Rethinking How to Extend Average Precision to Graded Relevance

Marco Ferrante[1], Nicola Ferro[2], and Maria Maistro[2]

[1] Dept. of Mathematics, University of Padua, Italy
ferrante@math.unipd.it
[2] Dept. of Information Engineering, University of Padua, Italy
{ferro,maistro}@dei.unipd.it

Abstract. We present two new measures of retrieval effectiveness, inspired by *Graded Average Precision (GAP)*, which extends *Average Precision (AP)* to graded relevance judgements. Starting from the random choice of a user, we define *Extended Graded Average Precision (xGAP)* and *Expected Graded Average Precision (eGAP)*, which are more accurate than GAP in the case of a small number of highly relevant documents with high probability to be considered relevant by the users. The proposed measures are then evaluated on TREC 10, TREC 14, and TREC 21 collections showing that they actually grasp a different angle from GAP and that they are robust when it comes to incomplete judgments and shallow pools.

1 Introduction

Average Precision (AP) [2] is a simple and popular binary measure of retrieval effectiveness, which has been longly studied and discussed. Robertson et al. [9] proposed *Graded Average Precision (GAP)*, an extension of AP to graded relevance together with a probabilistic interpretation of it, which allows for different emphasis on different relevance grades according to user preferences.

When it comes to graded relevance judgements, the need to develop systems able to better rank highly relevant documents arises but it also poses challenges for their evaluation. Indeed, unstable results may come up due to the relatively few highly relevant documents [10] and this may become further complicated when you consider also a user model as the one of GAP, where varying importance can be attributed to highly relevant documents according to the user view point.

In this paper we propose two extensions to GAP, called *Extended Graded Average Precision (xGAP)* and *Expected Graded Average Precision (eGAP)*, which reformulate the probabilistic model behind GAP putting even more emphasis on the user and which are able to better cope with the case when the user attributes high importance to few highly relevant documents. The experimental evaluation, in terms of correlation analysis and robustness to incomplete judgments, confirms that xGAP and eGAP take a different angle from GAP when it comes to users attributing high importance to few highly relevant documents

E. Kanoulas et al. (Eds.): CLEF 2014, LNCS 8685, pp. 19–30, 2014.

and that they are robust to incomplete judgments and shallow pools, thus not requiring costly assessments. Moreover, the evaluation provides also some more insights on GAP itself, not present in its original study [9].

The paper is organized as follows: Section 2 considers the general problem of passing from binary to multi-graded relevance; Section 3 briefly recalls the GAP measure and outlines some of its possible biases; Section 4 and Section 5 introduce, respectively, the xGAP and the eGAP metrics, outlining the difference with GAP; Section 6 conducts a thorough experimental evaluation of the proposed measures; finally Section 7 draws some conclusions and provides an outlook for future work.

2 Mapping Binary Measures into Multi-graded Ones

Given a ranked list of N documents for a given topic, we will denote by $r[j]$ the relevance of the document at the rank j. The relevance will be an integer belonging to $S(c) = \{0, \ldots, c\}$, where 0 denotes a not relevant document and the higher the integer the higher the relevance. A measure of retrieval effectiveness will be defined *binary* if $c = 1$ and *multi-graded* if $c > 1$.

The basic binary measures of retrieval effectiveness, recall and precision, can be defined as follows $Rec[n] = \frac{\sum_{i=1}^{n} r[i]}{RB}$ and $Prec[n] = \frac{\sum_{i=1}^{n} r[i]}{n}$, where $n \leq N$ is the rank and RB, the recall base, is the total number of relevant documents for the given topic. As a consequence, AP can be defined as follows

$$AP = \frac{1}{RB} \sum_{n=1}^{N} r[n] Prec[n] = Rec[N] \; \frac{1}{\sum_{n=1}^{N} r[n]} \sum_{n=1}^{N} r[n] Prec[n] \; . \qquad (1)$$

The last expression highlights how AP can be derived as the product of the recall and the arithmetic mean of the precision at each relevant retrieved document.

When you have to apply these binary measures in a multi-graded context, the typical approach is to map the multi-graded judgments into binary ones according to a fixed threshold $k \geq 1$ in the grade scale and then compute the binary measure according to its definition. This approach actually leads to a family of measures depending on the threshold used to map the multi-graded relevance scale into the binary one. For example, [10] studies the effect of setting this threshold at different levels in the grade scale.

We now show how the above mentioned approach can be directly embedded into evaluation measures, further highlighting that it gives raise to a whole family of measures. Indeed, instead of mapping the judgements to binary ones and then apply a binary measures, you can make a binary measure parametric on the mapping threshold and obtain a different version of it for each threshold. Following [9], we assume that any user owns a binary vision (relevant/not-relevant document), but at a different level of relevance, which is the mapping threshold k. Indeed, if for a given topic we denote by $R(k)$ the total number of documents with relevance k, their recall base is $RB(k) = R(k) + R(k+1) + \ldots + R(c)$. Note that $k \to RB(k)$ is a integer-valued, non negative and non increasing function and it is useful define $\tau := \max\{k : RB(k) > 0\}$.

There, a user with threshold k defines recall as $Rec[n](k) = \frac{\sum_{i=1}^{n} r[i](k)}{RB(k)}$ if $k \leq \tau$ and 0 otherwise, precision as $Prec[n](k) = \frac{\sum_{i=1}^{n} r[i](k)}{n}$ and AP as

$$AP(k) = \frac{1}{RB(k)} \sum_{n=1}^{N} r[n](k) Prec[n](k) = \frac{1}{RB(k)} \sum_{n=1}^{N} \frac{1}{n} \Big[\sum_{m=1}^{n} \delta_{m,n}(k) \Big] \qquad (2)$$

for $k \leq \tau$ and zero otherwise, where $r[n](k) = 1$ if $r[n] \geq k$, zero otherwise, and

$$\delta_{m,n}(k) = \begin{cases} 1 & \text{if } r[m] \geq k, r[n] \geq k \\ 0 & \text{otherwise .} \end{cases} \qquad (3)$$

As discussed above, this user's oriented vision leads to a family of measures, depending on the threshold k chosen by each user. In order to obtain a single measure of retrieval effectiveness (and not a family), [9] assumes that the users, and so their thresholds in the grade scale, are distributed in the total population according to a given probability distribution. This opens the way to two alternative approaches to define a multi-graded measures based on user thresholds:

1. To define a new multi-graded measure whose internals are based on some expected quantities dependent on user's thresholds;
2. To evaluate the expectation of a binary measure at different user's thresholds.

