Information Retrieval Evaluation with Partial Relevance Judgment

Shengli Wu and Sally McClean

School of Computing and Mathematics
University of Ulster, Northern Ireland, UK
{s.wu1, si.mcclean}@ulster.ac.uk

Abstract. Mean Average Precision has been widely used by researchers in information retrieval evaluation events such as TREC, and it is believed to be a good system measure because of its sensitivity and reliability. However, its drawbacks as regards partial relevance judgment has been largely ignored. In many cases, partial relevance judgment is probably the only reasonable solution due to the large document collections involved.

In this paper, we will address this issue through analysis and experiment. Our investigation shows that when only partial relevance judgment is available, mean average precision suffers from several drawbacks: inaccurate values, no explicit explanation, and being subject to the evaluation environment. Further, mean average precision is not superior to some other measures such as precision at a given document level for sensitivity and reliability, both of which are believed to be the major advantages of mean average precision. Our experiments also suggest that average precision over all documents would be a good measure for such a situation.

1 Introduction

Since the beginning of information retrieval research, the evaluation issue has been paid considerable attention because of its complexity. Recall (the fraction of all the relevant documents which are retrieved) and precision (the fraction of the retrieved documents which are relevant) are considered by many researchers as the two most important (but very different) aspects [6, 8, 12]. Using a single value measure for a comprehensive consideration of these two aspects is an attractive opinion [13]. Some such measures have been proposed: Borko's BK measure [12], Vickery's Q and V measures [12]. van Rijsbergen's E measure [12], the harmonic mean by Shaw, Burgin, and Howell [10], and cumulated gain by J*ä*velin and Kek*ä*l*ä*inen [5, 7], etc. However, most of them have not been used widely.

One exception to this is average precision over all relevant documents, which has been referred to as mean average precision recently. Mean average precision has been used in Text REtrieval Conferences[1] [11] since 1994 (TREC 3) and now it is widely used by many researchers to evaluate their systems, algorithms, etc.

[1] TREC is a major event for the evaluation of information retrieval. Since 1992, it has been held yearly by the National Institute of Standards and Technology of USA and USA Department of Defence.

D. Bell and J. Hong (Eds.): BNCOD 2006, LNCS 4042, pp. 86–93, 2006.

Some previous research [1,2,9,14,16] suggests that mean average precision is a good system measure for several reasons. First, it is a single value measure therefore convenient for use, especially for comparing the performances of several different information retrieval systems. Second, it is sensitive since its calculation uses the complete information of relevant documents: the total number of relevant documents in the whole document collection and the ranked positions of them in a resultant list. Third, it is reliable. The reason for this is the same as for the second point.

Compared with mean average precision, precision at a given document level is quite different and is believed to be a good user-oriented measure. First, very often users' major concern is how many relevant documents exist in the top k (say, 5 or 10) documents. Second, it is very convenient for evaluation and requires much less effort than mean average precision does. Third, its value is explicit and easy to understand, while a mean average precision value is abstract and cannot be explained explicitly.

The above conclusion on mean average precision should be true if all the relevance judgment information is available. However, this is not the case in some situations. For example, in TREC, a pooling strategy is used. For every information need statement (topic) the top 100 documents in all or some submitted runs are put in the pool. Only those documents in the pool are judged by human assessors and all the documents which are not in the pool are unjudged and assumed to be irrelevant. Therefore, many relevant documents may be missed using such a pooling strategy [18]. Results from a Web search service is another situation where complete relevance judgment is impossible. However, the harmful effect of the incompleteness of relevance judgment information on mean average precision has not been discussed.

In this paper we would like to investigate this issue through analysis and experimentation with TREC data. The analysis and experiments will reveal some drawbacks of mean average precision for incomplete relevance judgment besides the one already known – only obscure explanation available for any mean average precision value.

Furthermore, a new measure is introduced and investigated in this paper. It is average precision over all documents. It will be demonstrated in this paper that this measure has the advantages of both mean average precision and average precision at a given document level, but does not have some shortcomings of mean average precision.

2 Two Measures

Mean average precision has been used by TREC in TREC 3 and onwards [11]. Since then, mean average precision has been widely used by researchers to evaluate their information retrieval systems and algorithms. It uses the following formula:

$$map = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{r_i} \tag{1}$$

where n is the total number of relevant documents in the whole collection for that information need and r_i is the ranking position of the i-th relevant document in the list. For example, suppose there are 4 relevant documents for a topic, and these relevant documents are ranked in number 1, 4, 10, and 12 in a result, then this result's mean average precision is (1/1+2/4+3/10+4/12)/4=0.525.

Precision at a given document level is not very sensitive because it does not consider the positions of the relevant documents involved. For example, a relevant document appearing in rank 1 and in rank 100 has the same effect on precision at the 100 document level.

