Journal of Information Assurance and Security 2 (2007) 257-265

Several methods of ranking retrieval systems with partial relevance judgment

Shengli Wu [1] and Sally McClean[2]

[1] School of Computing and Mathematics, University of Ulster at Jordanstown
Shore Road, Newtownabbey, Northern Ireland, UK
s.wu1@ulster.ac.uk

[2]School of Computing and Information Engineering, University of Ulster at Coleraine,
Cromore Road, Coleraine, Northern Ireland, UK
si.mcclean@ulster.ac.uk

***Abstract*:** Some measures such as average precision over all relevant documents and recall level precision are considered as good system-oriented measures, because they concern both precision and recall that are two important aspects for effectiveness evaluation of information retrieval systems. However, such good system-oriented measures suffer from some shortcomings when partial relevance judgment is used. In this paper, we discuss how to rank retrieval systems based on partial relevance judgment, which is common in major retrieval evaluation events such as TREC conferences and NTCIR workshops. Four system-oriented measures, which are average precision over all relevant documents, recall level precision, normalized discount cumulative gain, and normalized average precision over all documents, are discussed. Our investigation shows that with partial relevance judgment, the evaluated results can be far from accurate and incomparable across queries. In such a situation, averaging values over a set of queries may not be the most reliable approach to rank a group of retrieval systems. Some alternatives such as Borda count, Condorcet voting, and the Zero-one normalization method, are investigated. Experimental results are also presented for the evaluation of these methods.

***Keywords*:** information retrieval systems, evaluation, system ranking, partial relevance judgment

1. Introduction

In information retrieval, to compare the effectiveness of a group of information retrieval systems, a test collection, which includes a set of documents, a set of query topics, and a set of relevance judgments indicating which documents are relevant to which topics, is required. Among them, "relevance" is an equivocal concept [1, 11, 12] and relevance judgment is a task which demands huge human effort. In some situations such as the Web search, a complete relevance judgment is not possible. In the Text REtrieval Conference (TREC), only partial relevance judgment is conducted due to the large number of documents in the whole test collection.

In the evaluation of information retrieval systems, precision (number of relevant documents retrieved/total number of documents retrieved) and recall (number of relevant documents retrieved/total number of relevant documents in the whole collection) are regarded as the two most important aspects and therefore both of them should be considered at the same time. On the other hand, a single value metric is required to rank a group of information retrieval systems according to their effectiveness. Average precision over all relevant documents (AP), recall level precision (RP), normalized discount cumulative gain (NDCG), and normalized average precision over all documents (NAPD) can be regarded as candidates of good system-oriented measures. Among them, AP and RP have been used in TREC for quite a few years and now they are widely used by researchers to evaluate their systems and algorithms; NDCG was proposed by Järvelin and Kekäläinen [5, 6]; and NAPD was proposed by Wu and McClean [19].

Without complete relevance judgment, only a subset of all relevant documents can be identified. This will affect recall and system-oriented measures whose precise values require complete relevant judgments. In the TREC conferences, a pooling method [13] is used. Since only the top 100 documents in all or a subset of the submitted runs are checked, a relatively large percentage of relevant documents may not be detected [21]. To find out the effect of these missing relevant documents on retrieval evaluation using some system-oriented measures is an issue worth investigation.

In this paper we would like to investigate how to fairly rank a group of retrieval systems using system-oriented measures based on partial relevance judgment. We find that partial relevance judgment does affect the values of system-oriented measures significantly when using the TREC's pooling method. The more incomplete the relevance judgment is, the bigger values we obtain for these measures. Moreover, different percentages of relevant documents may be identified by the pooling method for different topics. This means that the values calculated with the pooling method can be exaggerated at different rates for different topics. In such a situation, averaging these values over a set of queries might not be the best solution for ranking a group of systems. Some other reasonable options are discussed in this paper. Experiments are also conducted to evaluate these methods' reliability.

Received November 30, 2007.

The rest of this paper is organized as follows: in Section 2 we review some related work. Section 3 discusses the four measures used in this paper. Experimental results are presented in Section 4 to demonstrate that ranking a group of retrieval systems by averaging those values over a set of queries in the condition of partial relevance judgment may be questionable. In Section 5 we propose some alternative methods, namely, Borda count, Condorcet voting, and the Zero-one normalization method, for the ranking of a group of retrieval systems besides the method of averaging those values. Section 6 presents experimental results on the evaluation of these methods. Section 7 concludes the paper.

2. Related work

Zobel [21] investigated the reliability of some measures such as precision and recall (but none of the measures discussed in this paper were included) in TREC where partial relevance judgment was taken. He found that the results based on the relevance judgments formed from a limited pool were reliable--if the pool was sufficiently deep. However, he identified some limitations of the pooling method. The practice of using the top 1000 documents to measure systems when only the top 100 had contributed to the pool allows greater discrimination between systems, but introduces uncertainty. He also estimated that at best 50%-70% of the relevant documents could be found by the pooling method in TREC.

