Noname manuscript No.
(will be inserted by the editor)

Evaluating Information Retrieval System Performance Based on User Preference

Bing Zhou · Yiyu Yao

Received: date / Accepted: date

Abstract One of the challenges of modern information retrieval is to rank the most relevant documents at the top of the large system output. This calls for choosing the proper methods to evaluate the system performance. The traditional performance measures, such as precision and recall, are based on binary relevance judgment and are not appropriate for multi-grade relevance. The main objective of this paper is to propose a framework for system evaluation based on user preference of documents. It is shown that the notion of user preference is general and flexible for formally defining and interpreting multi-grade relevance. We review 12 evaluation methods and compare their similarities and differences. We find that the normalized distance performance measure is a good choice in terms of the sensitivity to document rank order and gives higher credits to systems for their ability to retrieve highly relevant documents.

Keywords Multi-grade relevance · Evaluation methods · User preference

1 Introduction

The evaluation of information retrieval (IR) system performance plays an important role in the development of theories and techniques of information retrieval (Cleverdon, 1962; Mizzaro, 2001; Jarvelin & Kekalainen, 2000; Kando, Kuriyams & Yoshioka, 2001; Sakai, 2003; Yao, 1995). Traditional IR models and associated evaluation methods make the binary relevance assumption (Cleverdon, 1966; van Rijsbergen, 1979; Buckley & Voorhees, 2000; Rocchio, 1971). That is, a document is assumed to be either relevant (i.e., useful) or non-relevant (i.e., not useful). Under this assumption, the information retrieval problem is implicitly formulated as a classification problem. Consequently, classical system performance measures, such as precision, recall, fallout, etc., are related to the effectiveness of such a two-class classification. In modern IR systems, users can easily acquire a large number of relevant documents for a query which exceed the number they want to examine. It is therefore important for a system to assign

B. Zhou · Y.Y. Yao
Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
E-mail: {zhou200b, yyao}@cs.uregina.ca

weights to the retrieved documents and provide a ranked list to the user. In other words, documents that are more relevant are ranked ahead of documents that are less relevant. This requires us to reconsider relevance as a continuous value rather than a dichotomous one. Many studies show that documents are not equally relevant to users, some documents are more relevant, and some documents are less relevant, relevance has multiple degrees (Cox, 1980; Cuadra & Katter, 1967; Mizzaro, 2001; Jacoby & Matell, 1971; Jarvelin & Kekalainen, 2000; Kando, Kuriyams & Yoshioka, 2001; Tang, Vevea & Shaw, 1999). For example, given a user who wants to study different information retrieval models, the classical book by van Rijsbergen (1979) is more relevant than a paper that only mentions the vector space model once. Given a user who wants to know the general definition of cognitive informatics, the explanation from Wikipedia might be more useful than a research paper that focuses on resolving a specific issue in cognitive informatics. To better identify user needs and preference, we should consider non-binary relevance judgments. In this case, the traditional IR evaluation measures are no longer appropriate. This calls for the effectiveness measures that are able to handle multiple degrees of relevance.

Two possible interpretations of non-binary relevance may exist. One view treats relevance as a relative notion. The relevance of a document is defined in comparison with another document. That is, some documents are more relevant than others. The second view treats relevance as a quantitative notion. One may associate a grade or a number to indicate the degree of relevance of a document. While the former has led to the notion of user preference of documents, the latter led to the notion of multi-grade relevance. It has been shown that multi-grade relevance can be formally defined in terms of user preference (Wong, Yao & Bollmann, 1988; Wong & Yao, 1990; Yao, 1995). Therefore, user preference may be used as a primitive notion based on which a new theory of information retrieval can be built. The main objective of this paper is to propose a framework for IR system evaluation based on user preference of documents. We compare 12 evaluation methods through theoretical and numerical examinations. We hope that our work can bridge the gap between the IR system evaluations based on binary and non-binary relevance, and provide evidence for choosing suitable evaluation methods.

The rest of the paper is organized as follows. In Section 2, we analyze multi-grade relevance judgments and user preference relations. In Section 3, some commonly used traditional IR system evaluation methods based on binary relevance are reviewed. In Section 4, a general evaluation strategy based on non-binary relevance is proposed. The main methodologies of 12 multi-grade evaluation methods are analyzed in Section 5. We analyze the similarities and differences between these methods in Section 6. Some practical issues and possible solutions in using multi-grade evaluation are discussed in Section 7. The findings of this study are given in Section 8.

2 Multi-grade Relevance and User Preference

The term relevance judgments indicate users' decision on whether a document satisfies their information needs of a specific topic. In the series of studies of relevance proposed by Cuadra and Katter (1967), 5 relevance-related aspects have been identified. One of these aspects is the interpretation of relevance judgments that is fundamental to the development of IR system evaluation. The difficulty in using binary relevance is that

it cannot adequately express the continuous nature of relevance. We investigate two non-binary relevance interpretations in this section.

2.1 Multi-grade Relevance Judgments

The multi-grade relevance judgments provide an alternative interpretation on why an IR system should rank documents. When the two-valued relevance judgments are used, an IR system ranks documents based on the probability ranking principle (Fuhr, 1989; Robertson, 1977; van Rijsbergen, 1979). That is, the system ranks documents according to their probability of being relevant. With multi-grade relevance, we have an utility ranking principle. Documents are ranked based on their utilities to users, which are represented by the multi-grade relevance judgments.

Over the years, many types of multi-grade relevance judgments have been introduced. In the experimental study of Katter (1968), relevance judgments are classified into category, ranking, and ratio scales. The category rating scales have been used by many evaluation methods, in which the relevance judgments is expressed by numbers from a finite, predetermined scale ranging typically from 2 to 11 points.

Category rating scale is an ordinal scale. There is an ordering relationship between documents, that is qualitative rather than quantitative. The traditional binary relevance can be seen as a two category scale consisting of relevant or non-relevant. The Text REtriveal Conference (TREC) data sets use a 3-point scale (relevant, partially relevant and non-relevant). The NII Test Collection for IR systems (NTCIR) project used a 4-point scale (highly relevant, relevant, partially relevant and non-relevant). Maron and Kuhns (1970) adopted a 5-point scale (very relevant, relevant, somewhat relevant, only slightly relevant, and non-relevant). Cuadra and Katter (1967) applied 2-, 4-, 6-, 8- and 9-point scales in their experiments. Eisenberg and Hu (1987) employed a 7-point scale (from extremely relevant to not at all relevant), and Rees and Schultz (1967) provided an 11-point scale. There is an ongoing debate over the optimal number of categories of multi-grade relevance (Jacoby & Matell, 1971; Rasmay, 1973; Champney & Marshall, 1939; Tang, Vevea & Shaw, 1999). A general agreement is that the optimal number of categories can vary under different situations because people use different standards (Cox, 1980).

The category rating scale is adequate to calculate multi-grade relevance in the evaluation functions. However, there are a few problems in using the category rating scales. First, user judgments are restricted by some fixed relevance scales. If the numbers or the descriptions of each degree are not clearly defined, a multi-grade relevance scale can be easily misused in the evaluation process. Second, there is no common agreement on the optimal number of degrees of relevance, it varies depending on people's intuitions in different scenarios. The limitations of category rating scale motivate a formal model to represent multi-grade relevance.

