Journal of Universal Computer Science, vol. 14, no. 7 (2008), 1020-1030
submitted: 1/10/07, accepted: 21/1/08, appeared: 1/4/08 © J.UCS

Ranking Retrieval Systems with Partial Relevance Judgements

Shengli Wu
(School of Computing and Mathematics
University of Ulster, United Kingdom
s.wu1@ulster.ac.uk)

Fabio Crestani
(Faculty of Informatics
University of Lugano, Switzerland
fabio.crestani@unisi.ch)

Abstract: Some measures such as mean average precision 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 judgments are used. In this paper, we discuss how to rank retrieval systems in the condition of partial relevance judgments, which is common in major retrieval evaluation events such as TREC conferences and NTCIR workshops. Four system-oriented measures, which are mean average precision, recall level precision, normalized discount cumulative gain, and normalized average precision over all documents, are discussed. Our investigation shows that 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.

Key Words: distributed information retrieval, ranking retrieval systems
Categories: H.3.1, H.3.3

1 Introduction

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 ambiguous concept (see for example [Barry 94], [Saracevic 75], and [Shanber et al 90]) and judging relevance is a task that demands huge human effort. In some situations such as Web search, a complete set of relevance judgments is not possible. In the Text REtrieval Conference (TREC), only partial relevance judgments are conducted due to the large number of documents in the whole test collection.

In the evaluation of information retrieval systems, precision (the number of relevant documents retrieved over the total number of documents retrieved) and recall (the number of relevant documents retrieved over the total number of relevant documents in the whole collection) are regarded as the two most important measures

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. Mean average precision (MAP), recall level precision (RP), normalized discount cumulative gain (NDCG), and normalized average precision over all documents (NAP) can be regarded as candidates of good system-oriented measures. Among them, MAP 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 in [Järvelin and Kekäläinen 02] and [Kekäläinen 05]; and NAP was proposed by Wu and McClean in [Wu and McClean 06].

Without complete relevance judgments, 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 is used [Spark Jones and van Rijsbergen 75]. 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 [Zobel 98]. 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 rank a group of retrieval systems using system-oriented measures in the condition of partial relevance judgments. We find that partial relevance judgments do affect the values of system-oriented measures significantly when using the TREC's pooling method. The more incomplete the relevance judgments are, the larger 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 options are discussed in this paper. Experiments are also conducted to evaluate these methods' reliability.

2 Related work

Zobel in [Zobel 98] 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 judgments were taken. 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 investigated in [Voorhees 98] and [Voorhees 00] the effect of varying relevance judgments to the evaluation results since very often different human assessors might have different opinions about documents' relevancy to an information need. Buckley and Voorhees in [Buckley and Voorhees 00] conducted an experiment to investigate the stability of different measures including MAP and RP when using different query formats. Voorhees and Buckley conducted in [Voorhees and Buckley 02] another experiment to investigate the effect of topic set size on retrieval result. They found that using precision at 10 documents level incurred higher error rate than

using MAP in their experiment. Sanderson and Zobel in [Sanderson and Zobel 05] reran the experiment that Buckley and Voorhees did with two more groups of results and had similar observations. However, they argued that P10 was as good as MAP if considering both error rate for relative difference and human judgmental effort. Buckley and Voorhees introduced in [Buckley and Voorhees 04] a measure *bpref* for partial relevance judgments.

Järvelin and Kekäläinen introduced in [Järvelin and Kekäläinen 02] the 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 MAP. Using cumulated gain-based evaluation measures; Kekäläinen in [Kekäläinen 05] compared the effect of binary and graded relevance judgments on the rankings of information retrieval systems. Wu and McClean introduced in [Wu and McClean 06] the measure of normalized average precision over all documents (NAP). Interestingly, NAP is a special case of NDCG.

3 Four measures

In this section we discuss the four measures proposed in this paper. MAP and RP have been used many times in TREC [Voorhees and Harman 00]. Both of them are defined with binary relevance judgments and are used widely by researchers to evaluate their information retrieval systems and algorithms (e.g., in [Bodoff and Robertson 04], [Lee and Lee 05], [Xu and Benaroch 05]). MAP uses the formula, $map = \frac{1}{R}\sum_{i=1}^{R}\frac{i}{p_i}$, 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.

