Program) under Grant No. 2007CB311103 and the National High Technology Research and Development Program of China (863 Program) under Grant No. 2006AA010105.

References

1. Allan J., Carterette B., Aslam J. A., Pavlu V., Dachev B., and Kanoulas E.: Overview of the TREC 2007 Million Query Track. In: Proceedings of TREC. (2007)
2. Aslam J. A., Pavlu V. and Yilmaz E.: A statistical method for system evaluation using incomplete judgments. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. (2006)
3. Aslam J. A. and Savell R.: On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. (2003)
4. Carterette B., Allan J. and Sitaraman R.: Minimal test collections for retrieval evaluation. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA (2006)
5. Nuray R. and Can F.: Automatic ranking of information retrieval systems using data fusion. Information Processing and Management: an International Journal, v.42 n.3, p.595-614, (2006)
6. Soboroff I., Nicholas C. and Cahan P.: Ranking retrieval systems without relevance judgments. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, p.66-73, New Orleans, Louisiana, United States (2001)
7. Spoerri A.: How the overlap between search results correlates with relevance. In: Proceedings of the 68th annual meeting of the American Society for Information Science and Technology (2005)

8. Spoerri A.: Using the structure of overlap between search results to rank retrieval systems without relevance judgments. Information Processing and Management: an International Journal, v.43 n.4, p.1059-1070, (2007)
9. Voorhees E. M.: Variations in relevance judgments and the measurement of retrieval effectiveness. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.315-323, Melbourne, Australia (1998)
10. Voorhees E. M. and Harman, D.: Overview of the eighth text retrieval conference (TREC-8). The 8th text retrieval conference (TREC-8), Gaithersburg, MD, USA (1999)
11. Wu S. and Crestani F.: Methods for ranking information retrieval systems without relevance judgments. In: Proceedings of the 2003 ACM symposium on applied computing. Melbourne, Florida (2003)
12. Zobel J.: How reliable are the results of large-scale information retrieval experiments? In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.307-314, Melbourne, Australia (1998)