ISSN 1346-5597

National Institute of Informatics

NII Technical Report

Property of Average Precision and its Generalization: An Examination of Evaluation Indicator for Information Retrieval Experiments

Kazuaki Kishida

NII-2005-014E
Oct. 2005

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Property of Average Precision and its Generalization: An Examination of Evaluation Indicator for Information Retrieval Experiments

Kazuaki Kishida
Faculty of Cultural Information Resources, Surugadai University /
National Institute of Informatics

Abstract

In information retrieval experiments, indicators for measuring effectiveness of the systems or methods are important. Average precision is often used as an indicator for evaluating ranked output of documents in standard retrieval experiments. This report examines some properties of this indicator. First, we clarify mathematically that relevant documents at a higher position in the ranked list contribute much more to increasing the score of average precision. Second, influence of detecting unknown relevant documents on the score is discussed. Third, we examine statistical variation of average precision scores caused by fluctuations of results from relevance judgments. Another issue of this report is to explore evaluation indicators using data of multi-grade relevance. After reviewing indicators proposed by other researchers, i.e., modified sliding ratio, normalized discounted cumulative gain (nDCG), Q-measure and so on, a new indicator, generalized average precision is developed. We compare these indicators empirically using a simple and artificial example.

1 Introduction

Laboratory experiments using test collections have played an important role in developing information retrieval theories and techniques until now. In the experiments, we have to evaluate correctly search results generated by retrieval systems. This is not an easy task. Traditionally, precision and recall ratio have been used as evaluation indicators for measuring the performance of Boolean searches. The meaning of these indicators is clear, i.e., the precision is the proportion of relevant documents that are retrieved, and recall is the proportion of retrieved documents that are relevant [1]. In contrast, it is more complicated to evaluate a list of documents ranked in decreasing order of relevance that the system estimates. In standard experiments at present, average precision is widely used for comparing the performance between retrieval techniques or systems. Unfortunately, computation of average precision is more intricate than that of traditional precision or recall (see section 2), and it is difficult to correctly understand what the score means in retrieval experiments. If the score of average precision is not correctly interpreted, the experiment becomes futile. Thus, we need more examinations on the properties of average precision. In particular, it is important to investigate average precision in terms of reliability as a measure to be used in scientific experiments.

Another point to consider is that the average precision is based on binary judgments of relevance (i.e., relevant or not) as well as the traditional indicators. This means that these indicators cannot distinguish between a highly relevant document and a normally relevant one even though there is difference in the degree of relevance between them because both the documents must be assumed to be equally relevant in order to compute these indicators. Consequently, a system that places highly relevant documents in higher order may not be evaluated sufficiently highly through the average precision. Therefore, it is worth developing an alternative indicator that enables us to take multiple degrees of relevance into consideration for assessing search results.

This report examines some properties of average precision, and explores an evaluation indicator based on multi-grade relevance judgments. First, in section 2, we discuss mainly the sensitivity of average precision when the ranking of a relevant document changes. Section 3 examines the reliability of average precision, focusing on problems of unknown relevant documents and variation of results in relevance judgments. In section 4, evaluation indicators directly utilizing multiple degrees of relevance are investigated. Finally, some concluding remarks are given. Note that the section 2 and 3 are partly based on discussion in Kishida[2].

2 Property of average precision

2.1 Formal definition of evaluation indicators

Suppose that we wish to evaluate a retrieval system by using a test collection, which consists of a set of documents, a set of search requests (topics), and results of relevance judgments. Let x_k be a variable representing the degree of relevance of the kth document in a ranked list that a system generates for a given topic. In this section, we assume binary relevance judgments, i.e.,

$$x_k = \begin{cases} 1 & \text{if } k\text{th document is relevant} \\ 0 & \text{if } k\text{th document is irrelevant.} \end{cases} \tag{1}$$

If we take the top-ranked m documents, the value of precision in the document set is computed such that

$$p_m = \frac{1}{m}\sum_{k=1}^{m} x_k. \tag{2}$$

It should be noted that p_m can be interpreted as an average of values, $x_1, ..., x_m$. We denote the average by $\bar{x}_m$.

Average precision is defined as "the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved [1]," which can be mathematically expressed by using p_m or $\bar{x}_m$ such that

$$v = \frac{1}{R}\sum_{i=1}^{n} I(x_i)p_i = \frac{1}{R}\sum_{i=1}^{n} I(x_i)\bar{x}_i \tag{3}$$

where R is the total number of relevant documents and n is the number of documents included in the list (usually $n = 1000$). $I(x_i)$ is a function such that

$$I(x_i) = \begin{cases} 0 & \text{if } x_i = 0 \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

In the case where x_i is a binary variable, we can simply set $I(x_i) = x_i$. Thus, the average precision can be represented as

$$v = \frac{1}{R} \sum_{i=1}^{n} x_i p_i = \frac{1}{R} \sum_{i=1}^{n} \frac{x_i}{i} \sum_{k=1}^{i} x_k. \tag{5}$$

Another indicator, R-precision, is sometimes used in retrieval experiments. It is defined as "precision after R documents have been retrieved where R is the number of relevant documents for the current topic [1]", which can be expressed using our notation as

$$r = \frac{1}{R} \sum_{i=1}^{R} x_i. \tag{6}$$

There are also other classical indicators for assessing ranked output, which were proposed in the 1960s. The sum of ranking of R relevant documents can be represented as $\sum_{i=1}^{N} iI(x_i)$ where N is the total number of documents included in the database. These relevant documents should be ranked from 1st to Rth positions in an "ideal" ranked list where the sum of ranking of relevant documents amounts to $\sum_{i=1}^{R} i$. Normalized recall (Salton and Lesk [3]) is based on a difference of the two sums of ranking, and its formal definition is

$$z_r = 1 - \frac{\sum_{i=1}^{N} iI(x_i) - \sum_{i=1}^{R} i}{R(N-R)}. \tag{7}$$

The maximum value of the difference of two sums amounts to $R(N-R)$ because the difference is calculated such that $(N-R+1) + (N-R+2) + \cdots + N - (1 + 2 + \cdots + R) = R \times (N-R)$. Therefore, $0.0 \leq z_r \leq 1.0$ where 1.0 indicates the best ranking.