NAP is introduced in [Wu and McClean 06]. First let us discuss a related measure - average precision over all documents (AP). AP uses the formula, $ap = \frac{1}{n}\sum_{i=1}^{n}\frac{r(i)}{i}$, 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 *ap_best* is the best possible AP score for the given query, then NAP can be defined as *nap=ap/ap_best*.

NDCG is introduced in [Järvelin and Kekäläinen 02]. 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 [Järvelin and Kekäläinen 02], which needs to take a specific integer 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 t documents, its discount cumulated gain (DCG) is $\sum_{i=1}^{t} w(i)*r(i)$. $r(i)$ 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: $ndcg = \frac{1}{dcg\_best}\sum_{i=1}^{t} w(i) * g(i)$. In summary, all these four measures are normalized since their values are always in the range of 0 and 1 inclusive.

Buckley and Voorhees [4] introduced a measure *bpref* for partial relevance judgments. *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. This implication in *bpref* is not reasonable for those submitted runs to TREC. Using the pooling method, most relevant documents have been identified since they are very likely to appear in the first 100 documents in one or more submitted runs. The probability for those un-judged documents to be relevant is very low. Let us see an example. In TREC 8, the TREC official pool includes 86830 documents, 4728 of them are relevant. Therefore, only 4728/86830 = 5.4% of the documents in the pool is relevant. For those documents that are not in the pool, their chance to be relevant should be much lower than 5.4%. Therefore, *bpref* favours those results that have fewer documents judged. One extreme situation is no documents are judged in a result, but that result can still get the maximum possible *bpref* value 1, no matter what its real performance is. Probably *bpref* may be applicable to some situations but is not suitable well for our purpose in this paper. Therefore, we do not include *bpref* in this study.

Group	Track	Num. of results	Num. 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**

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

4 Relationship between pool depths and measure values

In this section we investigate the effect of partial relevance judgments 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 summarised in [Tab. 1].

Considering that the pooling method in TREC is a reasonable method for partial relevance judgments, we conduct an experiment to compare the values of these measures by using pools of different depths. In every year, a pool of 100 documents in depth was used in TREC to generate its *qrels* (relevance judgments 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 can be calculated using $rel\_diff=(c_i-c_{100})/c_{100}$. [Fig. 1] shows the absolute and relative differences of MAP and RP values when different *qrels* are used. Every data point in [Fig. 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, MAP is the worst considering the difference rate. When using a pool of 10 documents in depth, the absolute difference rate is 44% and the relative difference rate is 31% for MAP. In the same condition, they are 32% and 10% for RP. In all cases, the relative difference is smaller than the corresponding absolute difference. In addition, similar conclusions are observed for NDCG and NAP. The difference rates for them are higher than that for RP, but are lower than that for MAP.

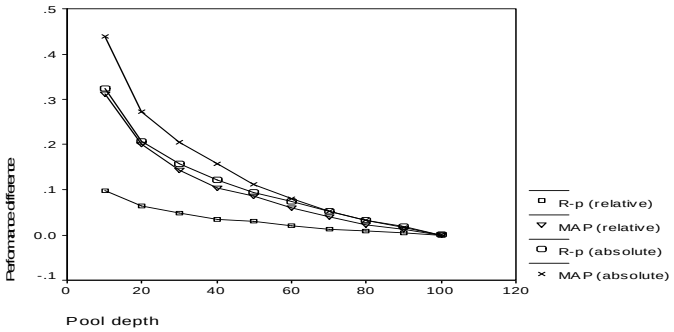

Figure 1: Value differences of two measures when using pools of different depth (the pool of 100 documents in depth is served as baseline)

Next, let us 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	G7	G8	G9	G10	G11	Total
47	16	79	76	49	33	39	27	25	17	165	699

For all these topic groups $G_1 \sim G_{11}$, we calculated the value differences of the two measures using pools of different depths. [Fig. 2] shows the experimental result for MAP and RP, respectively. Although not presented, the similar phenomena were observed for both NDCG and NAP. One common tendency for these two 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 MAP. For groups G_{10} and G_{11}, the value differences of MAP 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 MAP and RP values depends on the percentage of relevant documents identified for that topic. The larger percentage of relevant documents identified for a topic, the more accurate the estimated MAP and RP values for that topic. However, the numbers of relevant documents vary considerably from one topic to another: from 1 or 2 to several hundreds. Therefore, MAP and RP values obtained with a pool of certain depth are not comparable across topics.