Voorhees [15, 16] investigated the effect of varying relevance judgment to the evaluation of information retrieval systems since very often different human assessors might have different opinions about documents' relevancy to an information need. Two groups of results submitted to TREC 4 and TREC 6 were used for the experiments. Her experiments suggested that different relevance judgment profiles did affect evaluation using AP, but its effect on AP-based system ranking was slight.

Buckley and Voorhees [3] conducted an experiment to investigate the stability of different measures including AP and RP when using different query formats. Results submitted to the TREC 8 query track were used. In their experiment, recall at 1000 document level had the least error rate, which was followed by precision at 1000 document level, RP, and AP, while precision at 1, 10, and 30 document levels had the biggest error rates.

Voorhees and Buckley [16] conducted another experiment to investigate the effect of topic set size on retrieval result. 8 groups of results submitted to TREC ad hoc (TREC 3-8) and Web tracks (TREC 9 and TREC 2001) were used. They used all 50 queries and various subsets of them to check if they agreed as to which of the results was better. AP and P10 (precision at 10 document level) were used as effectiveness measures. They found that using precision at 10 document level incurred higher error rate than using AP in their experiment.

Buckley and Voorhees [4] introduced a measure *bpref* for partial relevance judgment. *bpref* is defined as

$$bpref = \frac{1}{R}\sum_{r} 1 - \frac{|n\_ranked\_higher\_than\_r|}{R}$$

Here R is the total number of relevant documents for the topic. The summation is over all such relevant documents. And $|n\_ranked\_higher\_than\_r|$ is the number of judged non-relevant documents whose ranks are higher than r. One characteristic of this measure is: it only concerns how many judged non-relevant documents there are before judged relevant documents, but it does not distinguish judged relevant documents from un-judged documents. In other words, it implies that all un-judged documents are relevant. Having noticed that, Sakai [10] proposed some alternatives to *bpref* for the partial relevance judgment environment.

Sanderson and Zobel [9] reran the experiment that Buckley and Voorhees did [16] with two more groups of results and had similar observations. However, they argued that P10 was as good as AP if considering both error rate for relative difference and human judgmental effort.

Järvelin and Kekäläinen [5] introduced cumulated gain-based evaluation measures. Among them, normalized discount cumulated gain (NDCG) concerns both precision and recall, which can be used as an alternative for AP. Using cumulated gain-based evaluation measures, Kekäläinen [6] compared the effect of binary and graded relevance judgment on the rankings of information retrieval systems. She found that these measures correlated strongly under binary relevance judgment, but the correlation became less strong when emphasizing highly relevant documents in graded relevance judgment.

In this paper, we focus on how to fairly rank a group of information retrieval systems based on system-oriented measures in the condition of partial relevance judgment, rather than define some new measures as in [4, 10].

3. Four measures

In this section we discuss the four measures used in this paper. AP and RP have been used many times in TREC [17]. Both of them are defined with binary relevance judgment and now they are used widely by researchers to evaluate their information retrieval systems and algorithms (e.g., in [2, 7, 20]). AP uses the equation, $ap = \frac{1}{R}\sum_{i=1}^{R}\frac{i}{p_i}$, to calculate scores. Here R is the total number of relevant documents in the whole collection for the given query and p_i is the ranking position of the i-th relevant documents in the resultant list. RP is defined as the percentage of relevant documents in the top R documents where R is the total number of relevant documents for the given query.

NAPD is introduced in [19]. First let us discuss a related measure - average precision over all documents (APD). APD uses the equation, $apd = \frac{1}{n}\sum_{i=1}^{n}\frac{r(i)}{i}$, to calculate scores. Here n is the total number of documents in the

resultant document list, and *r(i)* is the number of relevant documents in the first *i* documents of the resultant list. Suppose *apd_best* is the best possible APD score for the given query, then NAPD can be defined as *NAPD=apd/apd_best*.

NDCG is introduced in [5]. Each ranking position in a resultant document list is assigned a given weight. The top ranked documents are assigned the highest weights since they are the most convenient ones for users to read. A logarithmic function-based weighting schema was proposed in [5], which needs to take a particular whole number *b* (*b*=2 is used in this paper). The first *b* documents are assigned a weight of 1; then for any document ranked *k* which is greater than *b*, its weight is $w(k)=\log b/\log k$. Considering a resultant document list up to *n* documents, its discount cumulated gain (DCG) is $\sum_{i=1}^{n} w(i)*r(i)$. *r(i)* is defined as: if the *i*-th document is relevant, then *r*(*i*)=1; if the *i*-th document is irrelevant, then *r*(*i*)=0. DCG can be normalized using a normalization coefficient *dcg_best*, which is the DCG value of the best resultant lists. Therefore, we have: $ndcg = \frac{1}{dcg\_best}\sum_{i=1}^{n} w(i)*g(i)$. In summary, all these four measures are normalized since their values are always in the range of 0 and 1 inclusive.

