Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 212–225

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Evaluation of retrieval effectiveness with incomplete relevance data: Theoretical and experimental comparison of three measures

Per Ahlgren [a,*], Leif Grönqvist [b]

[a] *University College of Borås, Swedish School of Library and Information Science, Sweden*
[b] *Växjö University, School of Mathematics and Systems Engineering, Sweden*

Received 7 September 2006; received in revised form 9 January 2007; accepted 10 January 2007
Available online 6 March 2007

Abstract

This paper investigates two relatively new measures of retrieval effectiveness in relation to the problem of incomplete relevance data. The measures, *Bpref* and *RankEff*, which do not take into account documents that have not been relevance judged, are compared theoretically and experimentally. The experimental comparisons involve a third measure, the well-known *mean uninterpolated average precision*. The results indicate that *RankEff* is the most stable of the three measures when the amount of relevance data is reduced, with respect to system ranking and absolute values. In addition, *RankEff* has the lowest error-rate.
© 2007 Elsevier Ltd. All rights reserved.

Keywords: Cranfield; Incomplete judgments; Retrieval effectiveness; Rank effectiveness

1. Introduction

A typical experiment in information retrieval (IR) makes use of a test collection. Such a collection consists of three components:

- A database D of documents.
- A set T of topics (information needs, written in natural language).
- For each topic in T, a non-empty set of relevant documents from D.

In the Cranfield evaluation model (Voorhees, 2002) which prescribes the use of test collections, a completeness assumption is made: for each topic $t \in T$, each document $d \in D$ is judged for relevance in relation to t (Cleverdon, 1991). In the normal case, though, D contains a large number of documents, and it is practically

[*] Corresponding author. Tel.: +46 33 435 40 65; fax: +46 33 435 40 05.
E-mail addresses: per.ahlgren@hb.se (P. Ahlgren), leif.gronqvist@msi.vxu.se (L. Grönqvist).

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.01.011

impossible to judge each document for each topic. Instead, the documents that belong to a proper subset of D are relevance judged.

In the well known TREC environment, the proper subset is generated by the *pooling* technique. This technique can be described as follows. Several searches are executed for $t \in T$. For each search, the set of the (usually 100) top ranked documents is created. The union of the created sets forms the *pool*, say P_t, of documents to be relevance judged with respect to t (Voorhees, 2004).

When a *new* system (a system that did not participate in the pooling process) is tested against a TREC collection, it may be the case that documents that have not been relevance judged with respect to t, i.e., documents that do not belong to P_t, are retrieved within the 100 top ranking positions. Such documents are *assumed* to be irrelevant when a traditional measure like precision is applied. This assumption is less satisfactory, since it may favour systems that participated in the pooling process.

Issues in retrieval evaluation have received fairly much attention in the literature (Hull, 1993; Keen, 1992; Sanderson & Zobel, 2005; Tague-Sutcliffe, 1992), and alternatives to traditional measures have been proposed, for example in Järvelin and Kekäläinen (2002). In this paper, we treat two relatively new measures, presented in 2004 and 2005, respectively, of retrieval effectiveness which do not take into account documents that have not been relevance judged (Ahlgren & Grönqvist, 2006a, 2006b; Buckley & Voorhees, 2004; Grönqvist, 2005). These measures are compared, both theoretically and experimentally. In addition, we experimentally compare the two measures to a well-known evaluation measure, *mean uninterpolated average precision* (MAP).

The rest of the paper is organized as follows. Section 2 describes the test collections used in our experiments, while a motivation for inventing new evaluation measures is given in Section 3. In Section 4, we define two fairly new evaluation measures (one of the measures is given in two variants). Section 5 provides a theoretical comparison of the measures defined in Section 4, and in the following section these measures are compared to each other, and to MAP, experimentally. In Section 7, we discuss the outcome of the experiments and put forward conclusions.

2. Test collections

Our experiments used the same test collections as was used by Buckley and Voorhees (2004), TREC-8 (ad hoc task), TREC-10 (Web track) and TREC-12 (robust track). Table 1 gives an overview of these three collections. In the last two columns, the mean number of relevant documents per pool and the mean pool size are reported. Table 2 gives, for each test collection, the number of system runs we used in our experiments. In this work, we consider two distinct runs as two distinct systems. A group that participates in a given TREC may be associated with several runs, and thereby with several systems. Also given in Table 2 is the mean number of

Table 1
Overview of the data sets used in the experiments

TREC	Documents		Topics	Average pool	
	#	GB		Relevant	Judged
TREC-8	528k	1.9	50	94.6	1736.6
TREC-10	1700k	10.0	50	67.3	1408.0
TREC-12	528k	1.9	100	60.7	1288.0

Table 2
Overview of system top 1000 outputs

TREC	Runs	Top 1000 averages		
		Retrieved	Relevant	Unjudged
TREC-8	122	997.5	53.2	555.3
TREC-10	76	989.6	41.8	562.4
TREC-12	71	999.5	33.2	558.2