When we convert the index i representing each position of documents into $\log i$, the indicator is called "normalized precision," which is defined mathematically as

$$z_p = 1 - \frac{\sum_{i=1}^{N} I(x_i) \log i - \sum_{i=1}^{R} \log i}{\log {}_N C_R}. \tag{8}$$

where $\log {}_N C_R = \log(N-R+1) + \log(N-R+2) + \cdots + \log N - (\log 1 + \log 2 + \cdots + \log R)$.

Unfortunately, normalized recall and precision have been seldom used in recent retrieval experiments, perhaps because all N documents have to be ranked and checked for computing scores of these indicators, and this requirement is not practical.

It should be noted that there have been other efforts for evaluating ranked output such as "search length" by Cooper [4] and "measure E" by Swets [5].

2.2 Sensitivity of evaluation indicators

Suppose that an irrelevant document placed at the jth position in a ranked list is turned to be relevant (i.e., changing from $x_j = 0$ to $x_j = 1$). Let $\Delta v(j)$ be an amount of increment in average precision score with the change of the jth document from irrelevant to relevant. From Eq.(5), $\Delta v(j)$ can be computed as

$$\Delta v(j) = \frac{1}{R}\left(p_j + \sum_{t=j+1}^{n} \Delta p_t\right), \tag{9}$$

where Δp_t indicates an increment of p_t with this change ($t = j+1, ..., n$). Note that the change at the jth position not only adds a new value of precision p_j but also affects each p_t in lower positions below j. Since $\Delta p_t = t^{-1}x_t$, we obtain

$$\Delta v(j) = \frac{1}{R}\left(j^{-1}\sum_{k=1}^{j} x_k + \sum_{t=j+1}^{n} t^{-1}x_t\right). \tag{10}$$

Theoretically, Eq.(10) has a minimum when $j = N$ where N is the total number of documents in the database, and a maximum when $j = 1$. If R is assumed to be invariant with the change of jth document, $\sum_{k=1}^{N} x_k = R$. Hence, we obtain from Eq.(10) that $\Delta v(N) = 1/N$, which is usually very small in standard retrieval experiments. In contrast, the maximum value of Eq.(10) depends on each value of p_t.

We can easily conclude that

$$\Delta v(j) > \Delta v(j+1), \tag{11}$$

where the $(j+1)$th document is assumed to be irrelevant at the time of computing $\Delta v(j)$, and similarly, the jth document is assumed to be irrelevant when we consider $\Delta v(j+1)$. Under these assumptions, it is obvious that $p_j > p_{j+1}$ and

$$\sum_{t=j+1}^{n} \Delta p_t = \sum_{t=j+2}^{n} t^{-1}x_t = \sum_{l=j+2}^{n} \Delta p_l. \tag{12}$$

Therefore, we can obtain Eq.(11). This equation implies that a relevant document ranked at a higher position helps to increase the average precision score much more than a lower-ranked relevant document. This is a remarkable characteristic of average precision in comparison with R-precision. In the case of R-precision, an increment with the change is simply expressed by $\Delta r(j) = R^{-1}$, which is independent of the position of the document.

3 Reliability of evaluation by indicators

3.1 Sources of statistical variations

When evaluation indicators are employed in a standard situation of retrieval experiments, we have to consider two kinds of statistical errors: (1) sampling

errors and (2) non-sampling errors. For assessing the magnitude of sampling errors, a standard statistical theory can be applied. Let v_h be an average precision score for the hth search request (or topic). If we have L topics in an experiment, a sample mean of average precision scores $v_1, ..., v_L$, i.e.,

$$\bar{v} = L^{-1} \sum_{h=1}^{L} v_h, \tag{13}$$

can be used as an estimator of the population mean. The sample mean $\bar{v}$ is often called "mean average precision (MAP)." Similarly, a sample variance is computed as

$$s_v^2 = (L-1)^{-1} \sum_{h=1}^{L} (v_h - \bar{v})^2, \tag{14}$$

under the assumption that simple random sampling was carried out. The variance can be used for statistical tests.

On the other hand, in the case of retrieval experiments, there are two sources of non-sampling errors, i.e., (1) undiscovered relevant documents by adopting the so-called pooling method and (2) variations of results in the process of relevance judgments, which will be discussed below.

3.2 Influence of undiscovered relevant documents

Suppose that we select the top 100 documents from every search result submitted by research groups participating in a project such as TREC, NTCIR or CLEF, and create a pooled set of documents by merging them and removing duplications. A set of relevant documents is usually identified by assessing manually the pooled document set. Therefore, if a relevant document is ranked below the 100th position in all search results, it remains undiscovered and is supposed to be treated as an irrelevant document in the process of computing average precision scores since usually we set $n = 1000$ in Eq.(5).