GAP is defined following the first approach and in the next section we will argue that it presents some bias when few relevant documents are the only one considered relevant by a user.

To overcome this problem, we provide two solutions corresponding to the two alternative approaches above: in Section 4 we follow the first approach and introduce xGAP which defines a new multi-graded measure from scratch adopting a philosophy similar to GAP; in Section 5 we follow the second approach and introduce eGAP, which provides a new multi-graded extension of AP by taking the expectation of (2).

3 Graded Average Precision

Let Ω be the sample space of all the possible users and assume that a user fixes a threshold k strictly positive in $S(c)$ with probability g_k. This can be formalized defining the threshold of a user by a random variable K from Ω into $S(c)$ with distribution $(g_0, g_1, \ldots, g_c)$, where $g_0 = 0$. Using this notation, in [9] they evaluate the expected precision with respect to g of each relevant document in the ranked list, then sum up all these expected values and normalise the result dividing by its maximum. Their computation leads to the following definition:

$$GAP = \frac{\sum_{n=1}^{N} \frac{1}{n} \sum_{m=1}^{n} \Delta_{m,n}}{\sum_{k=1}^{c} R(k) \sum_{j=1}^{k} g_j}, \qquad (4)$$

where $\Delta_{m,n} = \sum_{h=1}^{\min\{r[m], r[n]\}} g_h$ with the convention that $\sum_{h=1}^{0} = 0$. If $\nu = \min\{i : g_i \neq 0\}$, the previous formula is well defined just for $\nu \leq \tau$. Indeed, if

$\nu > \tau$, none of the relevant documents is considered relevant by any user almost surely and for this reason we will define $GAP = 0$ in this case. Furthermore, it is easy to prove that $\mathbb{E}[RB(K)] = \sum_{k=1}^{c} RB(k) g_k = \sum_{k=1}^{c} R(k) \sum_{j=1}^{k} g_j$, and so

$$GAP = \frac{1}{\mathbb{E}[RB(K)]} \sum_{n=1}^{N} \mathbb{E}[r[n](K) Prec[n](K)].$$

GAP can thus be obtained substituting the expected values of the graded precision and the graded recall base in (1). Note that this is quite different from taking $\mathbb{E}[AP(K)]$, where $AP(K)$ is the composition of K with (2), since $RB(K)$ and $Prec[n](K)$ are not independent and, even if they were, Jensen's inequality ensures that $\mathbb{E}[1/X] < 1/\mathbb{E}[X]$ for any non trivial positive random variable X. This confirms that, to introduce a multi-graded measures, GAP adopts the first of the two approaches outlined in the previous section and not the second one.

[9] also defines GAP as the expectation of the following three steps random experiment: (i) select a document that is considered relevant by a user (accordingly to the user model described above) at random and let the rank of this document be n; (ii) select a document at or above rank n, at random and let the rank of that document be m; (iii) output 1 if the document at rank m is also considered relevant by the user.

In the first step, to avoid problems for the possible absence of highly relevant documents, [9] defines the slightly artificial probability to select at random the document at rank n as $\frac{\sum_{j=1}^{r[n]} g_j}{\sum_{i=1}^{c} R(i) \sum_{j=1}^{i} g_j}$. This choice leads to issues exactly in these corner cases. Indeed, consider the case where $c = 2$, $R(1) = 10$ and $R(2) = 1$. If the probabilities g_1 and g_2 are both $1/2$, we get that the probability to select one of the 10 documents of relevance 1 is equal to $1/12$, while the probability to select the only document with relevance 2 is $1/6$, that appears a reasonable set of values. However, if we increase the probability g_2 up to $9/10$, i.e. the unique relevant document for nine users over ten will be that of relevance 2, we get that the probability to select "at random" this document is just equal to $1/2$.

An additional bias in the definition of GAP can be observed in the following case. Let again $c = 2$ and assume that a run presents the first n documents of relevance 1 and then a unique document of relevance 2, followed possibly by additional non relevant documents. It is easy to prove, that for n that goes to infinity, the value of GAP tends to 1, independently from the values of g_1 and g_2. This means that, even when g_2 is close to 1 and the user is interested just in that highly relevant document appearing at the end of a (infinitely) long ranking, GAP will evaluate the system as approaching the performance of the ideal one instead of a very bad one.

4 Extended Graded Average Precision (xGAP)

To overcame the previous possible biases, we propose to define an extended version of GAP by reconsidering the "user" role in the previous three steps random experiment . So, to evaluate the probability to select at step 1 at random

the document d_n at rank n, we will assume to choose first at random a user and then to select at random among the documents considered relevant by this user. This new interpretation leads to the probability to select the document d_n equal to $\sum_{k=1}^{r[n]} \frac{1}{RB(k)}\ g_k$ where we take into account the different size of any relevance class (recall that $\sum_{k=1}^{0} = 0$). Note that, assuming again that $R(1) = 10$ and $R(2) = 1$, for $g_1 = g_2 = 1/2$ we get here that we choose at random any of the relevance 1 documents with probability $1/22$ and the only relevance 2 document with probability $12/22$, but when $g_2 = 9/10$, the probability to select the document of relevance 2 is now $100/110$.

Following the same computation in [9] for the steps 2 and 3, we again obtain that the probability that the document d_m at rank $m \leq n$ is relevant when d_n is relevant, is equal to $\frac{1}{n}\frac{\sum_{m=1}^{n} \Delta_{m,n}}{\sum_{k=1}^{r[n]} g_k}$ Collecting all the previous results and changing the order of the summation we define the *Extended Graded Average Precision (xGAP)* as:

$$xGAP = \sum_{n=1}^{N} \frac{1}{n} \left[\frac{\sum_{k=1}^{r[n]} \frac{g_k}{RB(k)}}{\sum_{k=1}^{r[n]} g_k} \Big(\sum_{m=1}^{n} \Delta_{m,n} \Big) \right] \tag{5}$$

when $\nu \leq \tau$ and 0 otherwise. Note that, in the case of a run with an increasing number of documents with relevance 1, followed by only one document with relevance 2, the value of xGAP as n tends to infinity converges to $1 - g_2^2$, a much more reasonable value.