Table 3
Overview of system top 100 outputs

TREC	Runs	Top 100 averages		
		Retrieved	Relevant	Unjudged
TREC-8	122	99.97	21.0	4.91
TREC-10	76	99.53	14.8	3.35
TREC-12	71	100.0	14.7	6.13

retrieved documents (in TREC, maximum 1000 per topic), relevant and unjudged documents across the runs. Since the measures we focus on in this work are dependent on the number of retrieved documents, each run that retrieved less than 95% of the maximal number of retrieved documents was excluded. For example, for TREC-8 (with 50 topics) the maximal number of retrieved documents for a run is 50 × 1000 = 50,000. Therefore, each TREC-8 run that retrieved less than 47,500 was excluded. Further, each run that did not retrieve any documents for one or more topics was excluded. These exclusion rules are the same as those employed in Buckley and Voorhees (2004). We excluded 7 runs each from TREC-8 and TREC-12, while 21 runs were excluded from TREC-10. Table 2 shows that on average (over all runs and topics) there is a relatively large number of retrieved documents that have not been relevance judged, irrespective of which of the three TRECs we consider. If we only take the top 100 documents into account, there is still a noticeable number of unjudged documents (Table 3).

3. Motivation for inventing alternative measures

We mentioned the issue of retrieved unjudged documents in the end of the preceding section. As can be seen in Fig. 1, where we exemplify with TREC-12 data, the share of documents that have not been judged is as large as 80% at ranking position 1000, over all runs and all 100 topics. For TREC-12, the top 125 documents were used during the pooling process. Despite that, the share of unjudged documents at ranking position 125 is about 13%. Even at ranking position 1, there are documents that have not been judged.

One concern with unjudged documents is that runs, which did not contribute to the pools, may retrieve unjudged relevant documents, and thereby be treated unfairly during evaluation. The same problem also applies to systems that did contribute to the pools, but to a lesser extent since the 100 top ranked documents are guaranteed to be relevance judged. Although it has been argued on empirical grounds that the TREC collections are not biased against such runs (Voorhees, 2002; Zobel, 1998), we believe that large shares of unjudged documents, especially at low ranking positions, is a sufficient motivation for the construction of measures that do not take unjudged documents into account.

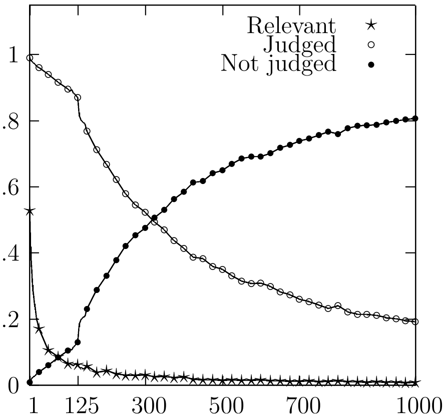

Fig. 1. Share of relevant, judged, and unjudged documents at different ranking positions, as an average over all systems and topics in TREC-12.

4. Two measures of retrieval effectiveness

In this section, we define two measures of retrieval effectiveness. One of them is put forward in two variants. Each measure is such that it relates all known relevant documents in a ranked list of documents to the same number of known irrelevant documents.

4.1. The Buckley and Voorhees measures

Buckley and Voorhees (2004) proposed a measure of retrieval effectiveness intended to be used when some retrieved documents have not been judged for relevance. To our knowledge, the measure is intended to be a measure of rank effectiveness. Such a measure favours IR methods that retrieve relevant documents early in the ranking. If retrieval effectiveness is measured by recall, an IR method may have high retrieval effectiveness, but its performance in terms of rank effectiveness may be poor.

Let R be the set of known relevant documents for a topic $t \in T$, and let r be the number of documents in R. If we assume that the pooling technique has been used, $R \subseteq P_t \subseteq D$. Let I_r be the set of the r most highly ranked (with respect to the IR-method M), known irrelevant documents for t. Further, let $I_r(d_i)$ be the number of documents d such that $d \in I_r$ and $Rank(d,t,M) < Rank(d_i,t,M)$, where $d_i \in R$. The measure, *binary preference* (*bpref*) for the IR method M with respect to the topic t, is defined as:

$$bpref(M,t) = \frac{1}{r}\sum_{d_i \in R} 1 - \frac{I_r(d_i)}{r} \tag{1}$$

It holds that $0 \leqslant bpref(M,t) \leqslant 1$. The highest performance is obtained when all the r most highly ranked, known irrelevant documents are ranked lower (higher ranks) than each known relevant document. Observe that the definition of this measure presupposes at least r known irrelevant documents for a given topic.

Since *bpref* is coarse when a topic has a small number (one or two) of known relevant documents, Buckley and Voorhees (2004) used a variant of the measure, *bpref*-10, in their experiments. Let I_{10+r} be the set of the $10 + r$ most highly ranked, known irrelevant documents for t. The variant is defined as

$$bpref\text{-}10(M,t) = \frac{1}{r}\sum_{d_i \in R} 1 - \frac{I_{10+r}(d_i)}{10+r} \tag{2}$$

where $I_{10+\mathrm{r}}(d_\mathrm{i})$ is the number of documents d such that $d \in I_{10+r}$ and $Rank(d,t,M) < Rank(d_i,t,M)$, where $d_i \in R$. Clearly, $0 \leqslant bpref\text{-}10(M,t) \leqslant 1$. With *bpref*-10, the evaluation involves at least 11 document pairs. This minimal number occurs when there is exactly one known relevant document ($r = 1$). In that case, the relevant document is related, with respect to ranking position, to 10 + 1 irrelevant documents.