2.2 User Preference Relation

Eisenberg (1988) introduced magnitude estimation as an open-ended scale where all positive rational numbers can be used to express the different relevance degrees of documents. Compatible with this scheme, the notion of user preference was adopted from decision and measurement theories to represent relevance (Wong, Yao and Bollmann,

1988; Yao, 1995). Under the user preference relation, a user only provides the relative relevance judgments on documents without referring to any predefined relevance scales.

A user preference relation can be formally defined by a pairwise comparison of documents (Bollmann & Wong, 1987; Wong & Yao, 1990; Wong, Yao & Bollmann, 1988; Yao, 1995). Given any two documents $d, d' \in D$, where D denotes a finite set of documents. We assume that a user is able to decide if one document is more useful than another document. The user preference relation can be defined by a binary relation $\succ$ on D as follows:

$$d \succ d' \quad \text{iff} \quad \text{the user prefers } d \text{ to } d'.$$

If a user considers d and d' to be equally useful or incomparable, an indifference $\sim$ relation on D can be defined as follows:

$$d \sim d' \quad \text{iff} \quad (\neg(d \succ d'), \neg(d' \succ d)),$$

where $\neg(d \succ d')$ means that a user does not prefer d to d'. If a preference relation $\succ$ satisfies the following two properties:

$$\text{Asymmetry: } d \succ d' \Rightarrow \neg(d' \succ d),$$

$$\text{Negative transitivity: } (\neg(d \succ d'), \neg(d' \succ d'')) \Rightarrow \neg(d \succ d''),$$

then it is a weak order (Fishburn, 1970). A weak order is transitive, that is, if there is a third document d'', then $d \succ d'$ and $d' \succ d''$ imply $d \succ d''$. A few additional properties are (Fishburn, 1970):

(a). the relation $\sim$ is an equivalence relation;

(b). exactly one of $d \succ d'$, $d' \succ d$ and $d \sim d'$ holds for every $d, d' \in D$;

(c). the set of all equivalence classes in D generated by the relation $\sim$ is denoted as $D/\sim$, the relation $\succ'$ on $D/\sim$ defined by $X \succ' Y \Leftrightarrow \exists d, d'(d \succ d', d \in X, d' \in Y)$ is a linear order, where X and Y are elements of $D/\sim$.

A linear order (total order) is a weak order in which any two different elements are comparable. If $\succ$ is a weak order, the indifference relation $\sim$ divides the document set into disjoint subsets. Furthermore, for any two equivalence classes X and Y of $\sim$, either $X \succ' Y$ or $Y \succ' X$ holds. In other words, it is possible to arrange documents into several levels so that documents in a higher level are preferred to documents in a lower level, and documents in the same level are indifferent (Cooper, 1968).

By quantifying a user preference relation with a utility function, multi-grade relevance can be formally defined in terms of user preference. If a preference relation $\succ$ is a weak order, there exists a utility function u such that:

$$d \succ d' \quad \Longleftrightarrow \quad u(d) > u(d'),$$

where $u(d)$ can be interpreted as the relevance value of document d.

Example 1 Suppose a user preference relation $\succ$ on $D = \{d_1, d_2, d_3, d_4, d_5\}$ is specified by the following weak order:

$$d_5 \succ d_4, \quad d_5 \succ d_1, \quad d_5 \succ d_2, \quad d_5 \succ d_3, \quad d_4 \succ d_1, \quad d_4 \succ d_2, \quad d_4 \succ d_3.$$

The indifference relation $\sim$ divides D into three subsets $\{d_5\}$, $\{d_4\}$ and $\{d_1, d_2, d_3\}$, and they can be arranged into three levels:

$$d_5 \succ d_4 \succ \begin{matrix} d_1 \\ d_2 \\ d_3 \end{matrix}$$

The utility function $u(d)$ quantifies this document ranking as follows:

$$u_1(d_5) = 0.2,\ u_1(d_4) = 0.1,\ u_1(d_1) = u_1(d_2) = u_1(d_3) = 0.0.$$

In this case, the user preference relation is able to represent a 3-point category rating scale including relevant (d_5), partially relevant (d_4), and non-relevant (d_1, d_2, d_3) documents.

If we use a 5-point category rating scale, another utility function may be used as:

$$u_2(d_5) = 0.4,\ u_2(d_4) = 0.2,\ u_2(d_1) = u_2(d_2) = u_2(d_3) = 0.0,$$

which provides a measurement of user preference relation $\succ$ including very relevant (d_5), somewhat relevant (d_4), and non-relevant (d_1, d_2, d_3) documents. Although u_1 and u_2 use different absolute values, they preserve the same relative order for document pairs.

From the above example, we can see that given a user preference relation, we can represent any degree of multi-grade relevance. On the other hand, some existing multi-grade relevance scales can be easily interpreted in terms of user preference relation. Compared to a category rating scale, the user preference relation is easier for users to make their judgments, as it is not restricted to a predefined scale, and is rich enough to represent any degrees of relevance.

3 Traditional IR System Evaluation

The six evaluation criteria suggested by Cleverdon provide a foundation for designing IR system evaluation methods (Cleverdon, 1966):

(1). The coverage of the collection;
(2). System response time;
(3). The form of the presentation of the output;
(4). User efforts involved in obtaining answers to a query;
(5). Recall;
(6). Precision.

Of these criteria, precision and recall were most frequently used and still are the dominant approach to evaluate the performance of information retrieval systems. Let L denote the size of the document collection, let R denote the number of relevant documents for a query, and let N denote the number of documents in the ranked output. Precision is defined as the proportion of retrieved documents that are actually relevant which can be expressed as:

$$\text{precision} = \frac{\sum_{i=1}^{n} d_i}{n},$$

where d_i is a variable representing the relevance level of the ith document in the ranked output to a certain query. In the binary relevance case, the possible relevance values of d_i are either 1 representing relevant, or 0 representing non-relevant, so the sum of d_i is the number of relevant documents up to the top n ($n \leq N$) of the ranked output. Recall is the proportion of relevant documents that are actually retrieved which can be expressed as:

$$\text{recall} = \frac{\sum_{i=1}^{n} d_i}{R}.$$

Many alternatives to precision and recall have been suggested. The fallout measure is the proportion of non-relevant documents that are retrieved, written as:

$$\text{fallout} = \frac{n - \sum_{i=1}^{n} d_i}{L - R}.$$

F-measure (van Rijsbergen, 1979) combines recall and precision in a single measure:

$$\text{F} = 2 * (\text{recall} * \text{precision})/(\text{recall} + \text{precision}).$$

The average precision combines precision, relevance ranking, and recall. It 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" (Buckley & Voorhees, 2000):

$$\text{average precision} = \frac{1}{R} \sum_{n=1}^{N} rel(d_n) \frac{\sum_{i=1}^{n} d_i}{n},$$

where $rel(d_n)$ is a function, such that $rel(d_n) = 1$ if the document is relevant and $rel(d_n) = 0$ otherwise. Another measure, R-precision, was found useful in many IR experiments. It is the precision at R, which can be expressed as:

$$\text{R-precision} = \frac{1}{R} \sum_{i=1}^{R} d_i.$$

Some other efforts have also been made for IR system evaluation. Cooper (1968) suggested that the primary function of an IR system is to save its users efforts of discarding irrelevant documents. Based on this principle, he proposed an expected search length measure which is defined as the number of irrelevant documents a user must scan before the desired number of relevant documents can be found.