We shall examine the change of average precision score when an undiscovered relevant document at the jth position happens to be newly detected. First, note that the total number of relevant documents increases to $R + 1$. Second, an increment with detection of a relevant document at the jth position can be computed by Eq.(10). Finally, we have to consider a decrease of original score v with increase of the number of total relevant documents, which can be evaluated as

$$v - Rv/(R+1) = [1 - R/(R+1)]v = (R+1)^{-1} v. \tag{15}$$

Therefore, the change of average precision score upon detecting an undiscovered relevant document at the jth position is computed as

$$\tilde{\Delta} v(j) = \frac{1}{R+1} \left(\frac{1}{j} \sum_{k=1}^{j} x_k + \sum_{t=j+1}^{n} \frac{x_t}{t} \right) - \frac{1}{R+1} v, \tag{16}$$

Table 1. Change of average precision score after finding a new relevant document at rank 101

	average precision: v		
R	0.1	0.3	0.5
10	0.00081	-0.01737	-0.03555
50	0.00794	0.00402	0.00010
100	0.00891	0.00693	0.00495

where v is an original score of average precision before the relevant document is detected. If we suppose that there is no relevant document below the jth position, we obtain $\sum_{k=1}^{j} x_k = R + 1$ and $\sum_{t=j+1}^{n} t^{-1} x_t = 0$. Thus, Eq.(16) becomes simpler, i.e.,

$$\tilde{\Delta} v(j) = j^{-1} - (R+1)^{-1} v. \tag{17}$$

Table 1 shows examples of values in Eq.(17) where $j = 101$. The amount of change is not so large. This can be explained by the fact that relevant documents in a lower position do not have a large effect on the average precision score as indicated in Eq.(10) and Eq.(11). As shown in Table 1, there is a possibility that the average precision score decreases inversely upon finding an additional undiscovered relevant document. It turns out from Eq.(17) that $\tilde{\Delta} v(j)$ becomes negative when $vj > R + 1$, which can be easily obtained by transforming the inequality $j^{-1} - (R+1)^{-1} v < 0$.

3.3 Influence of variation in relevance judgments

In order to compute an average precision score, we need to examine whether documents are relevant or not. Human assessors usually make relevance judgments based on the "topicality" of written search requests in standard retrieval experiments such as TREC, NTCIR or CLEF. If the value of x_i can be determined by the process of relevance judgments for a search topic, an average precision score can be calculated by Eq.(5) for each search run.

However, such judgment is subjective, and it can be reasonably assumed according to discussions on subjectivity of relevance (see Schamber [6]) that the results of judgments will vary depending on human assessors or situations in which the judgments are made. Thus, we need to investigate the influence of the variation on scores of average precision using a theory of statistical survey.

Suppose that the results from many independent repetitions of relevance judgments for each search topic are available, e.g., 10 or more different assessors separately made judgments for the same topic. In this case, according to a textbook of statistical sampling [7], we can introduce the model

$$v_{ha} = \mu_h + e_{ha}, \tag{18}$$

where v_{ha} is the average precision calculated by using the result from the ath repetition of relevance judgments for the hth topic ($a = 1, 2, ...$). If there is no

variation in the process of relevance judgments, all v_{ha} are inevitably equivalent (i.e., $v_{h1} = v_{h2} = ...$). However, when the judgments fluctuate, v_{ha} is conceptually broken into μ_h and e_{ha}, where μ_h is the true value of v_{ha} and e_{ha} is a kind of "error of measurement" (if there is no variation, e_{ha} is always zero).

Let $E_m(\cdot|h)$ be an expectation of any variable over repeated judgments ($a = 1, 2, ...$) for the hth topic. Since μ_h is independent of judgment fluctuations, we obtain

$$E_m(v_{ha}|h) = \mu_h + E_m(e_{ha}|h). \tag{19}$$

If e_{ha} follows a frequency distribution with mean zero, this equation becomes $E_m(v_{ha}|h) = \mu_h$, i.e., the error term e_{ha} is cancelled out by an average operation $E_m(\cdot|h)$. When $E_m(e_{ha}|h)$ cannot be assumed to be zero, the mean $E_m(v_{ha}|h)$ includes a statistical bias, and similarly the MAP score computed from $E_m(v_{ha}|h)$ is influenced by the bias. For simplicity, we assume that $E_m(e_{ha}|h) = 0$ (i.e., $E_m(v_{ha}|h) = \mu_h$). As for the variance of v_{ha}, we can compute it as

$$\begin{aligned} V_m(v_{ha}|h) &= E_m[(v_{ha} - \mu_h)^2|h] = E_m[(\mu_h + e_{ha} - \mu_h)^2|h] \\ &= E_m(e_{ha}^2|h) = V_m(e_{ha}|h), \end{aligned} \tag{20}$$

by assuming that $E_m(e_{ha}|h) = 0$. We denote the variance of e_{ha} by σ_h^2, i.e., $V_m(v_{ha}|h) = \sigma_h^2$.

When a particular sample S consisting of L topics is given, a score of MAP at the ath judgment can be calculated as $\bar{v}_a = L^{-1} \sum_{h=1}^{L} v_{ha}$. Then, the mean of $\bar{v}_a$ taken over repeated judgments ($a = 1, 2, ...$) is

$$E_m(\bar{v}_a|S) = E_m\left(\frac{1}{L} \sum_{h=1}^{L} v_{ha} \middle| S \right) = \frac{1}{L} \sum_{h=1}^{L} E_m(v_{ha}|h) = \frac{1}{L} \sum_{h=1}^{L} \mu_h \tag{21}$$

under the assumption that judgments between topics are independent.

Furthermore, since the result shown in Eq.(21) was obtained for a particular sample S, we must consider an expectation over all possible samples, and denote it by $E_p(\cdot)$. Using Eq.(21), we obtain

$$E_p[E_m(\bar{v}_a|S)] = E_p\left(\frac{1}{L} \sum_{h=1}^{L} \mu_h \right) = \frac{1}{L} \sum_{h=1}^{L} E_p(\mu_h). \tag{22}$$

According to elementary statistical theory, $E_p(\mu_h) = \mu$ where μ is the population mean of μ_h. Therefore, finally we obtain

$$E_p[E_m(\bar{v}_a|S)] = \mu, \tag{23}$$

which means that a MAP score including variations of relevance judgments is an unbiased estimator under the assumption that $E_m(e_{ha}|h) = 0$.