4.2. Rank effectiveness: a related measure

Grönqvist (2005) suggested a measure *RankEff* that is similar to the two measures given above. *RankEff* is intended to measure to what degree (known) irrelevant documents are ranked lower than (known) relevant documents. Let I be the set of all known irrelevant documents for topic t. The measure, *rank effectiveness* (*RankEff*) for the IR method M with respect to t, is then defined as:

$$RankEff(M,t) = \frac{\sum_{d_i \in R} I(d_i)}{r(n-r)} \tag{3}$$

where $I(d_\mathrm{i})$ is the number of documents d such that $d \in I$ and $Rank(d,t,M) > Rank(d_i,t,M)$, where $d_i \in R$, and n is the number of relevance judged documents for t.[1] As for the *bpref* measures, $0 \leqslant RankEff(M,t) \leqslant 1$. The

[1] Note that I refers to *all* known irrelevant documents, while I_r ($I_{10+\mathrm{r}}$) refers to the r ($10 + r$) most highly ranked, known irrelevant documents. Also note that the definition of *RankEff* uses '>', while the definition of *bpref* uses '<'. If $Rank(d,t,M) > (<) Rank(d',t,M)$, then d is ranked lower (higher) than d'.

highest performance is obtained when all the $n - r$ known irrelevant documents are ranked lower than each known relevant document.

The *RankEff* measure gives the mean number of known irrelevant documents that are ranked lower than a known relevant document, in relation to the number of known irrelevant documents. That this statement is true is shown by the following derivation:

$$RankEff(M,t) = \frac{\sum_{d_i \in R} I(d_i)}{r(n-r)} = \frac{\left(\sum_{d_i \in R} I(d_i)/r\right) \times r}{r(n-r)} = \frac{\sum_{d_i \in R} I(d_i)/r}{n-r}$$

One potential problem with *RankEff* concerns known irrelevant documents that are not retrieved. Let the number of judged documents for a topic t be 6. Assume that two of the documents are judged relevant, four irrelevant. Assume further that IR-method M_1 retrieves all six judged documents in the first six positions, with the two relevant ones in positions 1 and 2. Assume finally that IR-method M_2 retrieves only four documents, the two relevant and two of the irrelevant, with the two relevant ones in positions 1 and 2. Under these assumptions $RankEff(M_1, t) = 1$, while $RankEff(M_2, t) = 0.5$. M_2 will thus be punished for not retrieving known irrelevant documents, which is inappropriate. To avoid scenarios of this kind, we define, for $d, d' \in D$, $Rank(d,t,M) > Rank(d',t,M)$ as true if d is not retrieved and d' is retrieved. Given this definition, $RankEff(M_1,t) = RankEff(M_2,t) = 1$.

4.3. Practical adjustments

In the TREC environment, retrieval effectiveness is evaluated at 1000 retrieved documents. It may be the case, and often is, that some known relevant documents are not among the top 1000 retrieved, for a given system and a given topic. Because of this, the definitions of the measures need to be adjusted to fit practical situations. However, we only give definitions of *bpref*-10 and *RankEff*, since we compared the latter to the former (and not to the coarser variant *bpref*) in our experiments (Section 6).

Let R' be the set of *retrieved* known relevant documents for an IR method M and a topic $t \in T$, and let $r' \leqslant r$ be the number of documents in R'. Let $I_{10+r}(d'_i)$ be the number of documents d such that $d \in I_{10+r}$ and $Rank(d,t,M) < Rank(d'_i,t,M)$, where $d'_i \in R'$. Further, let $I(d'_i)$ be the number of documents d such that $d \in I$ and $Rank(d,t,M) > Rank(d'_i,t,M)$, where $d'_i \in R'$. We then give the following practical definitions of *bpref*-10 and *RankEff*:

$$bpref\text{-}10(M,t) = \frac{1}{r} \sum_{d'_i \in R'} 1 - \frac{I_{10+r}(d'_i)}{10+r} \tag{4}$$

$$RankEff(M,t) = \frac{\sum_{d'_i \in R'} I(d'_i)}{r(n-r)} \tag{5}$$

An IR method that retrieves only a small number, say k, of known relevant documents at the first k ranks receives a smaller value on the two measures than an IR method that retrieves a large number, say $l > k$, on the l first ranks. This is a desirable property, which the measures have in common with MAP. MAP is a measure that has been shown to be relatively stable with respect to capacity to distinguish between IR methods (Buckley & Voorhees, 2000).

Note that the two measures, like MAP, give a non-optimal value when not all known relevant documents have been retrieved. Further, Eqs. (4) and (5) define the same functions as Eqs. (2) and (3), respectively, given that $r' = r$.

5. Theoretical comparison of the measures

Let $\overline{I}_r(d_i)$ be the number of documents d such that $d \in I_r$ and $Rank(d,t,M) > Rank(d_i,t,M)$, where $d_i \in R$. Consider the case where *RankEff* is modified in such a way that the measure is only applied to the r most highly ranked, known irrelevant documents. In that case we define *RankEff* as

$$RankEff(M,t) = \frac{\sum_{d_i \in R} \overline{I}_r(d_i)}{rr} \tag{6}$$

Then we obtain the following:

$$bpref(M,t) = \frac{1}{r}\sum_{d_i \in R} 1 - \frac{I_r(d_i)}{r} = \frac{1}{r}\sum_{d_i \in R} \frac{r}{r} - \frac{I_r(d_i)}{r} = \frac{1}{r}\sum_{d_i \in R} \frac{\overline{I}_r(d_i)}{r} = \frac{\sum_{d_i \in R} \overline{I}_r(d_i)}{rr} = RankEff(M,t)$$

bpref and *RankEff* are thus identical when the latter measure is applied to the r most highly ranked, known irrelevant documents, or if $r = n/2$.