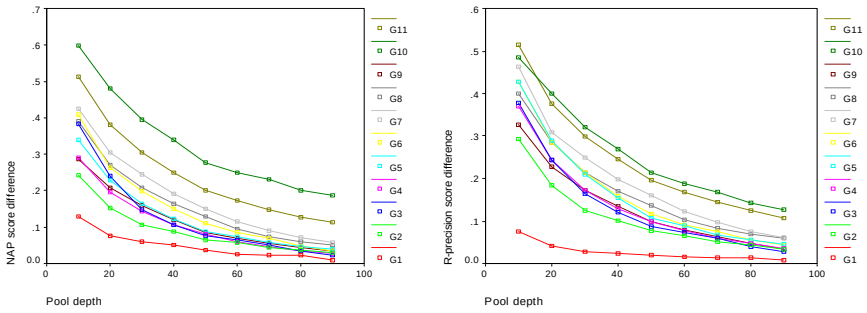

Figure 2: Difference in MAP and RP values using pools of different depths

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 MAP	Rate of exaggeration	Real MAP
A (Q1)	0.32	50%	0.32/(1+0.5)=0.1778
B (Q1)	0.25	50%	0.25/(1+0.5)=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 MAP values, we may conclude that *A* is better than *B*, because *A*'s MAP over two queries (0.32+0.45)/2=0.385 is greater than *B*'s MAP over two queries (0.25+0.50)/2=0.375. However, because Query 1's MAP is overestimated by 50% and Query 2's MAP is overestimated by 20%, a modification is needed for these MAP 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 [Section 5], we discuss some alternatives for such a task.

5 Other options than averaging all the values

Suppose for a certain collection of documents, we have a group of systems (r_1,r_2,…r_n) and a group of queries (q_1,q_2,…,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 over these queries (e.g., using any one of the 4 system-oriented measures). If complete relevance judgments are applied, then averaging these values over all the queries is no doubt the best solution. Under partial relevance judgments, the estimated values are far from accurate and are not comparable across queries, as we have demonstrated in [Section 4]. Considering a single query, if System *A* is better than System *B* with partial relevance judgments, then the same conclusion is very likely to be true with complete relevance judgments, though the difference may not be accurate. In such a situation, we may regard 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 (see its definition in [Wikipedia 07], for example) and Condorcet voting [Montague and Aslam 02] in political science can be used for this.

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 procedures 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 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 normalized 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 are used. For all the submissions in one-year group, we calculate 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, are used to rank them. For these rankings obtained using different methods, we calculate Kendall's tau coefficient for each pair of rankings obtained using the same measure but different ranking method. Table 2 show the results, each of which is for one of the four measures.

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

From [Tab. 2], we can observe that Kendall's tau coefficients in all cases are quite large. For any pair in any year group, the average is always larger 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.

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. [Tab. 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 180 pairs of rankings (20 pairs in each year group).

From [Tab. 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 larger, 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 (two-tailed t test). In some cases, the differences between Borda count and the Zero-one method are not significant.

	Averaging	Borda	Condorcet	Zero-one
1/5~all	0.7624	0.7855(.000)	0.7033	0.7765(.000)
2/5~all	0.8476	0.8658(.000)	0.7771	0.8597(.000)
3/5~all	0.8961	0.9115(.000)	0.8281	0.9071(.000)
4/5~all	0.9378	0.9454(.000)	0.8622	0.9438(.000)
Average	0.8610	0.8771[+1.87%]	0.7927[-7.93%]	0.8718[+1.25%]

Table 3: Kendall's tau coefficients for MAP (figures in parentheses indicate the significance level of difference compared with the averaging method, two-tailed t test)

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

Table 4: Kendall's tau coefficients for RP (figures in parentheses indicate the significance level of difference compared with the averaging method, two-tailed t test)

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

Table 5: Kendall's tau coefficients for NAP

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

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

7 Conclusions

In this paper we have discussed the issue of how to rank a group of information retrieval systems in the condition of partial relevance judgments. Four system-oriented measures, namely MAP, RP, NDCG, and NAP, are discussed in this paper. 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 across 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 of these four methods.