Similarly, the variance of $\bar{v}_a$ can be computed as

$$V_p[V_m(\bar{v}_a|S)] = \frac{1}{L}(\sigma_d^2 + \sigma_\mu^2) \tag{24}$$

where σ_μ^2 is the population variance of μ and

$$\sigma_d^2 = E_p\left(\frac{1}{L}\sum_{h=1}^{L}\sigma_h^2\right) \tag{25}$$

(see appendix for details). σ_d^2 is equal to the population mean of the variance of frequency distribution that the fluctuation of average precision with judgments (e_{ha}) follows.

If e_{ha} in Eq.(18) is caused by only careless mistakes in relevance judgments, we may be able to keep the variations small. Needless to say, as the sample variance is smaller, the estimated population mean of average precision is more reliable. As already mentioned, relevance judgment is essentially subjective and we may have to consider e_{ha} as an indispensable factor. In this case, we should try to reduce the standard error by using a large sample (i.e., a large number of topics) in order to obtain sufficient confidence in the statistical test.

For knowing more about the influence of judgment variation, we shall try a very simple simulation. Suppose that there are 50 topics and each has a search result consisting of 10 ranked documents respectively (i.e., $L = 50$ and $N = 10$). We also assume subjectivity of relevance, and introduce a relevance probability π_{hi} of the ith document for the hth topic. The probability is empirically interpreted as the ratio of judgments as relevant in M repeated trials for the same document, e.g., if $\pi_{hi} = 0.9$, this means that the ith document is assessed as relevant by 9 out of 10 judges for the hth search topic. The procedure of our simulation is as follows.

(1) Setting distribution of relevance probabilities $\pi_{h1}, ..., \pi_{h10}$ for the hth topic ($h = 1, ..., L$). In order to simplify this process, a random number from 0 to 9 following a uniform distribution is generated for each document, and if the number is less than 5 then a predetermined probability θ (e.g., $\theta = 0.9$) is assigned to the document. If not, $1 - \theta$ is allocated. That is, we set $\pi_{hi} = \theta$ or $\pi_{hi} = 1 - \theta$ randomly. The procedure is applied to every document for all topics.

(2) Estimating μ_h and σ_h^2 for a given hth topic by repeating M times the following operations (2-a) and (2-b).

(2-a) For the ith document, a uniform random number from 0.0 to 1.0 is generated, and if it exceeds the probability π_{hi} then the document is assumed to be relevant. If not, the document is irrelevant.

(2-b) The above procedure (2-a) is repeated for all documents (i.e., $i = 1, ..., N$), and an average precision score is computed under the assumption that a retrieval system outputs the documents from $i = 1$ to $i = N$ sequentially.
From a set of M scores, we can compute $\hat{\mu}_h$ and $\hat{\sigma}_h^2$ (the hat mark means that the quantity is an estimation).

(3) Procedure (2) is repeated for L topics ($h = 1, ..., L$). Consequently, σ_d^2 is estimated by averaging $\hat{\sigma}_h^2 (h = 1, ..., L)$, and σ_μ^2 is calculated from $\hat{\mu}_h (h =$

$1, ..., L$).

We execute two runs of the simulation where we set $\theta = 0.9$ and $\theta = 0.5$ respectively, i.e., in one case the degree of variation is relatively small and in the other it is large ("$\theta = 0.5$" means that each document is judged as relevant or irrelevant randomly). In both cases, the above procedure (2-a) and (2-b) is repeated 1000 times (i.e., $M = 1000$). The results are shown in Table 2.

Table 2. Results of simulation

	$\theta = 0.9$	$\theta = 0.5$
$\hat{\mu}(= L^{-1} \sum_{h=1}^{L} \hat{\mu}_h)$	0.62396	0.60679
$\hat{\sigma}_{\mu}^2$	0.01660	0.00004
$\hat{\sigma}_{d}^2(= L^{-1} \sum_{h=1}^{L} \hat{\sigma}_h^2)$	0.01314	0.03674
$\hat{\sigma}_{\mu}^2 + \hat{\sigma}_{d}^2$	0.02974	0.03678
$M^{-1} \sum_{a=1}^{M} \hat{V}_p(v_{ha}\|S)$	0.03006	0.03687

In the case that $\theta = 0.9$, the value of $\hat{\sigma}_d^2$ is less than that of $\hat{\sigma}_{\mu}^2$. It should be noted that differences of average precision score between topics are so small in the simulation unlike actual situations in standard test collection due to the fact that a fixed probability θ is uniformly used in all topics for simplicity. Therefore, $\hat{\sigma}_{\mu}^2$ is expected to be larger in real settings than the values shown in Table 2. In contrast, for the case where $\theta = 0.5$, $\hat{\sigma}_d^2$ is about three times as large as in the case of 0.9, as we expected. Note that the sum of these quantities, i.e., $\hat{\sigma}_{\mu}^2 + \hat{\sigma}_{d}^2$, almost equals $M^{-1} \sum_{a=1}^{M} \hat{V}_p(v_{ha}|S)$, which is an average of $\hat{V}_p(v_{ha}|S)$ that is an estimation of the population variation when the index a is fixed. Actually, only a value of $\hat{V}_p(v_{ha}|S)$ is observed in real situations, and is used as an estimation of the population variance. Therefore, if we can reduce the size of $\hat{\sigma}_d^2$ by preventing careless mistakes in judgments, statistical accuracy will be improved.