The traditional binary evaluation methods play a dominant role in the history of IR system evaluation. Each of these measures captures some important but distinct aspects of the IR system performance, and they are complementary to each other. The basic principles underlying these measures are similar: the total number of documents in the collection have been divided into four groups, the documents that are retrieved and relevant, the documents that are retrieved but not relevant, the documents that are not retrieved but relevant, and the documents that are not retrieved and not relevant. By checking the distribution of these four groups of documents, different perspectives of IR system performance can be evaluated by different measures, but all of these measures are based on a dichotomous relevance assumption, and ignore the variability and continuous nature of relevance.

4 A General Evaluation Strategy Based on Multi-grade Relevance

The first generation of IR systems is based on the Boolean model. Each document in the collection has been indexed by a set of terms, and it will be retrieved only if the terms exactly match the terms in the query. As a result, the retrieved results are either too many or too few unranked documents. The later models such as vector space model and probabilistic model assign a weight to each document in the collection based on the match between the query and documents. The retrieved results are partially matched ranked documents. The changes in information retrieval techniques have resulted in the transformation of IR system evaluation from binary relevance to continuous relevance and from binary retrieval to ranked retrieval. In spite of the successes of the traditional IR system evaluation methods, it is time to reconsider the evaluation strategy based on non-binary relevance. This will meet the evaluation needs of the current IR systems, and will provide satisfying evaluation results.

Ideally, a modern IR system is supposed to have the following two features. First, it should be able to return as many relevant documents as possible, and these retrieved documents should be ranked in a decreasing order based on their relevance degrees to a certain query, that is, the more relevant documents should always be ranked ahead of the less relevant documents. Second, it should be able to retrieve the few most relevant documents to a certain query. Among the large numbers of retrieved documents, the most useful ones are those which ranked higher and appear in the first few pages. Therefore, an evaluation method based on non-binary relevance should be able to credit the IR systems having the above features and distinguish them among the others. It should have the following three properties:

(a). The ability to evaluate document rank orders;

(b). The ability to give higher credits to IR systems that can retrieve the few most relevant documents;

(c). The flexibility of adapting to different multi-grade relevance interpretations.

Let us analyze these properties individually and find out how they can be considered in an evaluation function. For property (a), the document rank order provided by the user judgments is considered as the ideal ranking (i.e., user ranking) and the document rank order provided by the IR systems is considered as the system ranking. The IR system evaluation is based on the comparison of these two rankings. The more similar they are, the better the IR system. Generally speaking, there are two ways to compare two rankings. First, they can be compared by the ratio of the sums of their performances. The result ranges from 0.0 to 1.0 where 1.0 indicates the best rank, and 0.0 indicates the worst. Second, they can be compared by a distance function, where a smaller distance indicates a better performance. Compared to the IR system evaluation based on binary relevance, we can see that the general evaluation strategy based on multi-grade relevance is shifted from the distribution of relevant, non-relevant, retrieved and not retrieved documents to the comparison of ideal ranking and system ranking. For property (b), in order to differentiate the few most relevant documents from large numbers of partially relevant documents, the degrees of relevance need to be quantified in the evaluation function. The more relevant documents should be given more weight. For property (c), since there is no general agreement on which relevance expression or scale should be used for multi-grade evaluation, it is important to design

a method that is compatible with different relevance interpretations. Moreover, since the number of degrees of relevance varies, it is necessary for the evaluation method to have the flexibility to support any multi-grade scale of relevance. According to the analysis of these three properties, we can state the multi-grade evaluation methods in two general expressions, written as:

$$\text{IR system performance} = \frac{\text{performance of system ranking}}{\text{performance of ideal ranking}}, \tag{1}$$

or,

$$\text{IR system performance} = 1 - \text{distance function}, \tag{2}$$

where the distance function measures the distance between system ranking and ideal ranking. The abstractions of all the multi-grade evaluation methods analyzed in the next section belong to one of the above equations, although the details of each method vary.

5 Evaluation Methods Based on Multi-grade Relevance

In this section, we review the main methodologies of 12 existing multi-grade evaluation methods, and examine how they embody the properties analyzed in the last section. For a better understanding, we interpret these methods in a unified framework where we know both the actual relevance degrees of documents (i.e., user relevance judgment) and the relevance predicted by an IR system.

5.1 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient (1904) is the Pearson's correlation between ranks. Although it was not originally designed for measuring IR system performance, it has been used by many IR experimental studies for comparing rank correlations. A rank correlation is a number between -1.0 and +1.0 that measures the degree of association between two rankings x_i and y_i. The following formula can be applied to compute the Spearman's rank correlation coefficient when there are no ties between ranks:

$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{N(N^2-1)},$$

where $(x_i - y_i)$ measures the difference between the ranks of corresponding values x_i and y_i, and N is the total number of documents in the ranked output. A positive value of the measure implies a positive association between two rankings, a negative value implies a negative or inverse association, and two rankings are independent when the value is 0.0. The classic Pearson's correlation coefficient between ranks has to be used if tied ranks exist (Myers, 2003).

Example 2 Table 1 contains relevance scores of five documents given by user ranking (UR) and system ranking (IRS), respectively. Ranks are archived by giving "1" to the largest score, "2" to the second largest score and so on. The smallest score gets the lowest rank. The second last column shows the difference in the ranks: The rank of

Table 1 Data Table: Spearman's Rank Correlation Coefficient

Docs	UR	Rank	IRS	Rank	Difference between the ranks: (x_i - y_i)	$(x_i - y_i)^2$
d_1	0.6	1	0.5	2	-1	1
d_2	0.1	5	0.6	1	4	16
d_3	0.3	4	0.4	3	1	1
d_4	0.5	2	0.3	4	-2	4
d_5	0.4	3	0.2	5	-2	4

IRS is subtracted from the rank of UR. Calculating the coefficient using the formula above, we get:

$$\rho = 1 - \frac{6 * (1 + 16 + 1 + 4 + 4)}{5 * (5^2 - 1)} = -0.3,$$

which indicates a negative association between user ranking and system ranking.

5.2 Kendall Tau Rank Correlation Coefficient

Kendall Tau Rank Correlation Coefficient (Kendall, 1938) is another popular measure for rank consistency comparison. The basic principle behind this measure is based on the number of agreeing versus contradictory pairs between ranks. It can be computed by the following formula in absence of tied pairs:

$$Tau = \frac{C^+ - C^-}{\frac{1}{2}N(N-1)},$$

where the denominator is the total number of pairs in a rank containing N documents, and the numerator is the difference between the number of agreeing (C^+) and contradictory (C^-) pairs. Similarly to Spearman's rank correlation coefficient, the value of this measure also lies between -1.0 and 1.0 with -1.0 corresponding to the largest possible distance (obtained when one order is the exact reverse of the other order) and +1.0 corresponding to the smallest possible distance (obtained when both orders are identical). If tied pairs exist on ranks, the following measure, called Kendall Tau-b (Kendall, 1945; Stuart, 1953) can be used for the computation of associations between ranks:

$$Tau\text{-}b = \frac{C^+ - C^-}{\sqrt{(C^+ + C^- + T_Y)(C^+ + C^- + T_X)}},$$

where T_X is the number of pairs not tied on rank X, and T_Y is the number of pairs not tied on rank Y. Another measure called Kendall Tau-c (Kendall, 1945, Stuart, 1953) is used as a variant of Kendall Tau-b for nonsquare tables. It can be computed by the following formula:

$$Tau\text{-}c = (C^+ - C^-)[2m/(N^2(m-1))],$$

where m is the number of rows or columns of the table, whichever is smaller.