Suppose for a given query, there are 4 relevant documents in the whole collection. A resultant document list from a retrieval system comprises 10 documents:

$D_1^{*}\ D_2\ D_3\ D_4^{*}\ D_5^{*}\ D_6\ D_7\ D_8\ D_9\ D_{10}^{*}$

The documents with a * are relevant documents. Then we have:

$$ap = \frac{1}{4}\sum_{i=1}^{4}\frac{i}{p_i} = 0.25*(1+\frac{2}{4}+\frac{3}{5}+\frac{4}{10}) = 0.625$$

$$rp = \frac{2}{4} = 0.5$$

$$apd = \frac{1}{10}\sum_{i=1}^{10}\frac{r(i)}{i}$$

$$= 0.1*(1+\frac{1}{2}+\frac{1}{3}+\frac{2}{4}+\frac{3}{5}+\frac{3}{6}+\frac{3}{7}+\frac{3}{8}+\frac{3}{9}+\frac{4}{10}) = 0.4971$$

$$apd\_best = \frac{1}{10}\sum_{i=1}^{10}\frac{r(i)}{i}$$

$$= 0.1*(1+\frac{2}{2}+\frac{3}{3}+\frac{4}{4}+\frac{4}{5}+\frac{4}{6}+\frac{4}{7}+\frac{4}{8}+\frac{4}{9}+\frac{4}{10}) = 0.7383$$

$$napd = \frac{apd}{apd\_best} = \frac{0.4791}{0.7383} = 0.6489$$

$$dcg = \sum_{i=1}^{4} w(i)*r(i) = 1+\frac{\log 2}{\log 4}+\frac{\log 2}{\log 5}+\frac{\log 2}{\log 10} = 2.2316$$

$$dcg\_best = \sum_{i=1}^{4} w(i)*r(i) = 1+\frac{\log 2}{\log 2}+\frac{\log 2}{\log 3}+\frac{\log 2}{\log 4} = 3.1309$$

$$ndcg = \frac{dcg}{dcg\_best} = \frac{2.2316}{3.1309} = 0.7128$$

4. Relationship between pool depths and measure values

In this section we investigate the effect of partial relevance judgment on these system-oriented measures. We carry out an empirical study with TREC data. 9 groups of runs submitted to TREC (TREC 5-8: ad hoc track; TREC 9, 2001, and 2002: Web track; TREC 2003 and 2004: robust track) were used in the experiment. Their information is summarized in Table 1. Considering that the pooling method in TREC is a reasonable method for partial relevance judgment, we conduct an experiment to compare the values of these measures by using pools of different depths. For every year, a pool of 100 documents in depth was used in TREC to generate its *qrels* (relevance judgment file). Shallower pools of 10, 20,…, 90 documents in depth were used in this experiment to generate more *qrels*. For a resultant list and a measure, we calculate its value of the measure c_{100} using the 100 document *qrels*, then calculate its value of the measure c_i using the *i* document qerls (*i* = 10, 20, …., 90). Their absolute difference can be calculated using $asb\_diff=|c_i-c_{100}|/c_{100}$ and their relative difference value can be calculated using $rel\_diff=(c_i-c_{100})/c_{100}$.

Table 1. Information about 9 groups of submitted results in TREC

Group	Track	Number of results	Number of topics
TREC 5	ad hoc	61	50
TREC 6	ad hoc	71*	50
TREC 7	ad hoc	103	50
TREC 8	ad hoc	129	50
TREC 9	Web	105	50
TREC 2001	Web	97	50
TREC 2002	web	71	50
TREC 2003	robust	78	100
TREC 2004	robust	101	249**

Note: *Three submitted results to TREC 6 were removed since they include very few documents. **One topic in TREC 2004 was dropped since it did not include any relevant document.

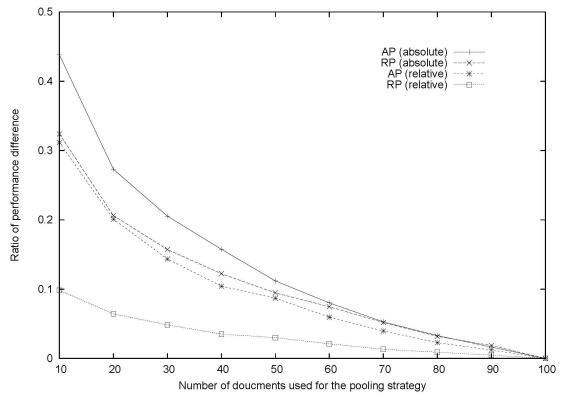

Figure 1. Value differences of two measures AP and RP when using pools of different depth (the pool of 100 documents in depth is served as baseline)

Figure 1 shows the absolute and relative differences of AP and RP values when different *qrels* are used. Every data point in Figure 1 is the average of all submitted runs in all year groups. One general tendency for the two measures is: the shallower the pool is, the bigger the difference is. However, AP is the worst considering the difference rate. When using a pool of 10 documents in depth, the absolute difference rate is as big as 44% and the relative difference rate is 31% for AP. In the same condition, they are 32% and 10% for RP. In all the cases, relative difference is smaller than corresponding absolute difference. In addition, similar conclusions are observed for NDCG and NAPD (Figure 2). The difference rates for them are close to that for RP, but are lower than that for AP.