A new measure, average precision over all documents, is introduced in this paper. It can be a better choice than precision at a given document level since it concerns with the positions of relevant documents. It uses the following formula to calculate scores:

$$ap\_all(m) = \frac{1}{m} \sum_{i=1}^{m} \frac{r(i)}{i} \qquad (2)$$

Where $r(i)$ is the number of relevant documents in the top i documents and m is the total number of documents considered. Comparing Formula 1 and Formula 2, they bear some similarities.

If a document in rank j is relevant, then its contribution to the final score is $ap\_all(j,m) = \frac{1}{m}\sum_{i=j}^{m}\frac{1}{i}$. $H(m) = \sum_{i=1}^{m}\frac{1}{i}$ is a Harmonic number [4], which has some interesting characteristics. Let us consider $ap\_all'(j,m) = ap\_all(j,m) * m = \sum_{i=j}^{m}\frac{1}{i}$, which is a tail of a Harmonic number and we have $ap\_all'(j,m) = H(m) - H(j-1)$. Actually, the measure of discounted cumulated gain proposed by Jävelin and J. Kekäläinen [5] is a "general" measure of weighting schemas. Also they suggested a weighting schema: 1 for rank 1 and 2, 1/2 for rank 3, 1/3 for rank 4,..... Average precision over all documents can be regarded as a specialised form of discounted cumulated gain. In the remainder of this paper, we will focus on average precision over all documents and will not discuss other weighting schema variations.

3 Experiments

Experimental results using TREC data are reported in this section. We hope this can help us to obtain a better understanding about these measures. Compared with previous work [2, 3, 9, 14, 16], our experiments have different goals: we would like to find out how mean average precision perform when only incomplete relevance judgment information is available, and we also would like to investigate the new measure introduced in this paper – average precision over all documents.

3.1 Experimental Setting

9 groups of results (TREC 5, 6, 7, and 8: ad hoc track; TREC 9, 2001, and 2002: web track; TREC 2003 and 2004: robust track) submitted to TREC ad hoc, web

and robust track are used in the experiments. Three measures, mean average precision, average precision over all documents, and precision at 10 document level, are used in the experiment. In order to eliminating the effect of pooling, only the top 100 documents are used for the evaluation of all the involved results.

3.2 Error Rates Using Different Measures

First we carry out an experiment to investigate the stability and sensitivity of different measures. For a given measure, we evaluate all the results in a year group and obtain the average performance of them. Then for those pairs whose performance difference is above 5%, we check if this is true for all the topics. Suppose we have two results A and B such that A's average performance is better than B's average performance by over 5% in all l topics. Then we consider these l topics one by one. We find that A is better than B by over 5% for m queries, and B is better than A by over 5% for n queries ($l \geq m + n$). In this case the error rate is $n/(m + n)$.

The result of this experiment is shown in Table 1. On average, average precision over all document levels (*ap_all*) is the best, precision at 10 document level (p10) is in the second place, while mean average precision (*map*) is the worst. The differences between these measures are not big (p10-map: 2.69%, ap_all-map: 3.86%, and ap_all-p10: 1.13%).

On the other hand, when using mean average precision, more pairs are selected than when using the two other measures. This suggests that mean average precision is more sensitive than the two others. However, the difference is not large here either (map-p10: 3.68% and map-ap_all: 3.57%). Our experimental results suggest that these three measures are close in sensitivity and stability.

A similar experiment was carried out by Buckley and Voorhees [2]. They used all results submitted to the TREC 8 query track and tested the stability of several measures over different query formats. The experimental result reported here is consistent with that of Buckley and Voorhees's [2] though the experimental settings are different. In their experiment, they considered the top 1000 documents for every result and they found that precision at 1000 document level has lower error rates than mean average precision, while precision at 10 and 30 document levels have higher error rates than mean average precision. This suggests that precision at certain document level can be as good as mean average precision if the same number of documents are used.

3.3 Correlation Among Different Measures

Our second experiment aims to investigate how similar or different these measures are. Given a group of results, we use different measures to evaluate them and rank them based on their performances. Then we compare those rankings generated by using different measures. The experimental result is shown in Table 2. Both Spearman and Kendall's tau ranking coefficients are calculated. In table 2, all Kendall's tau ranking coefficient values are lower than the corresponding Spearman coefficient values, though the difference does not affect their relative rankings in most cases.