5 Expected Graded Average Precision (eGAP)

Let us now apply the second approach to define a multi-graded extension of AP. Take the function (2), compose this with the random variable K that defines the relevance threshold of any user and take the expectation of this composed random variable. We will obtain the following new measure that we call *Expected Graded Average Precision (eGAP)*

$$eGAP = \mathbb{E}[AP(K)] = \sum_{n=1}^{N} \frac{1}{n} \Big[\sum_{k=1}^{\tau} \frac{g_k}{RB(k)} \Big(\sum_{m=1}^{n} \delta_{m,n}(k) \Big) \Big] \tag{6}$$

Note that eGAP can be also thought as an approximation of the mean areas under the Precision-Recall curves at any threshold k.

eGAP itself can be obtained as the expectation of a random experiment. The main issue will be again how to realise the random selection of a relevant document, that we will interpret here as "select at random a user, s/he fixes a threshold and select, at random, one document relevant for this user". This approach can be expressed as a four steps random experiment, whose expectation will provide an alternative definition of eGAP: (i) select at random a user and let k be his/her relevance threshold; (ii) select at random a document relevant to this user. Let its rank be n, if in the ranked list, or ∞ otherwise; (iii) in the

first case, select a document at or above rank n and let its rank be m; otherwise let the rank of this second document be ∞ as well; (iv) output 1 if the document at rank m, is also considered relevant by the user.

This differs from the random experiment used for defining GAP, because the first two steps, that we already implicitly used to derive xGAP, replace the single request to select at random a relevant document for the user. Moreover, in the fourth step the user who still considers relevant the document at rank m is the same user of the first step, something that was unclear in the definition in [9].

Let us now make explicit the random experiment: for simplicity, let us assume that all relevant documents are in the ranked list, so we have not to pay attention to the case of an ∞ rank. The first step corresponds to define the random variable K as above which takes values in $S(c)$. The second step consists in choosing a second random variable X, whose law conditioned by $\{K = k\}$ will be uniform on $\mathcal{R}(k) = \{j \in \{1, \ldots, N\} : r[j] \geq k\}$. In the third step we define a random variable Y thanks to its conditional law given that $X = n$ and $K = k$, with $Y|X = n, K = k$ uniformly distributed on the set $\{1, 2, \ldots, n\}$. The last step means to define the Binomial random variable $Z = 1_A$, where $A = \{\text{the document at rank } Y \text{ is considered relevant by the user}\}$. "Taking the expectation" of this random experiment means evaluate $\mathbb{E}[Z]$. This can be done using the smoothing property of the conditional expectation (see e.g [8], Chapter 10) and we obtain

$$\mathbb{E}[Z] = \sum_{k=1}^{c} \Big[\sum_{n=1}^{+\infty} \mathbb{P}[r[Y] \geq K | X = n, K = k] \; \mathbb{P}[X = n | K = k] \; g_k \Big] \tag{7}$$

As before, $\mathbb{P}[X = n|K = k] = \frac{1}{RB(k)} \; 1_{\{r[n] \geq k\}} \; g_k$ if $k \leq \tau$ and 0 otherwise, while

$$\mathbb{P}[r[Y] \geq K | X = n, K = k] = \frac{1}{n} \cdot \big| \{i \in \{1, \ldots, n\} : r[i] \geq k\} \big| = \frac{1}{n} \sum_{m=1}^{n} \delta_{m,n}(k)$$

with $\delta_{m,n}(k)$ defined in (3). Changing the order of the summation in (7), we obtain:

$$\mathbb{E}[Z] = \sum_{n=1}^{N} \frac{1}{n} \Big[\sum_{k=1}^{\tau} \frac{g_k}{RB(k)} \big(\sum_{m=1}^{n} \delta_{m,n}(k) \big) \Big]$$

which is exactly eGAP. As for xGAP the way to choose a relevant document at the first step fix the bias in the definition of GAP when few highly relevant documents are present in a topic, but most of the users considers only these as relevant. Moreover, going back to the example of a run with an increasing number n of low-relevance documents followed by a unique highly relevant one, as n approaches ∞ the value of eGAP converges to $1 - g_2 = g_1$ which is again a reasonable limit value for this very special situation.

6 Evaluation

Experimental Setup. We compare our proposed measures eGAP and xGAP to GAP [9] and AP [2], which are the main focus of the paper. We also consider

Table 1. Main features of the adopted data sets

Feature	TREC 10	TREC 14	TREC 21
Track	Web	Robust	Web
Corpus	WT10g	AQUAINT	ClueWeb09
# Documents	1.7M	1.0M	1040.0M
# Topics (few highly rel/key)	50 (17)	50 (6)	50 (10 / 7)
# Runs (above 1Q in terms of MAP)	95 (71)	74 (55)	27 (20)
Run Length	1,000	1,000	10,000
Relevance Degrees	3	3	4
Pool Depth	100	55	25 and 30
Minimum # Relevant	2	9	6
Average # Relevant	67.26	131.22	70.46
Maximum # Relevant	372	376	253

other measures of interest: *Normalized Discounted Cumulated Gain (nDCG)* [5], *Rank-Biased Precision (RBP)* [7], and, *Binary Preference (bpref)* [1].

We investigate the following aspects: (1) the correlation among measures using Kendall's tau [6,10]; (2) the robustness of the measures to incomplete judgements according to the stratified random sampling method [1].

We used the following data sets: TREC 10, 2001, Web Track [4]; TREC 14, 2005 Robust Tack [11]; and, TREC 21, 2012, Web Track [3], whose features are summarized in Table 1. For binary measures, we adopted a "lenient" mapping, i.e. every document above not relevant is considered as binary relevant. To prevent poorly performing systems from affecting the experiments, we considered only the runs above the first (lower) quartile as measured by MAP.

We explored two distinct cases: (a) considering all the topics in the collection; (b) considering only the topics for which $R(1) \geq 10 \cdot R(k), k = 2, 3$ and $R(k) \neq 0$, i.e. when there are few highly relevant/key documents with respect to the relevant ones and the bias of GAP, addressed by xGAP and eGAP, is more pronounced.