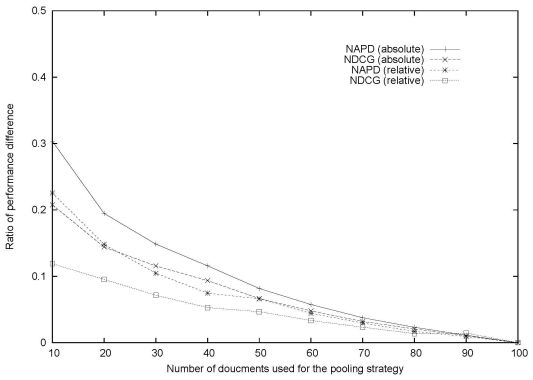

Figure 2. Value differences of two measures NAPD and NDCG when using pools of different depth (the pool of 100 documents in depth is serves as baseline)

Next we explore the relationship of pool depth and relevant documents identified. The result is shown in Figure 3. The curve increases quickly at the beginning and then slow down, but keeps increasing when the pool depth reaches 100. From the curve's tendency, it seems that the increase will continue for some time. Using some curve estimation techniques as used by Zobel we find that very likely 20%-40% of the relevant documents can be found if the pool expands from the point of 100 documents in pool depth. From these observations, we can derive that the values of all these four measures are over estimated using a pool of 100 documents compared with their actual values when complete relevance judgment is available.

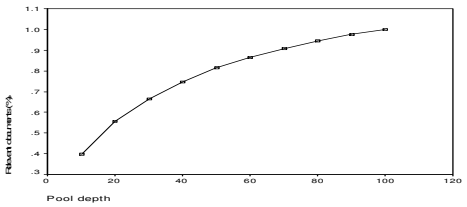

Figure 3. Percentage of relevant documents identified when the pool varies in depth (the pool of 100 documents served as baseline, 100%)

Furthermore, we investigate the impact of the number of identified relevant documents on these measures. For all 699 topics (queries) in 9 year groups, we divided them into 11 groups according to the number of relevant documents identified for them. Group 1 (G_1) includes those topics with fewer than 10 relevant documents, group 2 (G_2) includes those topics with between 10 and 19 relevant documents, …, group 11 (G_{11}) includes those topics with 100 or more relevant documents. The number of topics in each group is as follows:

G1	*G2*	*G3*	*G4*	*G5*	*G6*
47	16	79	76	49	33
G7	*G8*	*G9*	*G10*	*G11*	Total
39	27	25	17	165	699

For all these topic groups G_1 ~ G_{11}, we calculated the value differences of the same measure using pools of different depths. Figures 4 - 7 show the experimental result for AP, RP, NDCG, and NAPD, respectively. One common tendency for these four measures is: the fewer the relevant documents are identified, the less difference the values of the same measure have with pools of different depths. For example, the curves of G_1 are always below all other curves, while the curves of G_{10} and G_{11} are above all other curves. Comparing all these curves of different measures, we can observe that bigger differences occur for the measure of AP. For groups G_{10} and G_{11}, the value differences of AP are 0.93 and 0.84 between the pool of 10 documents and the pool of 100 documents, while the figures for RP are 0.48 and 0.52, respectively. From this experiment, we find that the error rate of the estimated values for any of the four measures depends on the percentage of relevant documents identified for that topic. The bigger percentage of relevant documents identified for a topic, the more accurate the estimated values for that topic. However, the numbers of relevant documents vary considerably from one topic to another. In TREC, some topics are much harder than the others and it is more difficult for information retrieval systems to catch relevant documents. Another reason is that some topics have more relevant documents than some other topics in the document collection. The numbers of relevant documents may differ greatly from one topic to another: from 1 or 2 to several hundreds. All these have considerable impact on the percentage of relevant documents identified for a given

topic by the TREC pooling method. Therefore, AP, RP, NDCG, and NAPD values obtained with a pool of certain depth are not comparable across topics.