Example 3 Suppose there are only four relevant documents for a given query, the user ranking gives the following order: UR = (d_1, d_3, d_2, d_4), and an IR system ranks the documents as: IRS = (d_1, d_3, d_4, d_2). The user ranking is composed of the following 6 ordered pairs:

$$\{(d_1, d_3), (d_1, d_2), (d_1, d_4), (d_3, d_2), (d_3, d_4), (d_2, d_4)\},$$

and the system ranking is composed of the following 6 ordered pairs:

$$\{(d_1, d_3), (d_1, d_4), (d_1, d_2), (d_3, d_4), (d_3, d_2), (d_4, d_2)\}.$$

The set of agreeing pairs between the user ranking and system ranking is:

$$\{(d_1, d_3), (d_1, d_2), (d_1, d_4), (d_3, d_2), (d_3, d_4)\},$$

and the set of contradictory pairs is:

$$\{(d_2, d_4)\}.$$

We can compute the Kendall rank correlation coefficient between UR and IRS as:

$$Tau = \frac{5-1}{\frac{1}{2} * 4 * (4-1)} = 0.67,$$

which indicates a high level association between user ranking and system ranking.

5.3 Sliding Ratio and Modified Sliding Ratio

The sliding ratio was proposed by Pollack (1968). In this method, the ranked system output is compared against the ideal ranking. Relevance score $d_i (i = 1, ..., n)$ is assigned to each ith ranked document in the output list. Different from the binary relevance case, the possible values of d_i ranges from 0.0 to 1.0 where 0.0 represents non-relevant, 1.0 represents the highest relevant level, and each intermediate value represents different relevant levels in between. For example, if a 3-point scale is chosen, the possible scores of d_i are 0.2, 0.1, and 0.0 indicating the relevance degree as highly relevant, relevant, and non-relevant, respectively. The overall ranking is quantified by the sum of d_i. The sliding ratio of the system ranking and ideal ranking is defined as:

$$sr = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} d_{I(i)}},$$

where $d_{I(i)}$ is the relevance score of the ith ranked document in the ideal ranking.

In a good rank order, the ith document should always be ranked ahead of the jth document if $d_i \geq d_j$. Unfortunately, the sliding ratio is not sensitive to the document rank order.

Example 4 Suppose we use a 4-point scale, and there are only five relevant documents whose relevance scores are 0.1, 0.1, 0.2, 0.2 and 0.3. These documents should be ideally ranked as:

$$d_{I(1)} = 0.3,\ d_{I(2)} = 0.2,\ d_{I(3)} = 0.2,\ d_{I(4)} = 0.1,\ d_{I(5)} = 0.1.$$

Suppose a rank predicted by an IR system IRS1 is:

$$d_1 = 0.3,\ d_2 = 0.2,\ d_3 = 0.1,\ d_4 = 0.1,\ d_5 = 0.0.$$

The value of sliding ratio between ideal ranking and IRS1 is $sr = 7/9 = 0.78$. However, suppose there is another IR system IRS2 that gives a different rank:

$$d_1 = 0.1,\ d_2 = 0.1,\ d_3 = 0.2,\ d_4 = 0.3,\ d_5 = 0.$$

Obviously IRS1 performs better than IRS2 in terms of document rank order, but according to the sliding ratio measure, the performance is the same for both systems ($sr = 0.78$).

To solve this problem, Sagara (2002) has proposed the modified sliding ratio that takes the document ordering into account, and it is defined as:

$$msr = \frac{\sum_{i=1}^{n} \frac{1}{i} d_i}{\sum_{i=1}^{n} \frac{1}{i} d_{I(i)}}.$$

That is, when a highly relevant document is ranked at the bottom of the output list, its contribution to the whole system performance drops. By using the modified sliding ratio, IRS1 ($msr = 0.90$) performs better than IRS2 ($msr = 0.57$).

5.4 Cumulated Gain and Its Generalizations

The cumulated gain proposed by Jarvelin and Kekalainen (2000) is very similar to the idea of sliding ratio. It assumes that the user scans the retrieved document list, and adds a relevance score each time he finds a relevant document. The cumulated gain is defined as:

$$cg = \sum_{i=1}^{n} d_i.$$

Jarvelin and Kekalainen (2000) also take the document ordering into consideration by adding a discounting function after the bth ranked document in the output, which progressively reduces the document relevance score as its rank increases. The discounted cumulative gain is defined as $dcg = \sum_{b<i\leq n} \frac{d_i}{\log_b i}$ for $i > b$ and $dcg = \sum_{1\leq i\leq b} d_i$ otherwise. This idea is similar to the modified sliding ratio, except that the latter uses divisions instead of logarithms which makes the reduction steep.

Suppose dcg_I is the discounted cumulated gain of an ideal ranking, the normalized discounted cumulated gain at document cut-off r is defined as (Jarvelin & Kekalainen, 2002):

$$ndcg = \frac{1}{r} \sum_{n=1}^{r} \frac{dcg}{dcg_I}.$$

5.5 Generalizations of Average Precision

Weighted average precision was introduced by Kando, Kuriyams and Yoshioka (2001) as an extension of average precision in order to evaluate multi-graded relevance. The average precision has been widely used in IR experiments for evaluating binary relevance. Since the average precision is based on binary relevance, the possible relevance scores of d_i are either 1 representing relevant, or 0 representing non-relevant. The sum of d_i is the number of relevant documents up to the nth ranked document.

The weighted average precision extends the average precision by assigning multi-grade relevance scores to d_i. The sum of d_i is the cumulated gain cg, and the cumulated gain of an ideal ranking up to the nth ranked document is denoted by cg_I. The weighted average precision is defined as:

$$wap = \frac{1}{R} \sum_{n=1}^{N} rel(d_n) \frac{cg}{cg_I}.$$

However, in the case of $n > R$, the cumulated gain of the ideal ranking cg_I becomes a constant after the Rth ranked document, so it cannot distinguish between two systems when one of the systems has some relevant documents ranked at the bottom of n.

Example 5 Suppose there are five documents in the ranked output ($n = 5$), and three relevant documents to a query ($R = 3$). If we use a 4-point scale, the sequence of relevance in an ideal ranking is:

$$(0.3, 0.2, 0.1, 0.0, 0.0).$$

The first IR system IRS1 retrieved only one highly relevant document and gives a rank as:

$$(0.0, 0.0, 0.2, 0.0, 0.0).$$

The second system IRS2 also retrieved one highly relevant document and ranks the documents as:

$$(0.0, 0.0, 0.0, 0.0, 0.2).$$

It is obvious that IRS1 performs better than IRS2 in terms of document rank order. However, according to the weighted average precision, their performance scores are the same ($wap = (2/6)/3 = 0.11$).