4 Generalized average precision

4.1 Indicators based on multi-grade relevance judgments

4.1.1 Multi-grade relevance judgments and indicators

In general, average precision is computed based on binary judgments on relevance (see Eq.(1)). However, it may be more natural to assign a multiple degree of relevance to each document in the process of assessment. For example, Tang et al. [8] concluded that a seven-point scale is optimal for relevance assessments. In retrieval experiments, the NTCIR project (CLIR task [9]) is using a four-point scale (highly relevant, relevant, partially relevant and irrelevant) and a three-point scale was employed at the TREC Web Track [10].

A standard technique for calculating average precision from data obtained by multi-grade judgments is to reduce the multi-grade point into a dichotomous value. For example, in the NTCIR project, a binary measure "rigid relevance"

is particularly defined by interpreting two grades of "highly relevant" and "relevant" as relevant and other grades as irrelevant (in this project, "relaxed relevance" is also used, in which "partially relevant" is included in the relevant class). This kind of conversion allows us to compute average precision scores, but some information contained in the original data is inevitably lost. An alternative strategy is to develop another indicator directly using multi-grade scores. Indeed, some researchers have proposed such indicators:

1. sliding ratio [11]
2. ranked half life [12]
3. normalized (discounted) cumulated gain [13, 14, 15]
4. generalized, nonbinary recall and precision [16]
5. weighted average precision [17]
6. Q-measure [18]
7. average distance measure [19]

In the following sections, we will discuss each of these indicators. In this discussion, we suppose that x_i represents a multiple degree of relevance for the ith document in a ranked output unlike Eq.(1) (e.g., in the case of a four-point scale, $x_i = 0, 1, 2, 3$).

4.1.2 Sliding ratio and its modification

The sliding ratio is a classical indicator proposed by Pollack [11] in 1968. We consider two sequences of relevance judgments, x_k and y_k $(k = 1, ..., n)$, where x_k indicates a relevance degree of the kth document in a ranked output, and y_k represents a relevance degree in an "ideal" ranking. If we have just four relevant documents whose relevance degrees are 1, 2, 2 and 3 respectively, these documents should be ideally ranked such that $y_1 = 3, y_2 = 2, y_3 = 2, y_4 = 1, y_5 = 0, ...$ The sliding ratio is defined as a ratio of two values at each position, i.e., x_k/y_k[11].

We can easily apply the sliding ratio to compute an indicator for evaluating the overall ranking by using two sums of x_k and y_k $(k = 1, ..., n)$, i.e.,

$$v_S = \frac{\sum_{k=1}^{n} x_k}{\sum_{k=1}^{n} y_k}. \tag{26}$$

For example, suppose that $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = ... = x_n = 0$ in a ranked document list generated by a system A for a query. Then, the value of v_S is $6/8 = 0.75$.

However, v_S in Eq.(26) is not sufficiently sensitive to ranking of documents. For example, suppose that another system B creates a ranked output such that $x_1 = 3, x_2 = 2, x_3 = 1, x_4 = ... = x_n = 0$. Although system B clearly outperforms system A, the value of v_S is the same in both systems (i.e., $v_S = 0.75$).

In order to solve this problem, Sagara [20] has proposed the "modified sliding ratio," which is defined as

$$v_{S'} = \frac{\sum_{k=1}^{n} \frac{1}{k} x_k}{\sum_{k=1}^{n} \frac{1}{k} y_k}. \tag{27}$$

When using $v_{S'}$, the score of system B amounts to about 0.88, which is greater than that of system A (i.e., about 0.61) in the above example.

We can easily obtain the differential of $v_{S'}$ by a continuous quantity x_j, i.e.,

$$\frac{dv_{S'}}{dx_j} = \frac{1}{j} \frac{1}{\sum_{k=1}^{n} \frac{1}{k} y_k}. \tag{28}$$

Since $\sum_{k=1}^{n} (1/k) y_k$ is a constant, the main factor of this equation is only j^{-1}. Hence, like the average precision (see Eq.(10)), a document ranked at a higher position in the list contributes much more to increasing the score of the modified sliding ratio.

4.1.3 Ranked half life

Borlund and Ingwersen [12] proposed the ranked half life (RHL) indicator, which is a kind of median of positions at which relevant documents are placed. For example, if four relevant documents whose relevance scores are 0.5, 1.0, 1.0, 0.5 are placed at 1st, 3rd, 9th and 14th positions respectively, the RHL indicator amounts to 3.0 because the total relevance score is 3.0 and half of the total score (i.e., 1.5) can be obtained up to the 3rd document. As this example shows, this indicator is not so sensitive to changes of order of relevant document, i.e., if these four documents are ranked at 1st, 3rd, 100th and 1000th positions, the value of the RHL indicator remains unchanged. Note that the RHL indicator is more precisely defined as a median for grouped data (see Borlund and Ingwersen [12] for details).

4.1.4 Cumulated gain

We denote the numerator of Eq.(26) by $c(n)$, i.e.,

$$c(n) = \sum_{k=1}^{n} x_k, \tag{29}$$

which is called "cumulated gain (CG)" by Järvelin and Kekäläinen [13, 14]. They also defined the "discounted cumulated gain (DCG)" as

$$d(i) = \sum_{1 \le k < b} x_k + \sum_{b \le k \le i} \frac{1}{\log_b k} x_k, \tag{30}$$

where b is the base of the logarithm (e.g., $b = 2$). The idea under the DCG is to reduce the weights of relevant documents as their ranking decreases, which is similar to that in the modified sliding ratio. However, the reduction is not steeper because x_k is divided by the logarithm of rank.