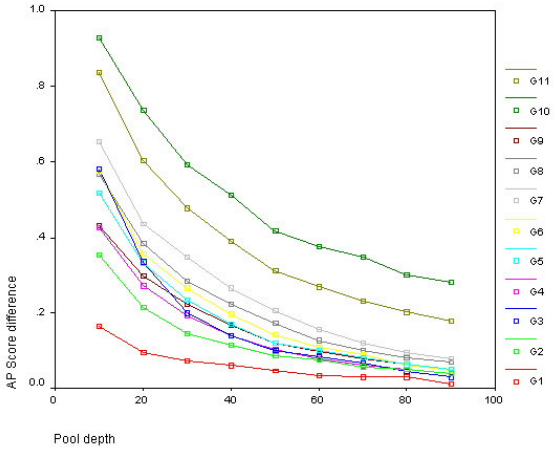

Figure 4. Difference in AP values using pools of different depths

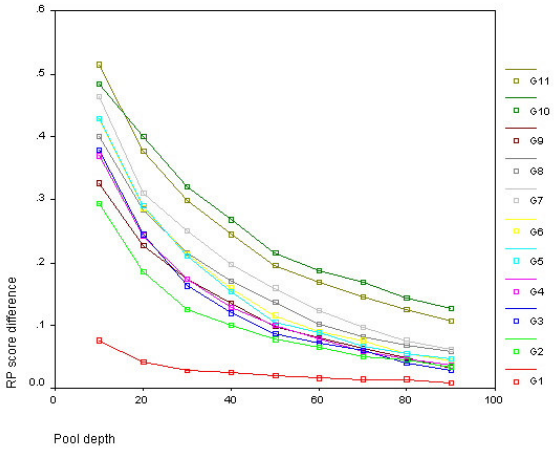

Figure 5. Difference in RP values using pools of different depths

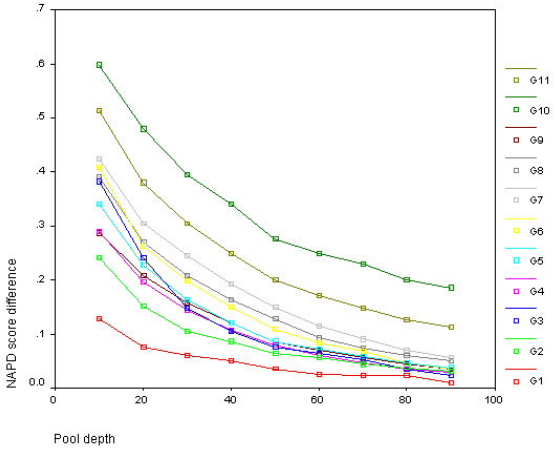

Figure 6. Difference in NAPD values using pools of different depths

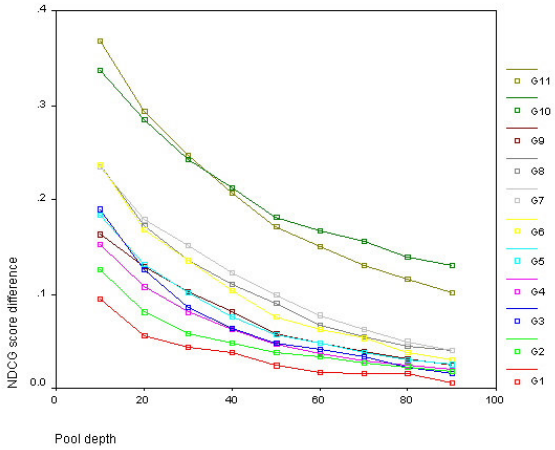

Figure 7. Difference in NDCG values using pools of different depths

Let us see an example to explain this further. Suppose that *A* and *B* are two systems under evaluation among a group of other systems. For simplicity, we only consider 2 queries. However, the same conclusion can be drawn if more queries are used to test their effectiveness. The results are as follows:

System (query)	Observed AP	Rate of exaggeration	Real AP
A (Q1)	0.32	80%	0.32/(1+0.8) =0.1778
B (Q1)	0.25	80%	0.25/(1+0.8) =0.1389
A (Q2)	0.45	20%	0.45/(1+0.2) =0.3750
B (Q2)	0.50	20%	0.5/(1+0.2) =0.4167

According to the observed AP values, we may conclude that *A* is better than *B*, because *A*'s AP over two queries (0.32+0.45)/2=0.385 is greater than *B*'s AP over two queries (0.25+0.50)/2=0.375. However, because Query 1's AP is overestimated by 80% and Query 2's AP is overestimated by 20%, a modification is needed for these AP values. After that, we find that System *A* ((0.1778+0.3750)/2=0.2764) is worse than System *B* ((0.1389+0.4167)/2=0.2778). This example demonstrates that averaging the values may not be the best solution for ranking a group of retrieval systems over a group of queries in such a condition. In Section 5, we will discuss some alternatives for such a task.