Sakai proposed the Q-measure (2004) in order to address this problem. In Q-measure, the relevance score or gain value d_i is replaced by the bonused gain $bg_i = d_i + 1$ if $d_i > 0$ and $bg_i = 0$ otherwise. The cumulated bonused gain is defined as:

$$cbg = \sum_{i=1}^{n} bg_i.$$

Q-measure is defined as:

$$Q = \frac{1}{R} \sum_{n=1}^{N} rel(d_n) \frac{cbg}{cg_I + n}.$$

The denominator $(cg_I + n)$ always increases after the Rth ranked document instead of remaining constant. In the above example, the performance of IRS1 evaluated by the Q-measure is: $Q = (3/(6 + 3))/3 = 0.11$, and the performance of IRS2 is: $Q = (3/(6 + 5))/3 = 0.09$. IRS1 performs better than IRS2 according to Q-measure.

5.6 Average Gain Ratio

In information retrieval experiments, one of the important properties that needs to be evaluated is how well the few most relevant documents are retrieved, but most evaluation methods treat them as same as the partially relevant documents. Since the amount of the partially relevant documents are usually much larger than the most relevant ones, most evaluation methods are affected by how well the partially relevant documents are retrieved. The average gain ratio (Sakai, 2003) is designed for giving more credit to systems for their ability to retrieve the most relevant documents. The relevant score or gain value is adjusted as:

$$d'_{l(i)} = d_{l(i)} - \frac{R_l}{R}(d_{l(i)} - d_{(l-1)(i)}),$$

where l denotes the relevance level, $d_{l(i)}$ denotes the relevance score for finding an l-relevant document at rank i, and R_l denotes the number of l-relevant documents. For example, if we use a 4-point scale, the possible values for $d_{l(i)}$ are 0.3, 0.2, 0.1, 0.0. Suppose there are 10 relevant documents, but a system only retrieved one highly relevant document at the first ranking position and the rest are non-relevant documents. According to the adjusted relevance score: $d'_{l(i)} = 3-(3-2)*1/10 = 2.9$. By employing this value to weighted average precision, the average gain ratio is defined as:

$$agr = \frac{1}{R}\sum_{n=1}^{N} rel(d_{l(n)})\frac{cg'}{cg'_I},$$

where cg' and cg'_I denote the cumulated gain and the cumulated gain of an ideal ranking calculated by $d'_{l(i)}$, respectively.

5.7 Normalized Distance Performance Measure

The normalized distance performance measure (Yao, 1995) is adopted from the distance function between two rankings used by Kemeny and Sneel (1962). It measures the distance between user ranking $\succ_u$ and system ranking $\succ_s$ by examining the agreement and disagreement between these two rankings, which is similar to the idea of Kendall tau rank correlation coefficient.

The distance between two rankings is defined with respect to the relationships between document pairs: two rankings agree on a pair of documents $d, d' \in D$ if both of them rank d and d' in the same order; they contradict each other if one ranking ranks d higher and the other ranking ranks d' higher; they are compatible with each other if one ranking ranks d or d' higher and the other ranking has d and d' tied. The numbers of agreeing pairs C^+, contradictory pairs C^-, and compatible pairs C^0 can be defined as:

$$\begin{aligned}
C^+ &= |\succ_u \cap \succ_s|,\\
C^- &= |\succ_u \cap \succ_s^c| = |\succ_u^c \cap \succ_s|,\\
C^0 &= |\succ_u \cap \sim_s| + |\sim_u \cap \succ_s| = C^u + C^s,
\end{aligned}$$

where $|\cdot|$ denotes the cardinality of a set, and $\succ_s^c$ denotes converse ranking of $\succ_s$, which can be obtained by reading the original ranking backwards. Let the distance count as 0 if two rankings agree on a document pair, count as 1 if they are compatible on a document pair, and count as 2 if they contradict on a document pair. The distance function between the user ranking $\succ_u$ and system ranking $\succ_s$ is defined as:

$$\beta(\succ_u, \succ_s) = 2 * C^- + 1 * C^0 + 0 * C^+ = 2C^- + C^0 = 2C^- + C^u + C^s.$$

The notion of acceptable ranking (Wong & Yao, 1990; Wong, Yao & Bollmann, 1988) was suggested to be more suitable for information retrieval, and this provides the possibility to derive a performance measure by using the distance between system ranking and acceptable ranking. There are many acceptable rankings with respect to $\succ_u$. For the definition of a fair measure, one should choose an acceptable ranking closest to $\succ_s$, which is defined as:

$$\succ_a \quad = \quad \succ_u \cup(\sim_u \cap \succ_s).$$

The following distance-based performance measure can be derived:

$$dpm(\succ_u, \succ_s) = min_{\succ \in \Gamma_u(D)} \beta(\succ, \succ_s) = \beta(\succ_a, \succ_s),$$

where $\Gamma_u(D)$ is the set of all acceptable rankings of $\succ_u$. For ranking $\succ_a$ and $\succ_s$, the number of agreeing, contradictory, and compatible pairs are:

$$\begin{aligned} &| \succ_a \cap \succ_s | = |(\succ_u \cap \succ_s) \cup (\sim_u \cap \succ_s)| = C^+ + C^s, \\ &| \succ_a \cap \succ_s^c | = | \succ_u \cap \succ_s^c | = C^-, \\ &| \succ_a \cap \sim_s | = | \succ_u \cap \sim_s | = C^u. \end{aligned}$$

Therefore, the distance function between $\succ_u$ and $\succ_s$ can be rewritten as:

$$dpm(\succ_u, \succ_s) = \beta(\succ_a, \succ_s) = 2C^- + C^0 = 2C^- + C^u.$$

The normalized distance performance measure was also proposed to measure the performance of every query equally. It is defined as:

$$ndpm(\succ_u, \succ_s) = \frac{dpm(\succ_u, \succ_s)}{\max_{\succ \in \Gamma(D)} dpm(\succ_u, \succ)},$$

where $\max_{\succ \in \Gamma(D)} dpm(\succ_u, \succ)$ is the maximum distance between $\succ_u$ and all rankings. Based on the definition of *dmp*, the converse ranking $\succ_u^c$ produce the maximum *dmp* value, that is,

$$\max_{\succ \in \Gamma(D)} dpm(\succ_u, \succ) = dpm(\succ_u, \succ_u^c) = 2\ |\succ_u^c| = 2\ |\succ_u| = 2C,$$

where C denotes the total number of document pairs qualifying the user preference relation in the user ranking. Combining the above results, the normalized distance performance measure can be computed as:

$$ndpm(\succ_u, \succ_s) = \frac{dpm(\succ_u, \succ_s)}{dpm(\succ_u, \succ_u^c)} = \frac{2C^- + C^u}{2C}.$$

Example 6 Let

$$\begin{matrix} d_1 \\ d_2 \end{matrix} \succ_u d_3 \succ_u \begin{matrix} d_4 \\ d_5 \end{matrix} \succ_u d_6$$

be a user ranking on a set of documents $D = (d_1, d_2, d_3, d_4, d_5, d_6)$, and

$$\begin{matrix} d_1 \\ d_3 \end{matrix} \succ_s \begin{matrix} d_2 \\ d_5 \end{matrix} \succ_s d_6 \succ_s d_4$$

be a system ranking. With respect to $\succ_u$, the closest acceptable ranking to $\succ_s$ is given by:

$$d_1 \succ_a d_2 \succ_a d_3 \succ_a d_5 \succ_a d_4 \succ_a d_6.$$

The contradict pairs between $\succ_a$ and $\succ_s$ are (d_2, d_3) and (d_4, d_6), $C^- = 2$. The compatible pairs are (d_1, d_3) and (d_2, d_5), $C^u = 2$. The value of the normalized distance performance measure is:

$$ndpm(\succ_u, \succ_s) = \frac{2C^- + C^u}{2C} = \frac{2 * 2 + 2}{2 * 13} = 0.23,$$

and the value of the IR system performance is $1 - 3/13 = 0.77$.