Table 1. Error rates of using different measures (numbers in parentheses are numbers of compared pairs)

Group	map	p10	ap_all
TREC 5	0.2731(1657)	0.2771(1694)	0.2710(1631)
TREC 6	0.2559(2262)	0.2650(2317)	0.2549(2257)
TREC 7	0.2451(4707)	0.2481(4837)	0.2429(4716)
TREC 8	0.2270(7096)	0.2304(7375)	0.2255(7119)
TREC 9	0.2351(5116)	0.2421(5114)	0.2315(5028)
TREC 2001	0.2839(4147)	0.2936(4261)	0.2798(4114)
TREC 2002	0.2641(2331)	0.2739(2374)	0.2595(2343)
TREC 2003	0.3006(2315)	0.3239(2478)	0.2916(2347)
TREC 2004	0.3223(4616)	0.3184(5053)	0.3241(4724)
Average	0.2675(3805)	0.2747(3945)	0.2645(3809)

Table 2. Correlation among rankings generated using different measures (S for Spearman coefficient and K for Kendall's tau coefficient)

Group	map vs. ap_all		map vs. p10		ap_all vs. p10	
	S	K	S	K	S	K
TREC 5	0.9683	0.8656	0.9628	0.8546	0.9822	0.9060
TREC 6	0.9551	0.8342	0.9482	0.8149	0.9773	0.8954
TREC 7	0.9754	0.8759	0.9523	0.8233	0.9797	0.8942
TREC 8	0.9710	0.8697	0.9466	0.8241	0.9807	0.8934
TREC 9	0.9689	0.8579	0.9526	0.8176	0.9851	0.9011
TREC 2001	0.9701	0.8621	0.9302	0.7934	0.9685	0.8565
TREC 2002	0.9243	0.7835	0.9538	0.8157	0.9036	0.7730
TREC 2003	0.9443	0.8069	0.8512	0.6830	0.8689	0.7362
TREC 2004	0.9800	0.8902	0.9460	0.8202	0.9588	0.8445
Ave.	0.9619	0.8496	0.9382	0.8052	0.9594	0.8556

The rankings generated using these three measures are strongly correlated with each other. On average the correlation is above 0.8 (Kendall's tau coefficient) or 0.9 (Spearman coefficient). In addition, the rankings generated using average precision over all documents are almost equally and very strongly correlated to the rankings generated using either of the two other measures, while the ranking correlation between precision at 10 document level and mean average precision is weaker.

3.4 Effect of Environment on Results Evaluation and Ranking

To evaluate information retrieval results using mean average precision demands much more efforts than using some other measures such as precision at the 10 or 100 document level, mainly because all relevant documents need to be identified. If complete relevance judgment is not available, then the performance of a result on mean average precision will depend on the relevant documents detected to a certain

Table 3. Correlation of rankings using full sets of results and rankings using partial sets of results (the numbers in parentheses indicate the performance difference of the same result in different environments)

Group	20%	40%	60%	80%
TREC 5	0.9606 (18.15%)	0.9724 (9.67%)	0.9832 (6.18%)	0.9867 (3.47%)
TREC 6	0.9582 (16.00%)	0.9793 (8.40%)	0.9905 (4.92%)	0.9941 (2.78%)
TREC 7	0.9768 (15.21%)	0.9870 (7.22%)	0.9930 (4.20%)	0.9963 (1.96%)
TREC 8	0.9690 (11.49%)	0.9833 (6.39%)	0.9922 (2.95%)	0.9968 (1.49%)
TREC 9	0.9714 (9.47%)	0.9934 (2.75%)	0.9944 (1.10%)	0.9970 (0.67%)
TREC 2001	0.9602 (15.40%)	0.9738 (7.54%)	0.9852 (3.86%)	0.9900 (2.02%)
TREC 2002	0.9604 (16.84%)	0.9810 (7.39%)	0.9856 (3.80%)	0.9879 (2.36%)
TREC 2003	0.9562 (16.39%)	0.9574 (12.10%)	0.9600 (10.10%)	0.9664 (8.95%)
TREC 2004	0.9740 (12.82%)	0.9797 (9.24%)	0.9862 (8.25%)	0.9849 (7.27%)
Average	0.9652 (14.64%)	0.9786 (8.73%)	0.9856 (5.04%)	0.9889 (3.44%)

degree. In TREC, only the documents in the pool are assessed and the pool comprises the top 100 documents from all or some of the submitted results. Therefore, a result's performance on mean average precision is affected by the other submitted results, and we refer to this phenomenon as the effect of environment.

We carry out an experiment to investigate this effect. For every year group, we evaluate and rank them as well by mean average precision. Then we randomly select a subset (20%, 40%, 60%, and 80%) of all the systems and assume these are all the results submitted, then we follow the TREC routine to generate a pool, and evaluate these systems by mean average precision and rank these results. We compare the ranking obtained from the subset of all the results and the one obtained from all the results to see if there is any ranking exchange for any two results appearing in both cases. Kendall's tau coefficient is calculated for them. Table 3 shows the experimental result. Each data point in Table 3 is the average of 10 runs.