In the remainder of this section we compare *bpref*-10, rather than *bpref*, to *RankEff*. One may ask if *bpref*-10 and *RankEff* are equivalent in the sense that they always agree with respect to the order of two compared IR methods. More formally, one may ask if the following statement holds:

$$bpref\text{-}10(M_1,t) < (>,=)\ bpref - 10(M_2,t)$$

if and only if

$$RankEff(M_1,t) < (>,=)\ RankEff(M_2,t)$$

However, the statement above is not true, which is shown by the following counterexample. Assume that $n = 30$ and $r = 2$, for a given topic t. Then $n - r = 28 = |I| =$ the number of known irrelevant documents for t. Assume that the methods M_1 and M_2 rank documents according to the data in Table 4, where "+" indicates a known relevant document, "–" a known irrelevant document, and "···" stands for the positions 16–29 in the lists of retrieved documents. Further, positions 16–29 are assumed to contain, for both M_1 and M_2, known irrelevant documents.

With the data in Table 4, we obtain the following for *bpref*-10 and the two methods:

$$bpref\text{-}10(M_1,t) = \frac{1}{2}\left(1 - \frac{0}{12} + 1 - \frac{12}{12}\right) = 0.500 = bpref\text{-}10(M_2,t)$$

For *RankEff* we obtain:

$$RankEff(M_1,t) = \frac{28+16}{2 \times 28} \approx 0.786$$

and

$$RankEff(M_2,t) = \frac{28+0}{2 \times 28} = 0.500$$

Thus, $RankEff(M_1,t) > RankEff(M_2,t)$. Given the data of the example, M_1 obviously performs better than M_2 with respect to rank effectiveness, and it is desirable that a measure for rank effectiveness detects this.

RankEff has some appealing properties in relation to *bpref*-10:

- It uses more information, since *each* known irrelevant document is taken into consideration.
- It can handle data sets with any number of relevant and irrelevant documents, except when the number of known relevant documents is 0, or the number of known irrelevant documents is 0 ($r = 0$ or $n = r$, respectively).
- It handles topics with a small number of relevant documents better than *bpref*-10 in the sense that unreasonably large differences in measurement values between IR methods are prevented.

Table 4
Rankings associated with two hypothetical IR methods

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	·	·	·	30
M_1	+	–	–	–	–	–	–	–	–	–	–	–	–	+	–	.	.	.	–
M_2	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	·	·	·	+

Another difference between *RankEff* and *bpref*-10 concerns the case where a known relevant document swaps position with a higher ranked known irrelevant. A measure of rank effectiveness should increase as a consequence of such a change. *RankEff* and uninterpolated average precision (AP) increase under the change. For *bpref*-10, this is not always the case. In Appendix, we give proofs of these statements.

6. Experimental comparison

In this section, we study (6.1) error rates associated with the measures, (6.2) the correlation between system rankings, generated from different measures, when the full pools of judged documents are used, and (6.3) the effects of gradually reducing the pools of judged documents on consistency of absolute average evaluation values and on the stability of system rankings. As mentioned above, the data used for the experiments come from TREC-8, TREC-10 and TREC-12. We are principally concerned with comparing *RankEff* and *bpref*-10. However, we compare the two measures to MAP, heavily used in TREC.

6.1. Error rates

One issue in retrieval evaluation is the following situation: two topic sets of the same size disagree as to which of two runs is the better. Let *EM* be an evaluation measure. An error rate in this respect can be obtained empirically by comparing the mean *EM* values of two runs on two disjoint topic sets of the same size z. If one of the runs performs better than the other for one of the two sets, but worse for the other set, a swap has occurred. When the error rates are computed, the difference in *EM* values between two runs (with respect to one of the two involved topic sets) is classified into one of (in our case) 101 bins, where each bin represents a range of differences in *EM* values: $[0, 0.002)$, $[0.002, 0.004), \ldots, [0.198, 0.2)$, $[0.2, 1]$. Now, the *error rate* for a bin b, in relation to a topic set size z, is defined as the share of the differences d in b such that d is associated with a swap. When the error rates are computed, one can obtain information, for a given topic set size, on the least difference in *EM* values for having at least 95% confidence in the conclusion. If we for example get a swap for at most 5% of the system pairs and disjoint topic sets when the initial *EM* difference is at least δ, we conclude that if the *EM* difference between two systems is at least δ, then we can be at least 95% certain that the system with the higher *EM* value actually is better. For more detailed information on the computing of error rates, see Voorhees and Buckley (2002).

We computed error rates for *bpref*-10, *RankEff* and MAP, using the maximal topic set sizes (25 for TREC-8 and TREC-10, 50 for TREC-12). We counted the share of swaps for all pairs of runs from TREC-8, TREC-10 and TREC-12, and we utilized 1000 different randomly selected pairs of disjoint topic sets for each pair of runs.