The full source code of the software used to conduct the experiments is available for download[1] in order to ease comparison and verification of the results.

Correlation Analysis. Table 2 reports the correlations among measures for the TREC 10 and TREC 14 collections while Table 3 reports those for the TREC 21 collection. Correlations greater than 0.9 should be considered equivalent and those "less than 0.8 generally reflect noticeable changes in rankings" [10]. GAP, xGAP, and eGAP share the same values of g_1 and g_2 (and g_3 in the case of TREC 21). For each measure, the n-ple $\tau_{\text{all\_topics}}/\tau_{\text{fewHRel\_topics}}$, (and also $\tau_{\text{fewKey\_topics}}$ in the case of TREC 21) is reported: the first value indicates the correlation computed considering all the topics; the second value indicates the correlation computed only on those topics with few highly relevant documents ($R(1) \geq 10 \cdot R(2)$); and, in the case of TREC 21, the third value indicates the correlation computed only on those topics with few key documents ($R(1) \geq 10 \cdot R(3)$).

The correlation among GAP, xGAP, and eGAP is always 1 when only one $g_i = 1.00$ and the others are zero, as a consequence of the fact that in these cases,

[1] `http://matters.dei.unipd.it/`

Table 2. Kendall's correlation analysis for TREC 10 and TREC 14

(g_1, g_2)	TREC 10, 2001, Web					TREC 14, 2005, Robust				
(0.0, 1.0)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.69/0.42	0.67/0.44	0.67/0.61	0.68/0.42	1.00/1.00	0.78/0.26	0.73/0.26	0.66/0.18	0.78/0.23
xGAP	1.00/1.00	0.69/0.42	0.67/0.44	0.67/0.61	0.68/0.42	1.00/1.00	0.78/0.26	0.73/0.26	0.66/0.18	0.78/0.23
eGAP	1.00/1.00	0.69/0.42	0.67/0.44	0.67/0.61	0.68/0.42	1.00/1.00	0.78/0.26	0.73/0.26	0.66/0.18	0.78/0.23
(0.1, 0.9)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.84/0.80	0.75/0.74	0.64/0.67	0.80/0.76	1.00/1.00	0.94/0.92	0.84/0.76	0.63/0.37	0.79/0.74
xGAP	0.86/0.67	0.73/0.48	0.71/0.51	0.69/0.66	0.73/0.49	0.86/0.61	0.83/0.55	0.77/0.51	0.66/0.27	0.79/0.48
eGAP	0.85/0.64	0.72/0.45	0.70/0.47	0.68/0.64	0.72/0.45	0.85/0.54	0.82/0.49	0.76/0.47	0.67/0.27	0.79/0.43
(0.2, 0.8)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.89/0.89	0.77/0.76	0.63/0.64	0.83/0.81	1.00/1.00	0.97/0.96	0.85/0.75	0.62/0.38	0.78/0.74
xGAP	0.84/0.64	0.76/0.54	0.73/0.56	0.68/0.68	0.75/0.54	0.88/0.68	0.86/0.65	0.79/0.58	0.67/0.29	0.80/0.55
eGAP	0.84/0.60	0.75/0.49	0.72/0.51	0.68/0.67	0.74/0.50	0.86/0.61	0.84/0.59	0.78/0.54	0.66/0.26	0.79/0.48
(0.3, 0.7)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.92/0.93	0.78/0.77	0.63/0.63	0.84/0.82	1.00/1.00	0.98/0.98	0.85/0.75	0.62/0.38	0.79/0.75
xGAP	0.86/0.68	0.80/0.61	0.77/0.63	0.68/0.70	0.79/0.60	0.90/0.74	0.89/0.72	0.82/0.63	0.66/0.32	0.81/0.61
eGAP	0.84/0.61	0.78/0.54	0.74/0.54	0.68/0.69	0.76/0.54	0.89/0.68	0.87/0.67	0.81/0.60	0.66/0.29	0.80/0.53
(0.4, 0.6)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.94/0.95	0.79/0.78	0.63/0.62	0.85/0.83	1.00/1.00	0.98/0.99	0.85/0.76	0.62/0.38	0.78/0.74
xGAP	0.88/0.71	0.83/0.66	0.78/0.67	0.68/0.70	0.81/0.65	0.93/0.80	0.92/0.78	0.83/0.69	0.65/0.34	0.80/0.66
eGAP	0.86/0.65	0.81/0.61	0.76/0.59	0.68/0.71	0.79/0.60	0.91/0.76	0.90/0.75	0.83/0.65	0.67/0.32	0.81/0.59
(0.5, 0.5)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.95/0.97	0.79/0.78	0.62/0.61	0.85/0.83	1.00/1.00	0.99/0.99	0.85/0.76	0.61/0.39	0.78/0.74
xGAP	0.90/0.76	0.86/0.73	0.79/0.72	0.67/0.69	0.82/0.71	0.94/0.84	0.93/0.83	0.84/0.72	0.64/0.35	0.80/0.70
eGAP	0.88/0.70	0.84/0.67	0.77/0.64	0.67/0.70	0.81/0.65	0.92/0.81	0.92/0.81	0.84/0.69	0.65/0.35	0.81/0.64
(0.6, 0.4)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.96/0.98	0.79/0.77	0.62/0.61	0.84/0.82	1.00/1.00	0.99/0.99	0.85/0.76	0.61/0.39	0.77/0.74
xGAP	0.92/0.83	0.89/0.80	0.80/0.76	0.65/0.66	0.83/0.76	0.96/0.87	0.95/0.86	0.85/0.73	0.64/0.35	0.79/0.73
eGAP	0.91/0.77	0.88/0.74	0.79/0.70	0.66/0.69	0.83/0.71	0.94/0.87	0.94/0.86	0.86/0.73	0.65/0.38	0.80/0.69
(0.7, 0.3)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.97/0.99	0.79/0.77	0.61/0.60	0.85/0.82	1.00/1.00	1.00/1.00	0.86/0.76	0.61/0.39	0.77/0.74
xGAP	0.94/0.86	0.92/0.85	0.80/0.77	0.65/0.66	0.85/0.79	0.97/0.91	0.97/0.90	0.85/0.75	0.62/0.36	0.78/0.74
eGAP	0.93/0.83	0.90/0.82	0.80/0.74	0.65/0.67	0.83/0.77	0.97/0.90	0.97/0.90	0.86/0.76	0.63/0.38	0.79/0.71
(0.8, 0.2)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.98/0.99	0.79/0.77	0.61/0.60	0.85/0.82	1.00/1.00	1.00/1.00	0.85/0.76	0.61/0.39	0.77/0.74
xGAP	0.95/0.91	0.94/0.90	0.81/0.78	0.64/0.63	0.85/0.81	0.98/0.94	0.99/0.93	0.86/0.75	0.62/0.37	0.78/0.74
eGAP	0.94/0.90	0.93/0.89	0.79/0.76	0.64/0.65	0.84/0.80	0.98/0.95	0.98/0.95	0.86/0.77	0.62/0.39	0.78/0.73
(0.9, 0.1)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	0.99/1.00	0.79/0.77	0.61/0.60	0.85/0.82	1.00/1.00	1.00/1.00	0.86/0.76	0.61/0.39	0.77/0.74
xGAP	0.97/0.95	0.96/0.94	0.80/0.78	0.63/0.62	0.86/0.82	0.99/0.97	0.99/0.97	0.86/0.76	0.61/0.38	0.77/0.75
eGAP	0.97/0.95	0.96/0.95	0.79/0.77	0.63/0.62	0.85/0.82	1.00/0.98	1.00/0.98	0.86/0.77	0.61/0.39	0.77/0.74
(1.0, 0.0)	GAP	AP	nDCG	RBP	bpref	GAP	AP	nDCG	RBP	bpref
GAP	1.00/1.00	1.00/1.00	0.79/0.77	0.60/0.60	0.85/0.82	1.00/1.00	1.00/1.00	0.86/0.76	0.61/0.39	0.77/0.74
xGAP	1.00/1.00	1.00/1.00	0.79/0.77	0.60/0.60	0.85/0.82	1.00/1.00	1.00/1.00	0.86/0.76	0.61/0.39	0.77/0.74
eGAP	1.00/1.00	1.00/1.00	0.79/0.77	0.60/0.60	0.85/0.82	1.00/1.00	1.00/1.00	0.86/0.76	0.61/0.39	0.77/0.74

