IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2022, accepted 15 September 2022, date of publication 3 October 2022, date of current version 10 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3211668

COMMENT

Batch Evaluation Metrics in Information Retrieval: Measures, Scales, and Meaning

ALISTAIR MOFFAT
School of Computing and Information Systems, The University of Melbourne, Melbourne 3010, Australia
e-mail: ammoffat@unimelb.edu.au

This work was supported by the Australian Research Council's Discovery Project DP190101113.

ABSTRACT A sequence of recent papers, including in this journal, has considered the role of measurement scales in information retrieval (IR) experimentation, and presented the argument that (only) uniform-step interval scales should be used. Hence, it has been argued, well-known metrics such as reciprocal rank, expected reciprocal rank, normalized discounted cumulative gain, and average precision, should be either discarded as measurement tools, or adapted so that their metric values lie at uniformly-spaced points on the number line. These papers paint a rather bleak picture of past decades of IR evaluation, at odds with the IR community's overall emphasis on practical experimentation and measurable improvement. Our purpose in this work is to challenge that pessimistic assessment. In particular, we argue that mappings from categorical and ordinal data to sets of points on the number line are valid provided there is an external reason for each target point to have been selected. We first consider the general role of measurement scales, and of categorical, ordinal, interval, ratio, and absolute data collections. In connection with the first two of those categories we also provide examples of the knowledge that is captured and represented by numeric mappings to the real number line. Focusing then on information retrieval, we argue that document rankings are categorical data, and that the role of an effectiveness metric is to provide a single value that summarizes the usefulness to a user or population of users of any given ranking, with usefulness able to be represented as a continuous variable on a ratio scale. That is, we argue that most current IR metrics are well-founded, and, moreover, that those metrics are more meaningful in their current form than in the proposed "intervalized" versions.

INDEX TERMS Information retrieval.

I. INTRODUCTION

Measurement is used to capture data about some attribute or observation of a real-world phenomenon. For example, a thermometer measures temperature so that we have guidance as to how hot or cold it is now; and weather forecasts and climate summaries predict temperatures based on datasets of past measurements so that we can make informed choices about our future activities. All other things being equal, a beach-side holiday at a location with an average daytime temperature of 25°C is likely to be preferable to one at a location with an average daytime temperature of 10°C.

If a measurement is to be useful, it should be connected to the attribute it refers to – that is, be *externally valid* – and allow inferences to be derived from sets of measurements that are both predictive of, and informative about, the reality that is being measured. The extent to which inferences can be derived from measurements is, in part, determined by the *scale* that is employed, defined as the way in which observed attributes are represented by measured surrogates. For example, patients in a hospital may be asked to rate their pain on a scale of one to ten, with "perceived patient pain" the attribute, and a set of ten ordered categories as the corresponding measurement. Analgesics might then in part be prescribed in accordance with the measurement a patient reports, even though patients are not calibrated "pain-o-meters".

Stevens [40] described four *scales of measurement*, and enumerated their *permissible operations* in terms of what can be legitimately concluded about the behavior of the underlying attribute, given knowledge of the measurement, or of a set of measurements. In the case of a "ten point" pain

The associate editor coordinating the review of this manuscript and approving it for publication was Ravibabu Mulaveesala.

assessment, for example, if yesterday a patient said "eight" and today they say "four", we can be confident that their perceived pain has decreased, but should not say that it has halved – the conclusion "has decreased" is permissible, but the conclusion "has halved" is not, and nor is the conclusion "has gone down by four". In the case of ordinal measures (which is what the ten point pain scale is), comparison of measurements is permitted as a valid reflection of the relative values of the corresponding underlying attributes, but the taking of ratios and differences is not (at least, not necessarily). Indeed, a patient's subjective pain scale might be highly nonlinear. Section II provides more detail of Stevens' hierarchy of scales of measurement.

A sequence of recent papers [14], [15], [16] has explored the way in which experimental measurement is carried out in information retrieval (IR), with particular emphasis on the use of *interval scales*, one of Stevens' four scale categories. The somewhat bleak assessment of these papers is that IR has, by and large, employed risky measurement techniques for the last several decades. In the most recent contribution, work that was published in this journal [16], it is argued that the IR community has the opportunity to "*strengthen the foundations of our field*" by "intervalizing" existing metrics in order to render them sound. The key proposal is that raw metric values should be mapped to evenly distributed points on the number line, deriving a corresponding adjusted metric on an equi-interval measurement scale.