In Tables 5–7, the least absolute differences (the δ values) for having at least 95% confidence in the conclusion are given for each measure and for TREC-8, TREC-10 and TREC-12, respectively. A column labeled with '%' reports the share in percent of the best observed MAP (average *bpref*-10, average *RankEff*) value obtained for a run and for a given TREC that δ represents. A σ column gives the standard deviation over the observed MAP (average *bpref*-10, average *RankEff*) values for a given TREC. For TREC-8 and TREC-10, the largest δ values are associated with *RankEff*. However, this measure also have the largest standard deviations, which indicates that one often obtains a larger absolute difference between two systems compared to MAP and *bpref*-10. Relative to the best observed values, the differences associated with *RankEff* are less than the differences for MAP and *bpref*-10. In this relative sense, *RankEff* outperforms the two other measures.

Table 5
Significant differences for each measure for TREC-8 ($z = 25$)

Measure	TREC-8			
	Best	δ	%	σ
MAP	0.413	0.030	7.3	0.0081
Bpref-10	0.455	0.034	7.5	0.0072
RankEff	0.745	0.038	5.1	0.0252

Table 6
Significant differences for each measure for TREC-10 ($z = 25$)

Measure	TREC-10			
	Best	δ	%	σ
MAP	0.283	0.050	17.7	0.0022
Bpref-10	0.332	0.058	17.5	0.0023
RankEff	0.777	0.074	9.5	0.0063

Table 7
Significant differences for each measure for TREC-12 ($z = 50$)

Measure	TREC-12			
	Best	δ	%	σ
MAP	0.311	0.028	9.0	0.0028
Bpref-10	0.326	0.032	9.8	0.0024
RankEff	0.763	0.030	3.9	0.0094

6.2. Full pools of relevance data – correlation between system rankings

Let S be the set of the n runs executed in a given evaluation context, and let EM be an evaluation measure. A *system ranking* with respect to S and EM is a list $(r_1, \ldots, r_n)$ of the runs in S such that $EM_{\text{Avg}}(r_i) \geqslant EM_{\text{Avg}}(r_j)$ if $i < j$, where $EM_{\text{Avg}}(r_k)$ is the EM average for run r_k over the involved topics. (For MAP, EM is AP, and $EM_{\text{Avg}}(r_k) = \text{MAP}(r_k)$.)

One way to empirically study if two evaluation measures measure the same thing is to compute the correlation between two system rankings, where one ranking is obtained from one of the measures, the other from the other measure. Table 8 gives, for each of the three TRECs, correlation data for *bpref*-10, *RankEff* and MAP. Correlations are measured by Kendall's τ (Kendall & Gibbons, 1990). The Kendall correlation values between *bpref*-10 and *RankEff* are fairly weak over all TRECs. *bpref*-10 has higher correlations with MAP than with *RankEff*. One of the correlation values between *bpref*-10 and MAP is (slightly) less than, while the remaining two values are greater than 0.9, which have been used as a cut-off for equivalent rankings (Buckley & Voorhees, 2004). Figs. 2–4, one for each TREC, give a picture of the level of agreement with regard to system rankings for the three measures. The x-axis in each figure represents systems, sorted by decreasing MAP values. Both *bpref*-10 and *RankEff* exhibit a decreasing trend. However, the low Kendall correlation between *RankEff* and MAP (Table 8) is clearly mirrored in the three graphs. *RankEff* disagrees with MAP (and with *bpref*-10) as to which system has the better performance for several pairs of systems. For a noticeable case, consider Fig. 3 (TREC-10). It is difficult to see the exact numbers in the figure, but one system ranked at position 42 by MAP is ranked as number 7 by *RankEff*. Similarly we can see a system in Fig. 4 (TREC-12) ranked number 17 by MAP and 49 by *RankEff*.

6.3. Effects of gradually reducing relevance data

For a given topic $t \in T$ and a given TREC, we started with the full pool, P_t, of judged documents. Then we gradually reduced P_t in the following way. Let P_t^{rel} be the set of known relevant documents for t, and let P_t^{irr} be

Table 8
Kendall correlations between system rankings for each measure

Measures	TREC-8	TREC-10	TREC-12
MAP – *bpref*-10	0.932	0.892	0.939
MAP – *RankEff*	0.856	0.760	0.770
bpref-10 vs. *RankEff*	0.823	0.698	0.741

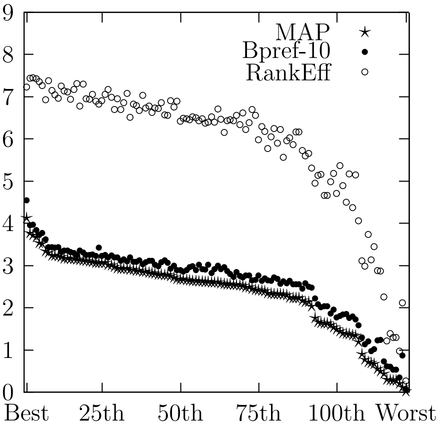

Fig. 2. Average values for each system, ordered by MAP values – TREC-8.