5.8 Average Distance Measure

The average distance measure (Mizzaro, 2001) measures the distance between user ranking and system ranking by examining the absolute differences between system relevance estimation and user relevance estimation. Suppose D is the whole document collection, for any document $d \in D$, let s_i denote the relevant score of the ith document estimated by the IR system, and let u_i denote the relevance score of the ith document estimated by the user. The average distance measure is defined as:

$$adm = 1 - \frac{\sum_{d \in D} |s_i - u_i|}{|D|}.$$

6 Comparison of Multi-grade Evaluation Measures

In this section, the properties of evaluation methods and their connections are examined.

6.1 General Observations

Most of these methods are based on cumulated gain (Jarvelin & Kekalainen, 2000; Jarvelin & Kekalainen, 2002; Kando, Kuriyams & Yoshioka, 2001; Sakai, 2004; Sakai, 2003), in which category rating scales are being used as their multi-grade relevance interpretation, and their evaluation functions match the general Equation (1) by using the ratio of system ranking and ideal ranking. The basic ideas underlying these methods are very similar to the sliding ratio method proposed back in 1968, that is, each retrieved document has been assigned a relevance score corresponding to a predefined relevance scale, the system ranking is quantified by the sum of relevance scores of each retrieved document, the user ranking is quantified by the sum of relevance scores of these documents in an ideal ranking, and the overall IR system performance is the proportion of these two rankings. The problem with the cumulated gain measure is that it is not sensitive to the document rank order. For example, if two IR systems retrieved the same document set with exactly opposite rank orders, their performances evaluated by cumulated gain measure will be the same since the sums of their relevant scores are the same. Some efforts have been made to reduce this problem. Discounted cumulated gain, normalized discounted cumulated gain, weighted average precision, Q-measure and average gain ratio all take the document ordering into consideration.

Some methods are generalized directly from the methods based on binary relevance in order to evaluate multi-grade relevance. The weighted average precision extends the widely used average precision by assigning multi-grade relevance scores to the retrieved documents. The problem of weighted average precision is that it is incapable of evaluating documents retrieved after rank R (i.e., the total number of relevant documents). Q-measure is proposed to address this problem by replacing the cumulated gain with bonused gain. The average gain ratio is also generalized from weighted average precision for the purpose of giving more credit to the systems that can retrieve the few most relevant documents, which is a very important issue in modern IR system evaluation.

Unlike cumulated gain-based methods, the normalized distance performance measure, Spearman correlation coefficient, Kendall tau rank correlation coefficient and average distance measure focus on measuring the distance between system ranking and user ranking. The evaluation functions of these four methods match the general Equation (2) by using the distance function, and they are more sensitive to the document rank order, compared with cumulated gain-based methods. The normalized distance performance measure uses a user preference relation as its multi-grade relevance interpretation, and the distance function is computed by considering the relationships of document pairs of system ranking and user ranking which is similar to the ideal of Kendall tau rank correlation coefficient. Spearman correlation coefficient uses rank positions instead of relevance scores to calculate the distance between two rankings. The average distance measure is still based on category rating scales, and the distance function is calculated based on the absolute differences between system relevance estimation and user relevance estimation of each document in the collection. It gives wrong evaluation results in some cases. For example, suppose that we are using a 7-point scale, and there are only three relevant documents to a given query, the sequence of the relevance score given by the user is (0.3, 0.2, 0.1), the sequence given by the first IR system IRS1 is (0.6, 0.4, 0.2), the sequence given by IRS2 is (0.1, 0.2, 0.3). It is obvious that the IRS1 performs better in terms of document rank order ($adm = 0.8$), but according to the average distance measure, the second system performs better ($adm = 0.87$).

6.2 Numerical Comparisons

In this section, we compare these multi-grade evaluation methods by employing them in some examples from two different perspectives. Since we already know that the sliding ratio methods and cumulated gain method did not take the document rank order into consideration, and the Spearman correlation coefficient and Kendall tau rank correlation coefficient have different value range, therefore we only compare the other 8 methods. The value of these methods lies between 0.0 and 1.0. The minimum value 0.0 represents the worst system performance, and the maximum value 1.0 represents the best system performance.

First, we compare these methods in terms of their sensitivities to the document rank order. Suppose that we are using a 7-point scale, and there are only five relevant documents to a given query. Let UR indicate the ideal ranking or user ranking, IRS1, IRS2, IRS3, and IRS4 represent four different IR systems, respectively. Their performance in terms of document rank order is that IRS1 is the best, IRS2 is better than IRS3, and IRS4 is the worst. Table 2 shows the actual evaluation results by the methods we discussed in Section 5. Let us briefly analyze these evaluation results. All methods are able to determine that IRS1 provides the best ranking and IRS4 provides the worst. The methods based on cumulated gain give an unreasonable evaluation results with respect to the performance of IRS2 and IRS3, because although IRS2 provides a better ranking, the sum of the relevance score of IRS3 is larger than IRS2. If we change the relevance score of each document in IRS2 and IRS3 so that their sums can be the same, Table 3 shows the evaluation results. All the cumulated gain-based methods except the discounted cumulated gain (dcg) are able to give the correct evaluation results at this time. Unfortunately, one cannot manually adjust the relevance scores given by the IR system which are usually decided by some retrieval algorithms automatically. Therefore, the best method in terms of the sensitivity to document rank order is the

Table 2 Evaluation results of document rank order

Docs	d_1	d_2	d_3	d_4	d_5	msr	dcg	ndcg	wap	Q	agr	ndpm	adm
UR	0.6	0.5	0.4	0.3	0.1								
IRS1	0.6	0.5	0.3	0.2	0.1	0.95	0.93	0.96	0.94	0.98	0.94	1.00	0.96
IRS2	0.5	0.3	0.4	0.2	0.1	0.79	0.77	0.78	0.79	0.93	0.78	0.90	0.92
IRS3	0.4	0.6	0.2	0.3	0.1	0.80	0.85	0.82	0.81	0.94	0.80	0.80	0.90
IRS4	0.1	0.2	0.2	0.4	0.5	0.43	0.54	0.34	0.40	0.80	0.36	0.05	0.70