all the three measures conflate to the same value. Moreover, the correlation with AP is always 1 when $g_1 = 1.00$ and the others are zero, since this corresponds exactly to the "lenient" strategy for mapping to binary relevance, when all these measures are equal, confirming that GAP, xGAP, and eGAP actually extend AP to graded relevance.

As a general behaviour, you can note that as g_1 increases from zero towards one (and thus the other g_i decrease correspondingly) the correlation between AP, xGAP, and eGAP increases. This is a consequence of the fact that increasing g_1 moves measures more and more toward the "lenient" mapping to binary relevance adopted for computing AP. For example, in TREC 10, moving from $(g_1, g_2) = (0.1, 0.9)$ and to $(g_1, g_2) = (0.3, 0.7)$ increases the correlations $\tau_{\text{AP,xGAP}} = 0.73$ and $\tau_{\text{AP,eGAP}} = 0.72$ to $\tau_{\text{AP,xGAP}} = 0.80$ and $\tau_{\text{AP,eGAP}} = 0.78$; similarly, in TREC 21, moving from $(g_1, g_2, g_3) = (0.2, 0.2, 0.6)$ to $(g_1, g_2, g_3) = (0.4, 0.2, 0.4)$ increases the correlations $\tau_{\text{AP,xGAP}} = 0.62$ and $\tau_{\text{AP,eGAP}} = 0.61$ to $\tau_{\text{AP,xGAP}} = 0.83$ and $\tau_{\text{AP,eGAP}} = 0.82$. In a similar fashion, as g_1 increases, also the correlation between GAP xGAP and eGAP increases, as an effect of the flattening towards a "lenient" mapping to binary relevance.