5. Other options than averaging all the values for ranking retrieval systems

Suppose for a certain collection of documents, we have a group of systems $(r_1,r_2,\ldots r_n)$ and a group of queries $(q_1,q_2,\ldots,q_m)$, and every system returns a ranked list of documents for every query. Now the task is to rank these systems based on their performances (e.g., using any one

of the four system-oriented measures) over these queries. If complete relevance judgment is applied, then averaging these values over all the queries is no doubt the best solution. Under partial relevance judgment, the estimated values are far from accurate and are not comparable across queries, as we have demonstrated in Section 4. Considering in a single query, if System *A* is better than System *B* with partial relevance judgment, then the same conclusion is very likely to be true with complete relevance judgment, though the difference in quantity may not be accurate. In such a situation, we may conclude that these systems are involved in a number of competition events, each of which is via a query. Then the task becomes how to rank these systems according to all these *m* competition events. Some voting procedures such as Borda count [18] and Condorcet voting [8] in political science can be used here.

The Borda count works as follow. For a fixed set of candidates (*n*) and voters (*m*), each voter ranks these candidates in order of preference. For each voter, the top-ranked candidate is given *n* points, the second-ranked candidates is given *n*-1 points, and so on. The candidates are ranked in order of total points from all voters, and the candidate with the most points wins the selection. Condorcet voting is used for majority voting. It considers all possible head-to-head ranking competitions among all possible candidate pairs. Then all the candidates can be ranked according to the number of competitions they have won. Both Borda count and Condorcet voting can be used here for the evaluation purpose if we regard information retrieval systems as candidates and retrieved results for every query as voters. These voting algorithms are useful when the rankings generated from all queries are reliable but the score information is not reliable or not available at all.

Both Borda count and Condorcet voting only consider the ranks of all involved systems, but not the score values. Another option is to linearly normalize the values of a set of systems in every query into the range of [0,1], which will be referred to as the Zero-one normalization method. Using this method, for every query, the top-ranked system is normalized to 1, the bottom-ranked system is normalized to 0, and all other systems are linearly normalized to a value between 0 and 1 accordingly. Thus every query is in an equal position to make contributions for the final ranking. Then all systems can be ranked according to their total scores.

6. Evaluation of the four ranking methods

In this section we present some experimental results on the evaluation of these four methods. As in Section 4, 9 groups of submitted runs to TREC were used. For all the submissions in one year group, we calculated their effectiveness for every query with different measures. Then different ranking methods, Borda count, Condorcet voting, the Zero-one normalization method, and the averaging method, were used to rank them. For these rankings obtained using different methods, we calculated Kendall's tau coefficient for each pair of rankings obtained using the same measure but different ranking method. Table 2 shows the results, each of which is for one of the four measures.

From Table 2, we can observe that Kendall's tau coefficients in all cases are quite big. For any pair in any year group, the average is always bigger than 0.8. Considering all single cases, the coefficients are less than 0.7 only occasionally. We also observe that for all the measures, the rankings from the averaging method and that from the Zero-one normalization method always have the strongest correlation. This demonstrates that the averaging method and the Zero-one normalization method are more similar with each other than any other pairs. In addition, the rankings from Borda count are strongly correlated with the rankings from either the averaging method or the Zero-one normalization method as well. On the other hand, the correlations between the rankings from Condorcet voting and any others are always the weakest. This demonstrates that Condorcet voting is quite different from the three other methods.

Table 2. Kendall's tau coefficients of rankings generated by different methods using different measures (A: averaging, B: Borda, C: Condorcet, S: Zero-one)

Measure	A-B	A-C	A-Z	B-C	B-Z	C-Z
AP	0.8798	0.8143	0.9337	0.8361	0.9173	0.8308
RP	0.9072	0.8276	0.9384	0.8480	0.9379	0.8435
NAPD	0.9316	0.8416	0.9703	0.8472	0.9416	0.8445
NDCG	0.9327	0.8503	0.9692	0.8567	0.9400	0.8556

Table 3. Kendall's tau coefficients for AP (figures in parentheses indicate the significance level of difference compared with the averaging method)

	Averaging	Borda	Condorcet	Zero-one
1/5~all	0.7624	0.7855(.000)	0.7033(.000)	0.7765(.000)
2/5~all	0.8476	0.8658(.000)	0.7771(.000)	0.8597(.000)
3/5~all	0.8961	0.9115(.000)	0.8281(.000)	0.9071(.000)

4/5~all	0.9378	0.9454(.000)	0.8622(.000)	0.9438(.000)
Average	0.8610	0.8771[+1.87%]	0.7927[-7.93%]	0.8718[+1.25%]

Table 4. Kendall's tau coefficients for RP (figures in parentheses indicate the significance level of difference compared with the averaging method)

	Averaging	**Borda**	**Condorcet**	**Zero-one**
1/5~all	0.7332	0.7418(.000)	0.6501(.000)	0.7367(.000)
2/5~all	0.8308	0.8401(.000)	0.7534(.000)	0.8387(.000)
3/5~all	0.8860	0.8943(.000)	0.8036(.000)	0.8912(.000)
4/5~all	0.9283	0.9329(.001)	0.8484(.000)	0.9311(.011)
Average	0.8446	0.8523[0.91%]	0.7639[-9.55%]	0.8494[0.57%]