Table 3 Changing the relevance score of Table 2

Docs	d_1	d_2	d_3	d_4	d_5	msr	dcg	ndcg	wap	Q	agr
UR	0.6	0.5	0.4	0.3	0.1						
IRS2	0.6	0.4	0.5	0.3	0.1	0.98	0.98	0.97	0.97	0.99	0.98
IRS3	0.5	0.6	0.3	0.4	0.1	0.95	0.99	0.95	0.95	0.98	0.95

Table 4 Evaluation results to retrieve highly relevant documents

Docs	d_1	d_2	d_3	d_4	d_5	msr	dcg	ndcg	wap	Q	agr	ndpm	adm
UR	0.3	0.3	0.2	0.1	0.1								
IRS1	0.2	0.2	0.0	0.0	0.0	0.53	0.49	0.58	0.54	0.91	0.24	0.89	0.88
IRS2	0.0	0.3	0.2	0.1	0.1	0.47	0.63	0.46	0.50	0.89	0.55	0.63	0.94
IRS3	0.0	0.0	0.2	0.1	0.1	0.24	0.33	0.16	0.24	0.84	0.25	0.19	0.86
IRS4	0.0	0.0	0.0	0.1	0.1	0.08	0.11	0.04	0.06	0.81	0.05	0.13	0.84

normalized distance performance measure (ndpm). The cumulated gain-based methods rely on the values of relevance and are not sensitive enough to document rank order in general. The average distance measure (adm) relies on the absolute differences of the relevance scores between the system estimation and user estimation, it cannot provide stable evaluation results in some cases.

Second, we compare these methods in terms of giving higher credits to the IR systems for their abilities to retrieve highly relevant documents. This time, we are using a 4-point scale, and there are only five relevant documents to a given query. Let IRS1, IRS2, IRS3, and IRS4 represent four different IR systems, respectively. Their performances for giving high credits to IR systems which can retrieve more highly relevant documents is in a decreasing order as: IRS1, IRS2, IRS3, and IRS4. Table 4 shows the actual evaluation results. The normalized distance performance measure (ndmp) provides the correct results again. All the cumulated gain-based methods except discounted cumulated gain (dcg) and average gain ratio (agr) are able to give the correct evaluation results. The average distance measure (adm) gives higher credit to IRS2 instead of IRS1 because the absolute difference between IRS1 and UR is higher than the absolute difference between IRS2 and UR.

According to the above numerical comparison, we can conclude that in terms of the sensitivity to document rank order and giving higher credits to the IR systems that can retrieve more highly relevant documents, the normalized distance performance measure gives the best evaluation results from both perspectives. The cumulated gain-based methods satisfy the second perspective, but fail in their sensitivities to the document rank order. The average distance measure gives unstable evaluation results for both perspectives.

7 Practical Issues in Using Multi-grade Evaluation

One difficulty with using the multi-grade evaluation is that there are still some practical issues on how to apply these methods. In this section, we discuss some of these critical issues and the possible solutions.

The first issue is how to acquire the user judgments for the ideal ranking. There are two types of rankings required for the computation of multi-grade evaluation function. The system ranking is given by the IR system via assigning weights to each document in the collection. The ideal ranking is supposed to be provided by the user directly and it is more subjective. There are some arguments about whether the judgments should be acquire from the experts of the corresponding field or from randomly selected users with common knowledge. Since most of the IR systems are not designed just for experts, it is fair that the judgments should be given by a group of real users. However, the judgements may vary depending on different users opinions and scenarios in which the judgments are made. The ideal ranking may be produced by merging different user judgments. Rank aggregation methods can be used to combine the rankings given by different users into a new ranked list of result (Borda, 1781; Dwork, Kumar, Naor & Sivakumar, 2001). These methods have been primarily used by meta-search engines. The rank aggregation function is computed by assigning scores to entities in individual rankings or by using orders of the entities in each ranking. In some IR experiments, the ideal ranking is obtained by merging the participating IR system ranking results without the users participation. For example, the Text REtrieval Conference (TREC) uses a pooling method, where each IR system submits their ranked document list (e.g., top 1000 documents), and the ideal ranking is generated by combining these ranking results through an automatic process.

The second issue is that some proposed methods require the user judgments over the entire document collection, in reality, this requirement is usually infeasible. It is important to find out how to use these methods when only partial user judgments are given (Frei & Schauble, 1991; Fuhr, 1989). The early attempt at solving this problem can be found in Cooper's paper (1968), where the expected search length measure indicates the stop point of scanning the entire document list. Nowadays, the general way of solving this problem is to ask the users to provide their judgements on selected samples. These samples could be the top-ranked retrieved documents, or randomly selected documents from the entire collection. In the Text REtrieval Conference (TREC), the document selection is first done by gathering the top 1000 ranked documents retrieved by each participating IR system in a list, and then the top n (e.g., 100) ranked documents of the list are evaluated by the invited experts or users (Voorhees, 2005). However, if there is a relevant document ranked below the 100th position, it will be treated as a non-relevant document in the computation of evaluation functions.

The third issue is how to define the boundaries of different levels of relevance in order to help the users make their judgments. In particular, when a relevance scale contains more than three levels, it is difficult to define the boundaries of the middle levels. For example, in a 4-point relevant scale (highly relevant, relevant, partially relevant and non-relevant), what kind of criteria should be used to differentiate the definition of relevant and partially relevant documents. In IR experiments, users are easily misled to make their judgments due to the poorly defined notion of middle levels of relevance. Some studies have been done with regard to this problem. Spink, Greisdorf and Bateman (1999) discovered 15 criteria used to define middle level relevant documents. Maglaughlin and Sonnenwald (2002) revealed 29 criteria used by participants when

determining the overall relevance of a document. A general agreement is that the more criteria a document satisfies, the higher relevance level it belongs to.

8 Conclusions

Relevance plays an important role in the process of information retrieval system evaluation. In the past, the variability and continuous nature of relevance were paid insufficient attention, and the traditional evaluation methods (e.g., precision and recall) only treat relevance as a two-leveled notion. One important feature of the modern IR system is the large amount of retrieved documents which vastly exceed the number of documents the user is willing to examine. Therefore, it is critical for the evaluation methods to favor those IR systems which can retrieve the most relevant documents and rank them at the top of the output list. This requires the reexamination of the multi-grade feature of relevance and the evaluation methods based on it.

In this paper, we reveal that multi-grade relevance can be formally defined in terms of the user preference relation. The main methodologies of 12 multi-grade evaluation methods, together with some commonly used traditional IR system evaluation methods, are reviewed and compared from different perspectives. Some interesting findings are discovered. We find that most evaluation methods are based on cumulated gain. They are able to give higher credits to IR systems for their abilities to retrieve highly relevant documents, but they are not sensitive enough to document rank order. The average distance measure is not reliable because it uses the absolute difference between system relevance estimation and user relevance estimation. Overall, the normalized distance performance measure provides the best performance in terms of the perspectives we are concerned with in this paper.

A general evaluation strategy based on multi-grade relevance is proposed. Some practical issues and possible solutions are discussed. We find that the evaluation criteria of multi-grade relevance changes compared to the traditional precision and recall. The evaluation strategy based on multi-grade relevance is shifted from the distribution of relevant, non-relevant, retrieved and not retrieved documents to the comparison of system ranking and ideal ranking. The evaluation methods based on multi-grade relevance should be able to credit the IR systems that can retrieve more highly relevant documents, provide better document rank order, and be adaptable to different types of relevance interpretation.