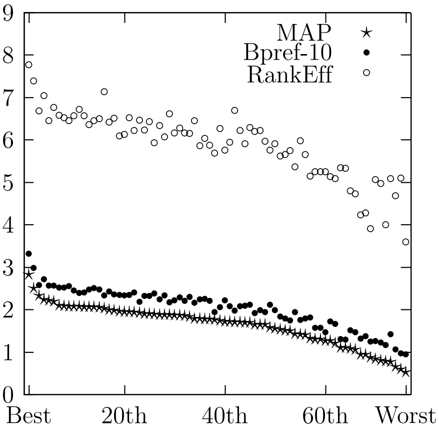

Fig. 3. Average values for each system, ordered by MAP values – TREC-10.

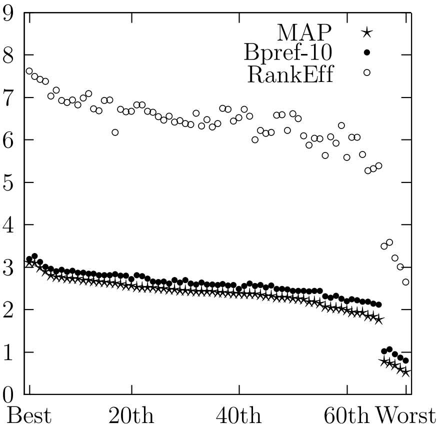

Fig. 4. Average values for each system, ordered by MAP values – TREC-12.

the set of known irrelevant documents for t. Then the union of P_t^{rel} and P_t^{irr} is equal to P_t. For each percentage value $\alpha \in V_\alpha = \{1, 2, \ldots, 10, 15, 20, \ldots, 95\}$, the corresponding number of relevant (irrelevant) documents was randomly selected from P_t^{rel} (P_t^{irr}). For example, if 95% of the relevant (irrelevant) documents was to be selected and if the number of relevant documents was 40 and the number of irrelevant ones was 200, $0.95 \times 40 = 38$ relevant and $0.95 \times 200 = 190$ irrelevant documents were selected. If the number, say x, of documents to be selected was a non-integer, we added 0.5 to x, and then the greatest integer less than $x + 0.5$ was selected. This approach yields, for each $\alpha \in V_\alpha$, α as the expectation value over the topics, for both categories of documents. As the minimal number of relevant and irrelevant documents to include in a reduced set, we used 1 and 10, respectively. If the number of relevant (irrelevant) documents corresponding to α was less than 1 (10), we used the minimum value.

6.3.1. Change in absolute average values

The graph of Fig. 5 shows, for the three measures and for TREC-12, the effects of gradually reducing the pools of judged documents on consistency of absolute average values. For each measure, a plotted value is the average over all topics and all runs, in relation to a certain level of relevance data incompleteness. The graphs for TREC-8 and TREC-10 are similar. The reduction has a clear impact on MAP and, to a lesser extent, on *bpref*-10. The former measure decreases as the level of incompleteness increases, with few exceptions. *bpref*-10 increases as the level of incompleteness increases. However, *bpref*-10 changes at a slower rate compared to MAP, with regard to $\alpha = 100$ down to $\alpha = 35$. *RankEff* is the measure that exhibits the most consistent behaviour: the average value stays approximately the same until 4% ($\alpha = 4$) of the relevance data are left.

One problem with inconsistency in the present sense is that it might affect systems that did not contribute to the pools. If such a system tends to retrieve unique relevant documents, and the used evaluation measure tends to decrease (or increase) when the relevant documents are reduced, then the system is treated unfairly (or favoured).

6.3.2. Correlations between system rankings

It is highly desirable that an evaluation measure is stable in the sense that it ranks IR methods (at least approximately) in the same relative order under different levels of incompleteness with respect to relevance data. Assume that the pooling method has been used for a given test collection, and that several relevant documents are not in the pools. Assume further that the retrieval effectiveness of a set of systems has been measured with respect to an evaluation measure *EM*, and that a system ranking is at hand. If *EM* is unstable, we might ask what would happen if more relevance data were added. It is possible that a hypothetical new ranking would disagree with the old one, perhaps to a large extent. Disagreement of this kind would be a serious problem for the Cranfield evaluation model.

The graphs in Figs. 6–8, which correspond to TREC-8, TREC-10 and TREC-12, plot the Kendall correlations between systems rankings obtained from MAP (*bpref*-10, *RankEff*) using the full relevance pools, and system rankings obtained from the same measure using reduced pools. The x-axis represents the level of relevance data incompleteness, while the y-axis shows the τ value. Optimal stability performance for a mea-

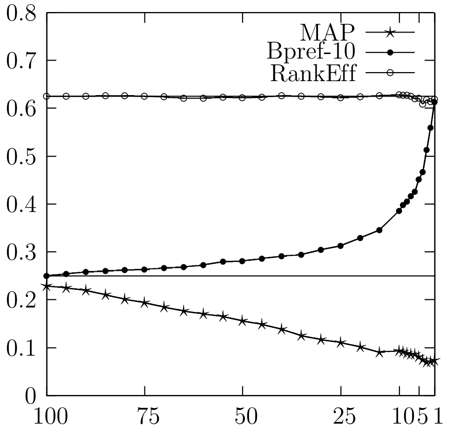

Fig. 5. Average values in TREC-12 when relevance data are reduced.