Similarly, we can consider the DCG for y_k representing a relevance degree in an ideal ranking, i.e.,

$$d_I(i) = \sum_{1 \le k < b} y_k + \sum_{b \le k \le i} \frac{1}{\log_b k} y_k. \tag{31}$$

Thus, an average of "normalized discounted cumulated gain (nDCG)" up to the position n is defined as

$$v_D = \frac{1}{n} \sum_{i=1}^{n} \frac{d(i)}{d_I(i)}. \tag{32}$$

If $j \ge b$, the differential of v_D by x_j is

$$\frac{dv_D}{dx_j} = \frac{1}{n} \sum_{k \ge j}^{n} \frac{1}{\log_b j} \frac{1}{d_I(k)}, \tag{33}$$

since $d(i)' = 1/\log_b j$. This equation contains a sort of inverse of rank, $1/\log_b j$, which is similar to the average precision or the modified sliding ratio. However, the differential of v_D also includes $d_I(k)$, which complicates the behavior of this function. Intuitively, as the rank j becomes smaller, the increment becomes greater, because the summation in Eq.(33) is computed for positions over j.

Kekäläinen and Järvelin [16] also proposed "generalized, nonbinary recall and precision." For computing these indicators, we need to normalize each value of x_k into a real number ranging from 0.0 to 1.0. An easy way is to divide x_k by the maximum value in its definition, i.e., $z_k = x_k / \max_k x_k$. The generalized precision g_P and recall g_R can be computed as

$$g_P = \frac{1}{n} \sum_{k=1}^{n} z_k \qquad \text{and} \qquad g_R = \frac{\sum_{k=1}^{n} z_k}{\sum_{k=1}^{N} z_k},$$

where N indicates the total number of documents in the database the same as before. Note that the maximum value of the generalized precision is not necessarily 1.0.

4.1.5 Weighted average precision and Q-measure

A variation of the sliding ratio has been proposed by Kando et al. [17], i.e.,

$$v_W = \frac{1}{R} \sum_{i=1}^{n} I(x_i) \frac{\sum_{k=1}^{i} x_k}{\sum_{k=1}^{i} y_k}, \tag{34}$$

which is called "weighted average precision." Unfortunately, the denominator $\sum_{k=1}^{i} y_k$ becomes partly a constant at low positions in the list because all relevant documents are placed in ranking within R (i.e., the number of relevant documents). This means that this indicator cannot be sensitive for assessing ranking below the Rth document, i.e., it gives the same score to two patterns of

ranking that are different in the part below the Rth position. To solve this problem, Sakai et al. [18] proposed an alternative indicator, "Q-measure," which is defined as

$$v_Q = \frac{1}{R}\sum_{i=1}^{n} I(x_i)\frac{\sum_{k=1}^{i} I(x_k)(x_k+1)}{i+\sum_{k=1}^{i} y_k}. \tag{35}$$

Since its denominator is added to the index number of the position, i, it always becomes larger as the order descends over the Rth position.

When we consider the differential of v_Q by x_j, it is necessary to assume that $x_j > 0$ (or $I(x_j) > 0$), and so

$$\frac{dv_Q}{dx_j} = \frac{1}{R}\left(\sum_{i=j}^{n} I(x_i)\frac{1}{i+\sum_{k=1}^{i} y_k}\right). \tag{36}$$

In contrast, for the case when x_j changes from 0 to a positive value (i.e., from $I(x_j) = 0$ to $I(x_j) = 1$), we have to use the following equation,

$$\Delta v_Q(j) = \frac{1}{R}\left(\frac{\sum_{k=1}^{j} I(x_k)(x_k+1)}{j+\sum_{k=1}^{j} y_k} + \sum_{i=j+1}^{n} I(x_i)\frac{x_j+1}{i+\sum_{k=1}^{i} y_k}\right). \tag{37}$$

Eq.(36) contains a summation starting from the index j, which means that the differential increases as j decreases. Hence, we can see that a relevant document at a higher position contributes much more to increasing the score, similar to the other indicators.

4.1.6 Average distance measure

The "average distance measure (ADM)" proposed by Mea and Mizzaro [19] is defined as

$$v_A = 1 - \frac{1}{N}\sum_{i:d_i \in D} |s_i - x_i|, \tag{38}$$

where D indicates the whole database, s_i is the relevance degree of the ith document d_i estimated by a retrieval system and x_i is the actual relevance degree. Note that information on the position of each document in a ranked list is not explicitly included in the ADM formula.

4.2 Generalization of average precision

It is easy to extend the traditional average precision for incorporating multiple degrees of relevance. First, we introduce a quantity s_R such that

$$s_R = \sum_{i=1}^{R} I(x_i)\bar{x}_i. \tag{39}$$

Then, average precision can be written as

$$v = \frac{1}{\max s_R} \sum_{i=1}^{n} I(x_i)\bar{x}_i = \frac{1}{\max s_R} \sum_{i=1}^{n} \frac{I(x_i)}{i} \sum_{k=1}^{i} x_k \tag{40}$$

since $\max s_R = R$ if x_i is a binary variable. In general, we can define $\max s_R$ such that

$$\max s_R = \sum_{i=1}^{N} I(y_i)\bar{y}_i, \tag{41}$$

using the ideal ranked list, $y_1, ..., y_N$. By substituting this equation into Eq.(40), we obtain

$$v_G = \frac{1}{\sum_{i=1}^{N} I(y_i)\bar{y}_i} \sum_{i=1}^{n} \frac{I(x_i)}{i} \sum_{k=1}^{i} x_k. \tag{42}$$

Needless to say, v_G can be used when x_i is not binary. We call this indicator "generalized average precision."