Our investigation also demonstrates that with partial relevance judgments, the evaluated results can be significantly different from the results with complete relevance judgments: 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 judgments, researchers need to be careful about the results.

References

[Barry 94] Barry, C. L.:User-defined relevance criteria: an exploratory study; *Journal of the American Society for Information Science*, 45, 3, (1994), 149-159.

[Bodoff and Robertson 04] Bodoff, D. and Robertson, S.: A new united probabilistic model; *Journal of the American Society for Information Science and Technology*, 55, 6, (2004), 471-487.

[Buckley and Voorhees 00] Buckley, C. and Voorhees, E. M.: Evaluating evaluation measure stability; In *Proceedings of ACM SIGIR'2000*, Athens, Greece (2000), 33-40.

[Buckley and Voorhees 04] Buckley, C. and Voorhees, E. M.: Retrieval evaluation with incomplete information. In *Proceedings of ACM SIGIR'2004*, Sheffield, United Kingdom, (2004), 25-32.

[Järvelin and Kekäläinen 02] Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques; *ACM Transactions on Information Systems*, 20, 4, (2002), 442-446.

[Kekäläinen 05] Kekäläinen, J.: Binary and graded relevance in IR evaluations - comparison of the efforts on ranking of IR systems; *Information Processing & Management*, 41, 5, (2005), 1019-1033.

[Lee and Lee 05] Lee, C. and Lee, G. G.: Probabilistic information retrieval model for a dependency structured indexing system; *Information Processing & Management*, 41, 2, (2005), 161-175.

[Montague and Aslam 02] Montague, M. and Aslam, J. A.: Condorcet fusion for improved retrieval; In *Proceedings of ACM CIKM'2002*, McLean, Virginia, USA, (2002), 538-548.

[Sanderson and Zobel 05] Sanderson, M. and Zobel, J.: Information retrieval system evaluation: Effort, sensitivity, and reliability; In *Proceedings of ACM SIGIR'2005*, Salvador, Brazil, (2005), 162-169.

[Saracevic 75] Saracevic, T.: 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, (1975), 321-343.

[Schamber et al 90] Schamber, L. and Eisenberg, M. B. and Nilan, M. S.:A re-examination of relevance: toward a dynamic, situational definition; *Information Processing & Management*, 26, 6, (1990), 755-776.

[Spark Jones and van Rijsbergen 75] Sparck Jones, K. and van Rijsbergen, C.: 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).

[Voorhees 98] Voorhees, E. M.: Variations in relevance judgments and the measurement of retrieval effectiveness; In *Proceedings of ACM SIGIR'1998*, Melbourne, Australia, (1998), 315-323.

[Voorhees 00] Voorhees, E. M.: Variations in relevance judgments and the measurement of retrieval effectiveness; *Information Processing & Management*, 36, 5, (2000), 697-716.

[Voorhees and Buckley 02] Voorhees, E. M. and Buckley C.: The effect of topic set size on retrieval experiment error; In *Proceedings of ACM SIGIR'2002*, Tampere, Finland, (2002), 316-323.

[Voorhees and Harman 00] Voorhees, E. M. and Harman, D.: Overview of the sixth text retrieval conference (TREC-6); *Information Processing & Management*, 36,1, (2000), 3-35.

[Wikipedia 07] Wikipedia: http://en.wikipedia.org/wiki/Borda_count; Last seen December 2007.

[Wu and McClean 06] Wu, S., and McClean, S.: Information retrieval evaluation of system measures for incomplete relevance judgment in IR; In *Proceedings of the 7th International conference on flexible Query Answering Systems*, Milan, Italy, (2006), 245-256.

[Xu and Benaroch 05] Xu, Y. and Benaroch, M.: Information retrieval with a hybrid automatic query expansion and data fusion procedure; *Information Retrieval*, 8, 1, (2005), 41-65.

[Zobel 98] Zobel, J.: How reliable are the results of large-scale information retrieval experiments; In *Proceedings of ACM SIGIR'1998*, Melbourne, Australia, (1998), 307-314.