Table 3. Kendall's tau correlation analysis for TREC 21

(g_1, g_2, g_3)	TREC 21, 2012, Web				
(0.0, 0.0, 1.0)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.41/0.49/-0.15	0.25/0.36/-0.12	0.58/0.57/-0.39	0.26/0.02/-0.18
xGAP	1.00/1.00/1.00	0.41/0.49/-0.15	0.25/0.36/-0.12	0.58/0.57/-0.39	0.26/0.02/-0.18
eGAP	1.00/1.00/1.00	0.41/0.49/-0.15	0.25/0.36/-0.12	0.58/0.57/-0.39	0.26/0.02/-0.18
(0.0, 0.2, 0.8)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.49/0.53/0.13	0.19/0.35/0.09	0.81/0.59/-0.03	0.34/0.06/-0.09
xGAP	0.68/0.99/0.54	0.44/0.54/-0.04	0.22/0.36/-0.05	0.63/0.57/-0.26	0.27/0.06/-0.24
eGAP	0.65/0.99/0.51	0.44/0.54/-0.05	0.23/0.36/-0.02	0.61/0.58/-0.25	0.28/0.07/-0.23
(0.0, 0.4, 0.6)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.52/0.52/0.15	0.14/0.34/0.12	0.82/0.61/-0.07	0.34/0.04/-0.07
xGAP	0.79/0.99/0.67	0.53/0.53/-0.01	0.28/0.35/0.00	0.78/0.60/-0.23	0.36/0.06/-0.19
eGAP	0.77/0.98/0.75	0.51/0.54/0.02	0.28/0.36/0.01	0.76/0.59/-0.22	0.35/0.06/-0.18
(0.0, 0.6, 0.4)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.51/0.48/0.15	0.13/0.31/0.12	0.81/0.58/-0.07	0.33/0.01/-0.07
xGAP	0.89/0.98/0.74	0.51/0.51/0.03	0.20/0.33/0.00	0.83/0.58/-0.21	0.34/0.03/-0.17
eGAP	0.85/0.96/0.79	0.47/0.53/0.06	0.20/0.35/0.03	0.85/0.58/-0.20	0.31/0.05/-0.16
(0.0, 0.8, 0.2)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.52/0.46/0.14	0.13/0.28/0.13	0.81/0.52/-0.04	0.33/-0.01/-0.06
xGAP	0.98/0.97/0.85	0.54/0.45/0.07	0.15/0.27/0.04	0.83/0.53/-0.15	0.33/-0.02/-0.13
eGAP	0.97/0.96/0.89	0.52/0.44/0.09	0.14/0.26/0.08	0.82/0.54/-0.13	0.32/-0.03/-0.11
(0.0, 1.0, 0.0)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.49/0.48/0.15	0.11/0.28/0.14	0.79/0.52/-0.05	0.31/-0.01/-0.05
xGAP	1.00/1.00/1.00	0.49/0.48/0.15	0.11/0.28/0.14	0.79/0.52/-0.05	0.31/-0.01/-0.05
eGAP	1.00/1.00/1.00	0.49/0.48/0.15	0.11/0.28/0.14	0.79/0.52/-0.05	0.31/-0.01/-0.05
(0.2, 0.0, 0.8)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.96/0.89/0.98	0.59/0.63/0.80	0.41/0.57/0.61	0.75/0.36/0.69
xGAP	0.60/0.79/0.28	0.58/0.71/0.28	0.37/0.53/0.27	0.64/0.59/-0.02	0.43/0.15/0.08
eGAP	0.59/0.73/0.23	0.57/0.66/0.23	0.37/0.46/0.22	0.63/0.61/-0.05	0.40/0.08/0.03
(0.2, 0.2, 0.6)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.94/0.89/0.94	0.57/0.61/0.74	0.43/0.59/0.67	0.73/0.34/0.63
xGAP	0.64/0.79/0.27	0.62/0.72/0.34	0.38/0.54/0.33	0.68/0.58/0.05	0.45/0.16/0.14
eGAP	0.63/0.75/0.24	0.61/0.67/0.31	0.35/0.47/0.29	0.71/0.62/-0.02	0.44/0.11/0.11
(0.2, 0.4, 0.4)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.91/0.87/0.93	0.54/0.61/0.75	0.46/0.61/0.66	0.69/0.29/0.62
xGAP	0.68/0.83/0.34	0.61/0.73/0.40	0.31/0.55/0.35	0.72/0.59/0.10	0.44/0.17/0.18
eGAP	0.72/0.78/0.22	0.64/0.65/0.27	0.29/0.47/0.24	0.73/0.62/-0.03	0.47/0.16/0.05
(0.2, 0.6, 0.2)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.90/0.87/0.92	0.51/0.61/0.76	0.49/0.61/0.65	0.69/0.29/0.63
xGAP	0.66/0.87/0.33	0.56/0.77/0.37	0.23/0.59/0.34	0.77/0.55/0.11	0.37/0.21/0.17
eGAP	0.73/0.77/0.32	0.63/0.65/0.38	0.26/0.47/0.33	0.74/0.57/0.07	0.44/0.16/0.16
(0.2, 0.8, 0.0)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.89/0.86/0.95	0.51/0.60/0.79	0.49/0.60/0.62	0.68/0.28/0.66
xGAP	0.72/0.84/0.35	0.61/0.71/0.36	0.22/0.53/0.33	0.78/0.51/0.09	0.40/0.15/0.16
eGAP	0.73/0.69/0.34	0.62/0.58/0.35	0.23/0.38/0.32	0.77/0.51/0.11	0.41/0.08/0.15
(0.4, 0.0, 0.6)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.97/0.93/0.99	0.60/0.64/0.79	0.40/0.56/0.62	0.76/0.37/0.68
xGAP	0.79/0.87/0.59	0.78/0.82/0.60	0.58/0.64/0.55	0.46/0.54/0.27	0.63/0.26/0.38
eGAP	0.80/0.82/0.46	0.79/0.77/0.47	0.53/0.57/0.42	0.53/0.61/0.17	0.62/0.19/0.27
(0.4, 0.2, 0.4)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.97/0.94/0.96	0.60/0.63/0.76	0.40/0.57/0.65	0.76/0.36/0.65
xGAP	0.82/0.88/0.56	0.83/0.82/0.61	0.53/0.64/0.55	0.52/0.54/0.27	0.66/0.26/0.39
eGAP	0.81/0.85/0.52	0.82/0.79/0.56	0.47/0.59/0.50	0.57/0.60/0.24	0.65/0.21/0.36
(0.4, 0.4, 0.2)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.97/0.93/0.96	0.58/0.63/0.76	0.42/0.58/0.65	0.76/0.35/0.65
xGAP	0.81/0.89/0.58	0.80/0.83/0.62	0.47/0.65/0.55	0.55/0.53/0.32	0.61/0.27/0.38
eGAP	0.81/0.87/0.61	0.78/0.80/0.65	0.41/0.60/0.56	0.61/0.58/0.31	0.59/0.22/0.43
(0.4, 0.6, 0.0)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.97/0.93/0.96	0.58/0.62/0.76	0.42/0.58/0.65	0.76/0.35/0.65
xGAP	0.76/0.89/0.63	0.73/0.82/0.65	0.34/0.62/0.58	0.66/0.54/0.33	0.52/0.24/0.41
eGAP	0.81/0.82/0.54	0.78/0.77/0.56	0.39/0.55/0.48	0.61/0.55/0.29	0.57/0.17/0.36
(0.6, 0.0, 0.4)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.97/0.97/1.00	0.60/0.62/0.80	0.40/0.56/0.61	0.76/0.37/0.69
xGAP	0.87/0.94/0.81	0.88/0.91/0.81	0.60/0.66/0.76	0.44/0.56/0.44	0.74/0.35/0.59
eGAP	0.89/0.92/0.82	0.91/0.88/0.82	0.56/0.62/0.75	0.48/0.60/0.45	0.74/0.31/0.62
(0.6, 0.2, 0.2)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.98/0.96/0.98	0.59/0.61/0.78	0.41/0.57/0.63	0.77/0.36/0.67
xGAP	0.89/0.88/0.82	0.89/0.86/0.84	0.55/0.68/0.79	0.47/0.54/0.45	0.71/0.31/0.62
eGAP	0.91/0.91/0.84	0.91/0.86/0.86	0.56/0.60/0.77	0.48/0.62/0.47	0.74/0.28/0.64
(0.6, 0.4, 0.0)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.98/0.96/0.97	0.59/0.61/0.77	0.41/0.57/0.65	0.77/0.36/0.66
xGAP	0.91/0.88/0.80	0.91/0.86/0.83	0.52/0.66/0.74	0.48/0.54/0.48	0.69/0.28/0.59
eGAP	0.89/0.87/0.79	0.87/0.85/0.80	0.48/0.59/0.68	0.52/0.61/0.47	0.66/0.25/0.58
(0.8, 0.0, 0.2)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.99/0.99/1.00	0.60/0.60/0.80	0.40/0.56/0.61	0.78/0.37/0.69
xGAP	0.97/0.94/0.93	0.96/0.93/0.93	0.61/0.66/0.83	0.41/0.54/0.54	0.79/0.35/0.66
eGAP	0.98/0.93/0.96	0.97/0.92/0.96	0.62/0.63/0.82	0.42/0.59/0.56	0.80/0.34/0.69
(0.8, 0.2, 0.0)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	0.99/0.98/0.99	0.60/0.60/0.79	0.40/0.56/0.62	0.78/0.35/0.68
xGAP	0.97/0.91/0.94	0.96/0.91/0.95	0.57/0.64/0.81	0.43/0.56/0.56	0.75/0.33/0.66
eGAP	0.95/0.91/0.96	0.94/0.91/0.97	0.55/0.58/0.81	0.45/0.62/0.60	0.73/0.28/0.68
(1.0, 0.0, 0.0)	**GAP**	**AP**	**nDCG**	**RBP**	**bpref**
GAP	1.00/1.00/1.00	1.00/1.00/1.00	0.61/0.59/0.80	0.41/0.57/0.61	0.79/0.36/0.69
xGAP	1.00/1.00/1.00	1.00/1.00/1.00	0.61/0.59/0.80	0.41/0.57/0.61	0.79/0.36/0.69
eGAP	1.00/1.00/1.00	1.00/1.00/1.00	0.61/0.59/0.80	0.41/0.57/0.61	0.79/0.36/0.69