If we assume that $x_j > 0$ (or $I(x_j) > 0$), the differential of v_G by x_j becomes

$$\frac{dv_G}{dx_j} = \frac{1}{\sum_{i=1}^{N} I(y_i)\bar{y}_i} \sum_{k=j}^{n} I(x_k)\frac{1}{k}. \tag{43}$$

When x_j changes from 0 to a positive value (i.e., from $I(x_j) = 0$ to $I(x_j) = 1$), we have to use the following equation,

$$\Delta v_G(j) = \frac{1}{\sum_{i=1}^{N} I(y_i)\bar{y}_i} \left(\frac{1}{j} \sum_{k=1}^{j} x_k + \sum_{t=j+1}^{n} \frac{I(x_t)x_j}{t} \right), \tag{44}$$

which corresponds to Eq.(10).

4.3 Numerical comparison

In this section, we briefly compare scores between some multi-grade indicators using a simple example. Since the purpose of this paper is to examine indicators appropriate for assessing a set of ranked documents in standard retrieval experiments, we select only rank-sensitive indicators, i.e., modified sliding ratio ($v_{S'}$), average of normalized DCG or nDCG (v_D), Q-measure (v_Q) and generalized average precision (v_G).

Suppose that there are just three relevant documents for a given query in a database, the relevance degrees of which are 3, 2, 1 respectively, and that a system presents to its users just 5 documents (i.e., $n = 5$) selected from the database. It is easy to compute the scores of each evaluation indicator for all patterns of the output. We denote each output pattern by a brief representation such as "32100", which means that $x_1 = 3, x_2 = 2, x_3 = 1, x_4 = 0$ and $x_5 = 0$. In total, there are 136 different patterns from "32100" to "00000" in this situation.

Table 3 shows values of average and standard deviation by each indicator for the sample of 136 patterns.

Clearly, there is no large difference of average and standard deviation between these evaluation indicators in this case. This means that the distributions of scores of these indicators are almost the same, although the Q-measure indicates a somewhat higher average score and that of generalized average precision is somewhat lower.

Table 3. Statistics for indicators (sample size is 136)

	$v_{S'}$	v_D	v_Q	v_G
average	.488	.443	.503	.410
std. dev.	.245	.250	.240	.228

Table 4. Correlation matrix (sample size is 136)

	$v_{S'}$	v_D	v_Q	v_G
$v_{S'}$	1.000			
v_D	0.969	1.000		
v_Q	0.885	0.840	1.000	
v_G	0.963	0.940	0.961	1.000
Ave. Pre.	0.857	0.829	0.928	0.894

Table 5. Scores of indicators for some characteristic patterns

	Pattern	$v_{S'}$	v_D	v_Q	v_G
a.	32000	0.923	0.933	0.667	0.733
b.	00123	0.331	0.184	0.513	0.304
	b/a	0.358	0.197	0.770	0.415
c.	03210	0.558	0.610	0.750	0.622
	c/a	0.604	0.654	1.125	0.848
d.	30000	0.692	0.640	0.333	0.400
	d/a	0.750	0.686	0.500	0.545
e.	00003	0.138	0.046	0.121	0.080
	e/d	0.200	0.072	0.364	0.200

Table 4 is a correlation matrix between a set of the four indicators and traditional average precision for our sample. Scores of the traditional (binary) average precision were computed assuming that a document whose relevance degree is more than 0 is relevant and otherwise irrelevant. It turns out that the Q-measure has significantly lower correlation with the modified sliding ratio and nDCG, i.e., the correlation coefficients are 0.885 and 0.840 respectively. We may consider that the pattern assessed highly by the Q-measure is somewhat different from those assessed highly by the modified sliding ratio and nDCG. In contrast, the modified sliding ratio and nDCG have the highest correlation (i.e., 0.969) among them, and the generalized average precision maintains high correlations with the other three indicators. We may thus represent the relationships between the four indicators as "($v_{S'}$, v_D) – (v_G) – (v_Q)."

In order to explore further the differences between the four indicators, we select some characteristic patterns from our example (see Table 5).

(a) Pattern "32000"
The score of Q-measure is relatively low, i.e., $v_Q = .667$, while those of the modified sliding ratio and nDCD are very high (.923 and .933, respectively). The generalized average precision is at a mid point near to the Q-measure ($v_G = .733$).

(b) Pattern "00123"
Unlike the case of pattern "32000", the score of Q-measure is the highest among all indicators, i.e., $v_Q = .513$, for the pattern "00123", in which all relevant documents appear in the list. The characteristics become clearer when we compare the score of v_Q with that of nDCG, which is .184, or only about 20 percent of the score for the pattern "32000". The Q-measure may be a kind of recall-oriented indicator in that the score shows higher performance in the case of more relevant documents within the list.

(c) Pattern "03210"
The recall-oriented nature of the Q-measure is again observed for the pattern "03210", which is a pattern in which the top-ranked document is irrelevant but relevant documents are successfully ranked from the second position in decreasing order of relevance. In the case of Q-measure, the score for "03210" outperforms that of "32000". In contrast, the three other indicators show higher performance for the pattern "32000" than "03210". However, only the generalized average precision shows a similar tendency as the Q-measure, i.e., the score ($v_G = .622$) is relatively large in comparison with that for "32000" (the ratio is 84.8 percent). The differential of these two indicators (Q-measure and generalized average precision) contains a summation having an indicator function $I(x_j)$ (see Eq.(36) and (43)). Hence, these indicators may take a higher score as the list contains many more documents whose relevance degree is more than zero (i.e., $I(x_j) = 1$).