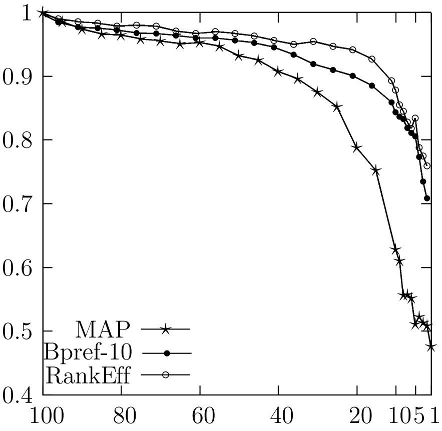

Fig. 6. Kendall correlation values for system rankings between full relevance pools and reduced for TREC-8.

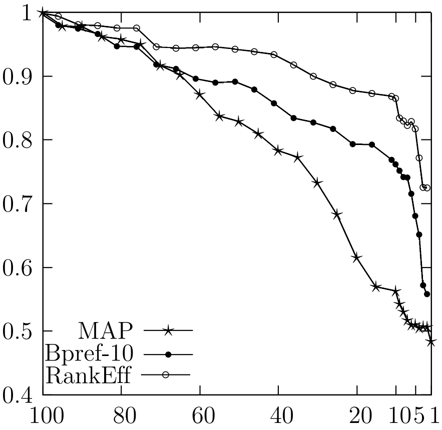

Fig. 7. Kendall correlation values for system rankings between full relevance pools and reduced for TREC-10.

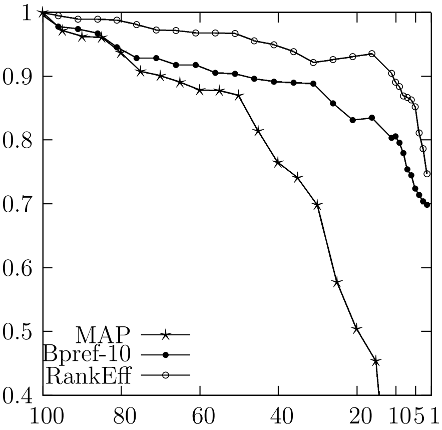

Fig. 8. Kendall correlation values for system rankings between full relevance pools and reduced for TREC-12.

sure would correspond to a straight line from the upper left corner to the upper right corner. For the test data used in this work, *RankEff* is more stable than MAP and *bpref*-10. For each of the three graphs, the plot for *RankEff* is flatter than the plots for the other two measures. As the relevance data are gradually reduced, *RankEff* goes on to rank systems in approximately the same relative order as when the relevance pools are not reduced. For TREC-8 and TREC-12, we have to move to incompleteness levels of 90% ($\alpha = 10$) and 91% ($\alpha = 9$) before the value of τ drops below 0.9. TREC-10 deviates from the other two TRECs in this respect, since its corresponding incompleteness level is 75% ($\alpha = 25$).

7. Discussion and conclusion

We have studied two relatively new measures of retrieval effectiveness, *bpref*-10 and *RankEff*. The measures have been constructed as a response to the problem of incomplete relevance data, a problem that is likely to be inherent in large test collections (like the TRECs). Both measures are such that they do not take into account documents that have not been relevance judged. We compared the measures theoretically, and we experimentally compared them to each other and to the well-known evaluation measure MAP.

The experimental results indicate that *RankEff* may be a better alternative to MAP than *bpref*-10 when the relevance data are incomplete. With respect to consistency of absolute average evaluation values as the relevance data are reduced, *RankEff* has by far the best performance. *bpref*-10 performs better than MAP, which is problematic from this perspective. As indicated in Section 6.3.1, inconsistency in this respect might affect systems that did not contribute to the pools.

System ranking stability, the ability of a measure to rank systems in the (approximately) same relative order under different levels of relevance data incompleteness, is a very important variable. For reasons given in Section 6.3.2, an evaluation measure that performs badly on this variable would constitute a serious problem for the Cranfield evaluation model. The three graphs of Section 6.3.2 (Figs. 6–8) show that *RankEff* exhibits a better performance than *bpref*-10. MAP has the worst performance, for all three considered TRECs.

The relative error rates of *RankEff* are lower than the corresponding rates for *bpref*-10, which in turn has rates similar to the rates of MAP. As pointed out in Buckley and Voorhees (2000), the error rate is only one property of an evaluation measure. A recall oriented measure with a low error rate would of course be improper to use when the retrieval of a few relevant documents at top positions is of interest. However, a low (relative) error rate is clearly a desirable property of an evaluation measure.

MAP and *bpref*-10 are in close agreement as to the ranking of systems when the full pools of judged documents are used for evaluation (Table 8), whereas the agreement between *RankEff* and *bpref*-10, and between *RankEff* and MAP, is fairly small. This indicates that *RankEff* measures something else than *bpref*-10 and MAP. We consider *RankEff* to be a measure of rank effectiveness. As pointed out in Section 5, one of the differences between *RankEff* and *bpref*-10 concerns the case where a known relevant document swaps position with a higher ranked known irrelevant. *RankEff* and AP increase under the change, while *bpref*-10 does not increase in all cases. One might argue that a measure that correlates poorly with an established measure, like MAP, is likely to be a poor measure (Buckley & Voorhees, 2004). However, if the established measure is problematic, it is questionable if a high degree of correlation is desirable.