The biases introduced in GAP and discussed in Section 3 become evident by looking at the correlation between GAP and AP. As soon as some weight is provided on g_1, the correlation between GAP and AP suddenly becomes quite high, even if it should not, since a low g_1 corresponds to a "hard" mapping strategy to binary relevance (all but highly relevant/key documents are considered not relevant) which is the opposite from the "lenient" one adopted for computing AP. For example, with $(g_1, g_2) = (0.1, 0.9)$ we have $\tau_{\text{AP,GAP}} = 0.84$ in TREC 10 and $\tau_{\text{AP,GAP}} = 0.94$ in TREC 14, which are already extremely high; while with $(g_1, g_2, g_3) = (0.2, 0.2, 0.6)$ we have $\tau_{\text{AP,GAP}} = 0.94$ in TREC 21. This indicates that GAP tends to overestimate the weight of g_1 and to saturate the ranking. On the other hand, for the same parameters, we have $\tau_{\text{AP,xGAP}} = 0.73$ and $\tau_{\text{AP,eGAP}} = 0.72$ in TREC 10, $\tau_{\text{AP,xGAP}} = 0.83$ and $\tau_{\text{AP,eGAP}} = 0.82$ in TREC 14, and $\tau_{\text{AP,xGAP}} = 0.62$ and $\tau_{\text{AP,eGAP}} = 0.61$ in TREC 21, which indicate how the weights g_i assigned by the user are more correctly taken into account.

This effect is even more exacerbated when you consider the topics with few highly relevant / key documents. For example, with $(g_1, g_2) = (0.1, 0.9)$, in TREC 10 the correlation $\tau_{\text{AP,GAP}} = 0.84$ on all topics is quite similar to the correlation $\tau_{\text{AP,GAP}} = 0.80$ on topics with few highly relevant documents, indicating a lack of sensitivity of GAP to this important case and its flattening on AP. On the other hand, the correlations for xGAP and eGAP fall from $\tau_{\text{AP,xGAP}} = 0.73$ and $\tau_{\text{AP,eGAP}} = 0.72$ to $\tau_{\text{AP,xGAP}} = 0.48$ and $\tau_{\text{AP,eGAP}} = 0.45$, indicating that they treat the case when the user attributes more weight (g_2 high) to the few high relevant documents quite differently from AP, which flattens out everything with a "lenient" mapping to binary relevance. Similar behaviors can be observed also in the case of TREC 14 and TREC 21.

The correlation with nDCG, the only other graded measure, increases as the value of g_1 increases, i.e. the more you move away from an "hard" strategy for mapping to binary relevance. Moreover, in the case of topics with few highly relevant / key documents and with a low g_1, the correlation between GAP and nDCG is always higher than the one between xGAP/eGAP and nDCG, indicating that both GAP and nDCG are less sensitive to this case.

Robustness to Incomplete Judgments. The *stratified random sampling* of the pools allows us to investigate the behavior of the measures as relevance judgment sets become less complete following the methodology presented in [1], which is here adapted to the case of multi-graded relevance.

The plots in Figure 1 and 2 show the Kendall's tau correlations between the system rankings produced using progressively down-sampled pools from 100% (complete pool) to 5%. Each line shows the behavior of a measure; the flatter (and closer to 1.0) the line, the more robust the measure. In fact, a flat line indicates that the measure continues to rank systems in the same relative order with different levels of relevance judgments incompleteness. In this respect, nDCG and bpref exhibit the best behaviour.

As an example of the main case of interest in the paper (g_1 low), when all the topics are considered (Figure 1, on the left), xGAP and eGAP behave similarly to GAP for TREC 10 and 14, even if they improve for quite shallow pools (10%