(d) Pattern "30000"
The pattern "30000" is the case in which only the most relevant document is included in the list and is top-ranked. If users wish to get only the most relevant document, they can do so with this pattern even though other relevant documents are not included in the list. For such pattern, the modified sliding ratio shows high performance, i.e., $v_{S'} = .692$, which is just 75 percent of the score for the pattern "32000". A similar tendency is also observed in the result of nDCG. In contrast, the Q-measure does not assign a particularly high score to the pattern "30000", and the generalized average precision is again at a mid point between $(v_{S'}, v_D)$ and (v_Q).

(e) Pattern "00003"
The score of the nDCG is the lowest ($v_D = 0.046$). A similar tendency is also

observed in the patter "00123" as discussed above. Intuitively, $d_I(k)$ in the differential of this indicator (see Eq.(33)) affects this tendency.

Although we cannot obtain a clear conclusion from a simple numerical comparison using a small and artificial example, our restricted simulation reveals that the Q-measure is recall-oriented, whereas the modified sliding ratio and nDCG are precision-oriented. The generalized average precision may be at a mid point between the two sides. The nDCG tends to estimate lower those patterns having relevant documents only in the bottom of the ranked list.

5 Concluding remarks

In this paper, we have discussed some properties of average precision and explored indicators directly using results from multi-grade judgments of relevance. Some interesting findings were obtained through theoretical and empirical examinations. However, all empirical examinations in this report were executed just on small samples that were artificially generated under some assumptions. There is room for further research using larger samples obtained in more realistic situations of retrieval experiments.

References

[1] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2000.

[2] Kazuaki Kishida. Property of average precision as performance measure for retrieval experiment. *IPSJ Transactions on Databases*, 43(SIG 2):11–26, 2002. (in Japanese).

[3] G. Salton and M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of the Association for Computer Machinery*, 15(1):8–36, 1968.

[4] W. S. Cooper. Expected search length: a single measure of retrieving effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1):30–41, 1968.

[5] John A. Swets. Effectiveness of information retrieval methods. *American Documentation*, 20:72–89, 1969.

[6] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.

[7] William G. Cochran. *Sampling Techniques, 3rd ed.* John Wiley and Sons, 1977.

[8] Rong Tang, Jr. William M. Shaw, and Jack L. Vevea. Toward the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science and Technology*, 50(3):254–264, 1999.

[9] Kazuaki Kishida, Kuang Hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, Sung Hyon Myaeng, and Eguchi Koji. Overview of CLIR task at the fourth NTCIR Workshop. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 1–59, 2004.

[10] E. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, 2001.

[11] S. M. Pollack. Measures for the comparison of information retrieval systems. *American Documentation*, 19(4):387–397, 1968.

[12] P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–331, 1998.

[13] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, 2000.

[14] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaulation of IR techniques. *ACM Transactions on Information Systems*, 20:422–446, 2002.

[15] Jaana Kekäläinen. Binary and graded relevance in IR evaluations: comparison of the effects on ranking of IR systems. *Information and Processing Management*, 41:1019–1033, 2005.

[16] Jaana Kekäläinen and Kalervo Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[17] N. Kando, K. Kuriyama, and M. Yoshioka. Information retrieval system evaluation using multi-grade relevance judgments: discussion on averageable single-numbered measures. *IPSJ SIG Notes*, FI-63-12:105–112, 2001. (in Japanese).

[18] T. Sakai, M. Koyama, and A. Kumano. Toshiba BRIDJE at NTCIR-4 CLIR: monolingual/bilingual IR and flexible feedback. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 65–72, 2004.

[19] Vincenzo Della Mea and Stefano Mizzaro. Measuring retrieval effectiveness: a new proposal and a first experimental validation. *Journal of the American Society for Information Science and Technology*, 55(6):513–529, 2004.

[20] Y. Sagara. Performance measures for ranked output retrieval systems. *Journal of Japan Society of Information and Knowledge*, 12(2):22–36, 2002. (in Japanese).

Appendix

As is well-known in statistical science, $V_p[V_m(\cdot|S)]$ can be decomposed as

$$V_p[V_m(\cdot|S)] = E_p[V_m(\cdot|S)] + V_p[E_m(\cdot|S)] \tag{45}$$

(see p.275 in the textbook by W.G. Cochran [7]). First, we obtain

$$V_p[E_m(\bar{v}_a|S)] = \frac{1}{L}\sigma_\mu^2 \tag{46}$$

since $E_m(\bar{v}_a|S)$ is a sample mean of μ (see Eq.(21)) and it is clear that the variance of a sample mean can be obtained by dividing population variance σ_μ^2 by sample size L.

Regarding the first term $E_p[V_m(\bar{v}_a|S)]$, we obtain

$$V_m(\bar{v}_a|S) = V_m\left[L^{-1}\sum_{h=1}^{L} v_{ha}\middle| S\right] = L^{-2}\sum_{h=1}^{L} V_m(v_{ha}|S). \tag{47}$$

Under the assumption of independent relevance judgments between search topics,

$$L^{-2}\sum_{h=1}^{L} V_m(v_{ha}|S) = L^{-2}\sum_{h=1}^{L} V_m(v_{ha}|h) = L^{-2}\sum_{h=1}^{L}\sigma_h^2. \tag{48}$$

Therefore,

$$E_p[V_m(\bar{v}_a|S)] = E_p\left[L^{-2}\sum_{h=1}^{L}\sigma_h^2\right] = L^{-1}E_p\left[L^{-1}\sum_{h=1}^{L}\sigma_h^2\right]. \tag{49}$$

If we define σ_d^2 as in Eq.(25), it is clear that Eq.(24) can be obtained (note that $V_p[E_m(\bar{v}_a|S)] = L^{-1}\sigma_\mu^2$).