The main contributions of this paper can be summarized as follows. We identify that multi-grade relevance can be formally defined in terms of the user preference relation. We propose a general evaluation strategy based on multi-grade relevance. We recommend that the normalized distance performance measure is a good choice in terms of the perspectives we are concerned with in this paper.

Acknowledgements

The authors are grateful for the financial support from NSERC Canada, constructive comments from professor Zbigniew W. Ras during the ISMIS 2008 conference in Toronto, and for the valuable suggestions from anonymous reviewers.

References

1. Bollmann, P. Wong, S.K.M. (1987). Adaptive linear information retrieval models. *SIGIR*, pp. 157-163.
2. Borda, J.C. (1781). Memoire sur les elections au scrutin. In Histoire de lAcademie Royale des Sciences.
3. 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.
4. Champney, H., and Marshall, H. (1939). Optimal refinement of the rating scale, *Journal of Applied Psychology*, *23*, pp. 323-331.
5. Cleverdon, C. (1962). Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems, Cranfield Coll. of Aeronautics, Cranfield, England.
6. Cleverdon, C., Mills, J., and Keen, M. (1966). Factors dermnining the performance of indexing systems, *Aslib Cranfield Research Project*, Cranfield, UK.
7. Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems, *Journal of the American Society for Information Science*, *19(1)*, pp. 30-41.
8. Cox, E.P. (1980). The optimal number of response alternatives for a scale: a review, *Journal of Marketing Research*, pp. 407-422.
9. Cuadra, C.A., and Katter, R.V. (1967). Experimental studies of relevance judgments: final report, System Development Corp, Santa Monica, CA.
10. Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the web. In WWW 01: Proceedings of the 10th International Conference on World Wide Web, pp. 613622.
11. Eisenberg, M., and Hu, X. (1987). Dichotomous relevance judgments and the evaluation of information systems, in: *Proceeding of the American Scoiety for Information Science, 50th Annual Meeting*, Medford, NJ.
12. Eisenberg, M. (1988). Measuring relevance judgments. *Information Processing and Management*, *24(4)*, pp. 373-389.
13. Fishburn, F.C. (1970). Utility Theory for Decision Making. New York: Wiley.
14. Frei, H.P., and Schsuble, P. (1991). Determine the effectiveness of retrieval algorithms, *Information Processing and Management*, *27*, pp. 153-164.
15. Fuhr, N. (1989). Optimum polynomial retrieval functions based on probability ranking principle. *ACM Transactions on Information System*, it 3, pp. 183-204.
16. Katter, R.V. (1968). The influence of scale form on relevance judgments, *Information Storage and Retrieval*, *4(1)*, pp. 1-11.
17. Kemeny, J.G., Snell, J.L. (1962). *Mathematical Models in the Social Science*. New York: Blaisdell.
18. Kendall, M. (1938). A new measure of rank correlation, *Biometrika*, *30*, pp. 81-89.
19. Kendall, M. (1945). The treatment of ties in rank problems, *Biometrika*, *33*, pp. 239-251.
20. Maglaughlin, K. L., and Sonnenwald, D. H. (2002) User perspectives on relevance criteria: a comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology*, *53(5)*, pp. 327-342.
21. Maron, M.E., and Kuhns, J.L. (1970). On relevance, probabilistic indexing and information retrieval, in *T. Saracevis (Ed.), Introduction to Information Science*, New York: R.R. Bowker Co., pp. 295-311.
22. Myers, J.L., and Arnold D.W. (2003). Research Design and Statistical Analysis. Lawrence Erlbaum.
23. Mizzaro, S. (2001). A new measure of retrieval effectiveness (Or: What's wrong with precision and recall), *International Workshop on Information Retrieval*, pp. 43-52.
24. Jacoby, J., and Matell, M.S. (1971). Three point likert scales are good enough, *Journal of Marketing Research*, *8*, pp. 495-500.
25. Jarvelin, K., and Kekalainen, J. (2000). IR evaluation methods for retrieving highly relevant documents, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
26. Jarvelin, K., and Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques, *ACM Transactions on Information Systems*, *20*, pp. 422-446.
27. Kando, N., Kuriyams, K., and Yoshioka, M. (2001). Information retrieval system evaluation using multi-grade relevance judgments: discussion on averageable single-numbered measures, *JPSJ SIG Notes*, pp. 105-112.

28. Pollack, S.M. (1968). Measures for the comparison of information retrieval system, *American Documentation*, *19(4)*, pp. 387-397.
29. Rasmay, J.O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values, *Psychometrika*, *38(4)*, pp. 513-532.
30. Rees, A.M., and Schultz, D.G. (1967). A field experimental approch to the study of relevance assessments in relation to document searching, Case Western Reserve University, Cleverland, Ohio.
31. Robertson, S.E. (1977). The probability ranking principle, in: *IR Journal of Documentation*, *33(4)*, pp. 294-304.
32. Rocchio, J.J. (1971). Performance indices for document retrieval, in: *G. Salton (Ed.), The SMART Retrieval System-experiments in Automatic Document Processing*, pp. 57-67.
33. Sagara, Y. (2002). Performance measures for ranked output retrieval systems, *Journal of Japan Society of Information and Knowledge*, *12(2)*, pp. 22-36.
34. Sakai, T. (2003). Average gain ratio: a simple retrieval performance measure for evaluation with multiple relevance levels, *Proceedings of ACM SIGIR*, pp. 417-418.
35. Sakai, T. (2004). New performance matrics based on multi-grade relevance: their application to question answering, *NTCIR-4 Proceedings*.
36. Spearman, C. (1904). General intelligence: objectively determined and measured, *American Journal of Psychology*, *15*, pp. 201293.
37. Spink, A., Greisdorf, H., and Bateman, J. (1999). From highly relevant to not relevant: examining different regions of relevance, *Information Processing and Management*, *34(4)*, pp. 599-621.
38. Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika*, *40*, pp. 105-10.
39. Tang, R., Vevea, J.L., and Shaw, W. M. (1999). Towards the identification of optimal number of relevance categories, *Journal of American Society for Information Science (JASIS)*, *50(3)*, pp. 254-264.
40. van Rijsbergen, C.J. (1979). *Information Retrieval*, Butterworth-Heinemann, Newton, MA.
41. Voorhees E.M. (2005). Overview of TREC 2004, in: *Voorhees, E., Buckland, L. (Eds.) Proceedings of the 13th Text Retrieval Conference*, Gaithersburg, MD.
42. Wong, S.K.M., Yao, Y. Y., and Bollmann, P. (1988). Linear structure in information retrieval, in*Proceedings of the 11th Annual Interna- tional ACMSIGIR Conference on Research and Development in In- formation Retrieval*, *2*, pp. 19-232.
43. Wong, S.K.M., Yao, Y. Y. (1990). Query formulation in linear retrieval models. *Journal of the American Society for Information Science*, *41*, pp. 334-341.
44. Yao, Y.Y. (1995). Measuring retrieval effectiveness bsed on user preference of documents, *Journal of the American Society for Information Science*, *46(2)*, pp. 133-145.