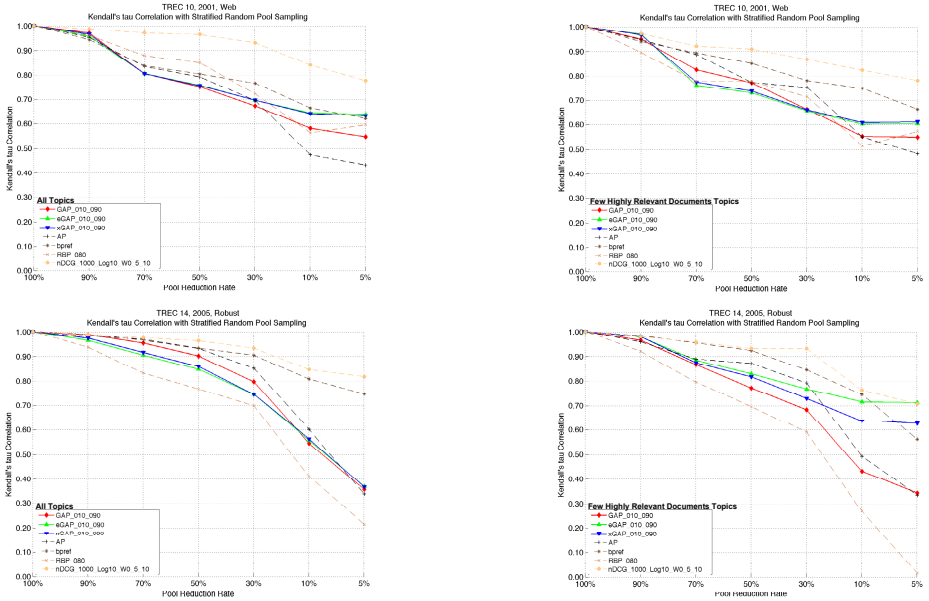

Fig. 1. Kendall's rank correlation at pool reduction rates on TREC 10 (top row) and TREC 14 (bottom row) for all topics (left) and topics with few highly relevant documents (right). GAP, xGAP, and eGAP with $(g_1, g_2) = (0.1, 0.9)$.

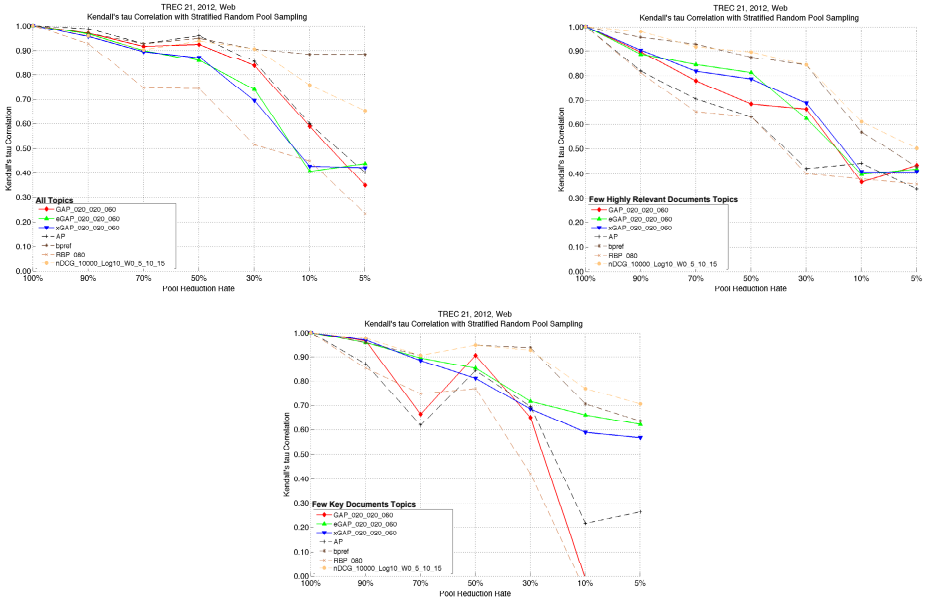

Fig. 2. Kendall's rank correlation at pool reduction rates on TREC 21 for all topics (top left), topics with few highly relevant documents (top right), and topics with few key documents (bottom center). GAP, xGAP, and eGAP with $(g_1, g_2, g_3) = (0.2, 0.2, 0.6)$.

and 5% reduction rates), which are important for avoiding costly assessment. This behaviour is even more evident when it comes to topics with few highly relevant documents (Figure 1, on the right). In the case of TREC 21, GAP exhibits better properties than xGAP and eGAP when all topics are considered (Figure 2, top left), even it almost follows the behaviour of AP, thus indicating again its tendence to overestimate the weight of g_1. However, xGAP and eGAP improve with respect to GAP when topics with few highly relevant documents (Figure 2, top right) and topics with few key documents (Figure 2, bottom center) are considered; in this latter case, it can be noted how GAP become unstable for quite shallow pools (10% and 5% reduction rates).

7 Conclusions and Future Work

In this paper we have introduced the xGAP and eGAP measures which extend GAP and are able to further push the focus on the user perception of relevance. We have shown how they take a different angle from GAP addressing its biases and how they are robust to incomplete judgements.

Future work will consist of a more extensive evaluation on different experimental collections, taking into account also the possibility of using xGAP and eGAP as objective metric for learning to rank algorithms, as well as exploring their discriminative power.

References

1. Buckley, C., Voorhees, E.M.: Retrieval Evaluation with Incomplete Information. In: SIGIR 2007, pp. 25–32. ACM Press, USA (2004)
2. Buckley, C., Voorhees, E.M.: Retrieval System Evaluation. In: TREC. Experiment and Evaluation in Information Retrieval, pp. 53–78. MIT Press, USA (2005)
3. Clarke, C.L.A., Craswell, N., Voorhees, H.: Overview of the TREC 2012 Web Track. In: TREC 2012, pp. 1–8. NIST, Special Publication 500-298, USA (2013)
4. Hawking, D., Craswell, N.: Overview of the TREC-2001 Web Track. In TREC 2001, pp. 61–67. NIST, Special Publication 500-250, USA (2001)
5. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) 20(4), 422–446 (2002)
6. Kendall, M.G.: Rank correlation methods. Griffin, Oxford (1948)
7. Moffat, A., Zobel, J.: Rank-biased Precision for Measurement of Retrieval Effectiveness. ACM Transactions on Information Systems (TOIS) 27(1), 2:1–2:27 (2008)
8. Resnick, S.I.: A Probability Path. Birkhäuser, Boston (2005)
9. Robertson, S.E., Kanoulas, E., Yilmaz, E.: Extending Average Precision to Graded Relevance Judgments. In: SIGIR 2010, pp. 603–610. ACM Press, USA (2010)
10. Voorhees, E.: Evaluation by Highly Relevant Documents. In: SIGIR 2001, pp. 74–82. ACM Press, USA (2001)
11. Voorhees, E.M.: Overview of the TREC 2005 Robust Retrieval Track. In: TREC 2005. NIST, Special Pubblication 500-266, USA (2005)