Table 5. Kendall's tau coefficients for NAPD (figures in parentheses indicate the significance level of difference compared with the averaging method)

	Averaging	**Borda**	**Condorcet**	**Zero-one**
1/5~all	0.7981	0.8031(.003)	0.7312(.000)	0.8036(.001)
2/5~all	0.8716	0.8761(.003)	0.7974(.000)	0.8758(.000)
3/5~all	0.9138	0.9193(.001)	0.8414(.000)	0.9187(.001)
4/5~all	0.9472	0.9504(.003)	0.8742(.000)	0.9507(.002)
Average	0.8816	0.8872[+0.64%]	0.8111[-8.00%]	0.8872[+0.64%]

Table 6. Kendall's tau coefficients for NDCG (figures in parentheses indicate the significance level of difference compared with the averaging method)

	Averaging	**Borda**	**Condorcet**	**Zero-one**
1/5~all	0.7910	0.7980(.004)	0.7315(.000)	0.7962(.002)
2/5~all	0.8670	0.8751(.000)	0.8020(.000)	0.8722(.000)
3/5~all	0.9125	0.9177(.004)	0.8462(.000)	0.9165(.003)
4/5~all	0.9458	0.9504(.001)	0.8824(.000)	0.9494(.002)
Average	0.8791	0.8853[+0.71%]	0.8155[-7.23%]	0.8836[+0.51%]

Table 7. Kendall's tau coefficients for all the four measures when comparing the two rankings, one of which is generated with a pool of 100 documents, the other is generated with a shallow pool of 10-90 documents

	Averaging	**Borda**	**Condorcet**	**Zero-one**
AP	0.6607	0.6800	0.4855	0.6771
RP	0.6309	0.6568	0.4851	0.6550
NAPD	0.7095	0.7167	0.5267	0.7134
NDCG	0.6981	0.7107	0.5013	0.7077

Another thing we can do is to compare those Kendall's tau coefficient values using the same ranking method but different measures. Using NDCG, all Kendall's tau coefficients are the biggest (0.9006 on average). NDCG is followed by NAPD (0.8961) and RP (0.8687), while AP is at the bottom (0.8687). This indirectly suggests that NDCG is the most reliable measure, which is followed by NAPD and RP, while AP is the least reliable measure.

Next we investigate the issue of system ranking using different number of queries. For the same group of systems, we rank them using all the queries and using a subset of all the queries (1/5, 2/5, 3/5, and 4/5 of all the queries), then we compare these two rankings by calculating their Kendall's tau coefficient. Tables 3-6 present the experimental results. In all the cases, a random process is used to select a subset of queries from all available queries. Every data point in these tables is the average of 20 pairs of rankings.

From Tables 3-6, we can see that on average Borda count and the Zero-one method are the most reliable methods, the averaging method is in the middle, and Condorcet voting is the least reliable method. The difference between Condorcet voting and the others is bigger, while the three others are much closer with each other in performance. Although the differences between the averaging method and Borda, and between the averaging method and Zero-one, are small, the differences are always significant for all four measures. Condorcet is worse than all three others at a significance level of .000. In some cases, the differences between Borda count and the Zero-one method are not significant.

Finally we conducted an experiment to compare the rankings using different pools. One ranking was generated with the pool of 100 documents, and the other ranking was

generated with a shallower pool of less than 100 documents. In the shallower pool, each query might be assigned a different pool depth, which was decided by a random process to choose a number from 10, 20, …, 80, and 90. The results are shown in Table 7, which is the average of 9 year groups, and 20 runs were performed for each year group. Again, we can observe that Condorcet is the worst, Borda count and the Zero-one method is slightly better than the averaging method.

7. Conclusions

Since the Web and digital libraries have more and more documents on these days, there is a need to test and evaluate information retrieval systems with larger and larger collections. In such a situation, how to make the human judgment effort reasonably low becomes a major issue. Partial relevance judgment is the solution to this in information retrieval evaluation events TREC and NTCIR. However, some further questions come up:

- What is the effect of partial relevance judgment on the evaluation process?
- Are there any partial relevance judgment methods other than the pool strategy can be applied?
- Which measures should we use for such a process?
- What can be done to make the evaluation process more reliable in the condition of partial relevance judgment?

In order to answer some of these questions, some previous research [4, 10] tried to define some new measures which are suitable for partial relevance judgment. Besides the pooling strategy, some other partial relevance judgment methods were also investigated. This paper has taken a different approach. We have investigated the effect of partial relevance judgment (especially the pooling method) on those extensively used system-oriented measures such as AP and RP. Based on that, we have further investigated how to fairly rank a group of retrieval systems based on those system-oriented measures. We have stuck to the pooling method since it has been successfully used in TREC and NTCIR and other information retrieval evaluation events for many years.