In Table 3, the ranking correlation coefficient values are close to 1 all the time. this means that the relative rankings of a group of results do not change much when some new results are included. Though it can be regarded as a good news, it is not good enough. Since no ranking position exchanging at all is a norm with other measures such as precision at 10 or 100 document level and average precision over all documents.

On the other hand, considerable difference exists for the performance of the same result when the environment changes. When 20% of all results are considered, the difference is over 10% compared with the environment in which all the results are involved.

4 Conclusions

In this paper we have discussed three information retrieval evaluation measures, which are average precision over all relevant documents (mean average precision), precision at a given document level, and average precision over all documents, under the condition of incomplete relevance judgment.

Though it has been believed that average precision over all relevant documents is a good measure, our investigation shows that it suffers from several drawbacks when only partial relevance judgment is available. First, the correct mean average precision value can never be calculated. Hence complete relevance judgment is required for a correct calculation of average precision over all relevant documents. Second, when a pair of results take part in an information retrieval evaluation event such as TREC, their relative ranking positions may reverse if other results involved are different at each time. Though the possibility for such a contradiction is very small, there is no guarantee that it does not happen. Besides, a mean average precision value is difficult to explain, and to calculate average precision values demands great effort. These are two drawbacks of mean average precision even with complete relevance judgment.

Then what about these measures' stability and sensitivity? Our experiment suggests that mean average precision's stability and sensitivity is not superior to the two other measures: average precision over all documents and precision at a given document level, if we use the same (or similar) number of documents for the calculation of these measures. This observation is consistent with previous research [2, 9, 16]. Buckley and Voorhees in [2] find that precision at 1000 documents is more stable than mean average precision, and mean average precision is more stable than precision at 10 documents. The last point is also echoed in [9, 16].

We argue that mean average precision is not a very good measure when relevance judgment is severely incomplete. Although in theory mean average precision has some advantages, its use within TREC evaluation methodology has led to the anomalies discussed above. The difficulties are inevitable in modern IR contexts such as retrieval over the Web. Meanwhile, precision at a given document level and especially average precision over all documents are good measures in such situations. Average precision over all documents has been introduced in this paper and it is more reasonable than precision at a given document level since it distinguishes relevant documents' position. In addition, the similarity between mean average precision and average precision over all documents is more than that between mean average precision and precision at a given document level. Therefore, we consider that average precision over all documents would be a good measure for information retrieval evaluation events such as TREC as well as for researchers to evaluate information retrieval systems and algorithms when the document collection is too big for a complete relevance judgment.

References

1. J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analysing retrieval measures. In *Proceedings of ACM SIGIR'2005*, pages 27–34, Salvador, Brazil, August 2005.
2. C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of ACM SIGIR'2000*, pages 33–40, Athens, Greece, July 2000.
3. C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of ACM SIGIR'2004*, pages 25–32, Sheffield, United Kingdom, July 2004.

4. R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete mathematics.* Addison-wesley publishing company, 1989.
5. K. Jävelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):442–446, October 2002.
6. Y. Kagolovsky and J. R. Moehr. Current status of the evaluation of information retrieval. *Journal of Medical Systems*, 27(5):409–424, October 2003.
7. J. Kekäläinen. Binary and graded relevance in IR evaluations – comparison of the effects on ranking of IR systems. *Information Processing & Management*, 41(5):1019–1033, September 2005.
8. S. E. Robertson and M. M. Hancock-Beaulieu. On the evaluation of IR systems. *Information Processing & Management*, 28(4):457–466, July-August 1992.
9. M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of ACM SIGIR'2005*, pages 162–169, Salvador, Brazil, August 2005.
10. W. M. Shaw, R. Burgin, and P. Howell. Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing & Management*, 33(1):1–14, January 1997.
11. TREC. http://trec.nist.gov/.
12. C. J. van Rijsbergen. *Information Retrieval.* Butterworths, 1979.
13. V. G. Voiskunskii. Evaluation of search results: A new approach. *Journal of the American Society for Information Science*, 48(2):133–142, February 1997.
14. E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of ACM SIGIR'1998*, pages 315–323, Melbourne, Australia, August 1998.
15. E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, September 2000.
16. E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of ACM SIGIR'2002*, pages 316–323, Tampere, Finland, August 2002.
17. S. Wu and S. McClean. Modelling rank-probability of relevance relationship in resultant document list for data fusion, submitted for publication.
18. J. Zobel. How reliable are the results of large-scale information retrieval experiments. In *Proceedings of ACM SIGIR'1998*, pages 307–314, Melbourne, Australia, August 1998.