We stress the fact that we have only worked with a limited set of test data, consisting of three test collections. Therefore, the results of the experimental part of this work should be interpreted with caution. We have obtained evidence for the relative merits of RankEff, but the measure should be tested against *bpref*-10 and MAP (and perhaps against other evaluation measures) using other test collections with incomplete relevance data.

Finally, if measures like *RankEff* and *bpref*-10 are used for evaluation in environments like TREC, one might consider instructing the participating groups to rank the full pool of judged documents for a topic. In that way, there would be no need for the practical adjustments described in Section 4.3, and *RankEff* and *bpref*-10 could be used as they are intended to be used. Moreover, since *RankEff* correlates poorly with both *MAP* and *bpref*-10 (Table 8), the issue of what aspect of retrieval effectiveness *RankEff* is actually measuring should be further addressed.

Appendix. The three measures and the swap property

A desirable property of an evaluation mesaure *EM* is the following: (SWAP) If a known relevant document swaps position with a higher ranked known irrelevant, then the value on *EM* increases. Below, we prove that *RankEff* and AP satisfy SWAP, and that it is not always the case that *bpref*-10 satisfies SWAP.

Proposition 1. *RankEff satisfies SWAP.*

Proof. Let d' be the known relevant document that swaps position with a higher ranked known irrelevant, say d''. It is clear that $I(d_{\mathrm{r}})$, the number of documents d such that $d \in I$ (the set of known irrelevant documents) and $Rank(d) > Rank(d_r)$, is not affected by the swap, for each known relevant d_r such that d_r is ranked lower than d' or higher than d''. Now, since d' swaps position with d'', which is irrelevant and higher ranked, $I(d')$ increases with at least 1. Moreover, for each known relevant document d_{r} such that $Rank(d'') < Rank(d_r) < Rank(d')$, $I(d_{\mathrm{r}})$ increases with at exactly 1. But then the value on *RankEff* increases. □

Proposition 2. *AP satisfies SWAP.*

Proof. Let d' be the known relevant document that swaps position with a higher ranked known irrelevant, d''. The precision at each known relevant document d such that d is ranked lower than d' or higher than d'' is not affected by the swap. We consider two possible cases. (1) There is no known relevant document between d' and d'' in the ranking. Then, since the new rank for d' is less than its original rank, and the number of known relevant documents up to d' is the same as it was before the swap, the new precision at d' is greater than the original precision at d'. Therefore, the value on AP increases. (2) There is at least one known relevant document between d' and d'' in the ranking. Let $d_1, \ldots, d_m$ $(1 \leqslant m)$ be the known relevant documents between d' and d'' in the ranking, i.e., $Rank(d'') < Rank(d_i) < Rank(d')$, for $1 \leqslant i \leqslant m$. Further, assume that $Rank(d_i) > Rank(d_j)$ if $i < j$. Let $P(d)$ be the original precision at d, and $P'(d)$ the new (after the swap) precision at d. Then

$$P'(d_1) > P(d'), P'(d_2) > P(d_1), \ldots, P'(d_m) > P(d_{m-1}), P'(d') > P(d_m).$$

But then the value on AP increases. □

Proposition 3. *It is not always the case that bpref-*10 *satisfies SWAP.*

Proof. Let d' be the known relevant document that swaps position with a higher ranked known irrelevant, d''. Assume that $Rank(d'') > Rank(d)$, for each $d \in I_{10+r}$, the set of the $10 + r$ most highly ranked, known irrelevant documents (where r is the number of known relevant documents for the topic). In such a situation, $I_{10+\mathrm{r}}(d_{\mathrm{r}})$, the number of documents d such that $d \in I_{10+r}$ and $Rank(d) < Rank(d_r)$, is the same after the swap as it was before the swap, for each known relevant document d_{r}. Therefore, the value on *bpref*-10 is the same after the swap as it was before the swap, and thus the value does not increase. □

References

Ahlgren, P., & Grönqvist, L. (2006a). Retrieval evaluation with incomplete relevance data: a comparative study of three measures (poster abstract). In *Proceedings of the 15th ACM international conference on information and knowledge management* (pp. 872–873).

Ahlgren, P., & Grönqvist, L. (2006b). Measuring retrieval effectiveness with incomplete relevance data. In *InSCit2006, Current research in information sciences and technologies: Multidisciplinary approaches to global information systems, Vol. I* (pp. 74–78).

Buckley, C., & Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 33–40).

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 25–32).

Cleverdon, C. W. (1991). The significance of the Cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 3–12).

Grönqvist, L. (2005). Evaluating latent semantic vector models with synonym tests and document retrieval. In *ELECTRA workshop: Methodologies and evaluation of lexical cohesion techniques in real-world applications beyond bag of words, in association with ACM SIGIR*.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 329–338).

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions of Information Systems, 20*(4), 422–446.

Keen, E. M. (1992). Presenting results of experimental retrieval comparisons. *Information Processing and Management, 28*(4), 491–502.

Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). Edward Arnold.

Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 162–169).

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management, 28*(4), 467–490.

Voorhees, E. M. (2002). The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems. Proceedings of CLEF 2001* (pp. 355–370).

Voorhees, E. M. (2004). Overview of TREC 2003. In *Proceedings of the twelfth text retrieval conference (TREC 2003)* (pp. 1–13).

Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 316–323).

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 307–314).