As we have seen, in such a situation the averaging method may be questionable, since the values of system-oriented measures obtained from different queries are not quite comparable cross multiple queries. Several alternative methods including Borda count, Condorcet voting, and the Zero-one normalization methods are investigated. Our experimental results suggest that Borda count and the Zero-one normalization method are slightly better than the averaging method, while Condorcet is the worst in these four methods.

Our investigation also demonstrates that with partial relevance judgment, the evaluated results can be significantly different from the results with complete relevance judgment: from their values on a system-oriented measure to the rankings of a group of information retrieval systems based on such values. Therefore, when conducting an evaluation with partial relevance judgment, we need to be careful about the results.

References

[1] C.L. Barry. “User-defined relevance criteria: an exploratory study”, *Journal of the American Society for Information Science*, 45(3), pp. 149-159, 1994.

[2] D. Bodoff, S. Robertson. “A new united probabilistic model”, *Journal of the American Society for Information Science and Technology*, 55(6), pp. 471-487, 2004.

[3] C. Buckley, E.M. Voorhees “Evaluating evaluation measure stability”. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 33-40, 1999.

[4] C. Buckley, E.M. Voorhees “Retrieval evaluation with incomplete information”. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 25-32, 2004.

[5] K. Järvelin, J. Kekäläinen. “Cumulated gain-based evaluation of IR techniques”, *ACM Transactions on Information Systems,* 20(4), pp 442-446, 2002.

[6] J. Kekäläinen “Binary and graded relevance in IR evaluations - comparison of the efforts on ranking of IR systems”, *Information Processing & Management*, 41(5), pp 1019-1033, 2005.

[7] C. Lee, F.G. Lee. “Probabilistic information retrieval model for a dependency structured indexing system”, *Information Processing & Management*, 41(2), pp 161-175, 2005.

[8] M. Montague, J. A. Aslam. “Condorcet fusion for improved retrieval”. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pp 538-548, 2002.

[9] M. Sanderson, J. Zobel. “Information retrieval system evaluation: Effort, sensitivity, and reliability”. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 162-169, 2005.

[10] R. Sakai. “Alternatives to Bpref”. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 71-78, 2007.

[11] T. Saracevic. “Relevance: A review of and a framework for thinking on the notion in information science”, *Journal of the American Society for Information Science*, 26(6), pp 321-343, 1975.

[12] L. Schamber, M.B. Eisenberg, M.S. Nilan. “A re-examination of relevance: toward a dynamic, situational definition”, *Information Processing & Management*, 26(6), pp 755-776, 1990.

[13] K. Sparck Jones, C. van Rijisbergen. “Report on the need for and provision of an ‘ideal’ information retrieval test collection”. *Technical report, British library research and development report 5266*, Computer laboratory, University of Cambridge, Cambridge, UK, 1975.

[14] E.M. Voorhees. “Variations in relevance judgments and the measurement of retrieval effectiveness”. In *Proceedings of the Annual ACM Conference on*

Research and Development in Information Retrieval (SIGIR), pp 315-323, 1998.

[15] E.M. Voorhees. "Variations in relevance judgments and the measurement of retrieval effectiveness", *Information Processing & Management*, 36(5), pp 697-716, 2000.

[16] E.M. Voorhees, C. Buckley. "The effect of topic set size on retrieval experiment error". In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp 316-323, 2002.

[17] E.M. Voorhees, D. Harman. "Overview of the sixth text retrieval conference (trec-6)", *Information Processing & Management*, 36(1), pp 3-35, 2000.

[18] Wikipedia: http://en.wikipedia.org/wiki/Borda_count.

[19] S. Wu, S. McClean. "Information retrieval evaluation of system measures for incomplete relevance judgment in IR". In *Proceedings of the 7th International conference on flexible Query Answering Systems*, pp 245-256, 2006.

[20] Y. Xu, M. Benaroch. "Information retrieval with a hybrid automatic query expansion and data fusion procedure". *Information Retrieval*, 8(1), pp 41-65, 2005.

[21] J. Zobel. "How reliable are the results of large-scale information retrieval experiments". In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 307-314, 1998.

Author Biographies

Shengli Wu was born in Nanjing, China in 1963. He received his Bachelor's degree in computer science in The University of Science and Technology of China in 1983, and received both his Master's and Ph. D. degrees in computer science in Southeast University, China, in 1989, and 1996, respectively. Since 1996, he has been working in a few different universities and research institutions in China, USA, Singapore, and UK. At present he is a lecturer in the University of Ulster, UK. His major fields of study include information systems, machine learning, and information retrieval.

Sally McClean is Professor of Mathematics at the University of Ulster. She has published and edited five books and has published over 200 research papers. In addition, she has brought in over half a million pounds worth of funding from the European Union, for research projects in the area of Statistical Databases distributed over the Internet. She has also received funding from the UK Medical Research Council for the evaluation of Intelligent Systems in Medicine and currently, from the EPSRC as part of the RIGHT project for Healthcare Modeling. Professor McClean is a Fellow of the Royal Statistical Society, and a past President of the Irish Statistical Association.