Our goal here is to express a more optimistic view of IR evaluation and IR measurement. In particular, we argue that the majority of existing IR metrics are in fact well-founded in terms of measurement and, for the most part, achieve that which they were designed to capture. Central to that argument is the belief that any given metric corresponds to a certain type of user search behavior, and hence may be used (and should only be used) whenever the search task and user cohort can be argued as being a match for the properties of the metric. We also differ in regard to the worth of intervalization as a corrective mechanism.

In Section II Stevens' measurement scale typology is explained, and examples of categorical, ordinal, and interval scales are provided. Section III then focuses on IR evaluation, explains how IR measures are defined, and assesses them against Stevens' hierarchy. The final part of Section III examines and responds to the suggestion [16] that metrics be intervalized. Section IV then considers other related work, and other aspects of IR experimental methodology that researchers and practitioners alike should bear in mind.

II. BACKGROUND

The distinction between *categorical*, *ordinal*, *interval*, and *ratio* measures was introduced by Stevens [40] and is now widely summarized in textbooks and online resources.[1] This section provides an overview of those four classes, starting with categorical and ordinal measures, and examples thereof; then taking a diversion into a parable in which professors' salaries are tabulated; and finishing with a description of interval and ratio measures. Section III then considers the ways in which measures are applied to evaluation in information retrieval.

TABLE 1. Two sets of categorical (or nominal) labels: (a) the standard set of country codes, used perhaps to record country of birth in personnel data; and (b) a set of academic work categories at a university.

Label	Class
AU	Australia
CL	Chile
CN	China
IT	Italy
JP	Japan

(a) Countries of birth

Label	Class
AM	administrative and management
RO	research focused
TO	teaching focused
TR	teaching and research

(b) Academic employment types in a university

A. CATEGORICAL DATA

In a categorical (or nominal) measure each object in the data collection of interest (a *dataset*) is assigned to a single class, with the classes identified via a set of *class labels*. Table 1 provides two examples. In Table 1(a) the classes are *countries*, perhaps reflecting birth locations of university employees, and the class labels are two-letter abbreviations; and in Table 1(b) the classes are *academic work categories* in a university, and the labels are again two-letter surrogates. For example, the birth countries of a group of ten professors at a university might be represented by the dataset:

$$\{\mathsf{AU}, \mathsf{CN}, \mathsf{AU}, \mathsf{IT}, \mathsf{CL}, \mathsf{CL}, \mathsf{CN}, \mathsf{AU}, \mathsf{IT}, \mathsf{JP}\}.$$

The cardinality of each class across the dataset can be tabulated and used to summarize fractions; and the mode (most frequent class) can be reported, as can other class ranks based on frequency of occurrence. For example, in the example dataset, the most common birth country is AU, accounting for 3/10 of the professors; similarly (in the context of Table 1(b)), a university might report that "research focused" RO staff are its second largest category.

In categorical data there is no meaningful ordering between the classes, and the only operations that can be applied to data items are equality and inequality testing ($=$ and $\neq$). The classes themselves can be ordered by considering their labels to be strings, as has been done in the two table sections, but that is purely for presentational convenience, and not an intrinsic feature of the data. If the class labels changed (or even if they didn't) the table rows could be reordered, without affecting the validity or accuracy of any conclusions drawn from the dataset being discussed.

This absence of ordering between the classes means that the median (and other percentiles), and arithmetic and geometric averages, are meaningless concepts – a fact immediately grasped when considering the questions

[1] See, for example, https://en.wikipedia.org/wiki/Level_of_measurement.

TABLE 2. **Four sets of ordinal labels, and for each set, one possible mapping between the set's categories and the real number line: (a) the set of available options (radio buttons) in a survey about frequency of alcohol consumption, with $N(\cdot)$ in "risk points"; (b) the set of professorial ranks at some university, with $N(\cdot)$ in "salary, dollars per week"; (c) the set of recommendations (radio buttons) available to referees during a peer review process, with $N(\cdot)$ a numeric mapping; and (d) the set of relevance labels employed in a document judging process as part of an IR evaluation, with $N(\cdot)$ a gain expressed in units of "utility".**

Label	Class	$N(\cdot)$
A0	never	0
A1	monthly or less	1
A2	two to four times a month	2
A3	two to three times a week	3
A4	four or more times per week	4

(a) Screening for alcohol consumption (one question of several)
Source: https://www.uptodate.com/contents/calculator-alcohol-consumption-screening-audit-questionnaire-in-adults-patient-education

Label	Class	$N(\cdot)$
P1	junior professor	\$700
P2	assistant professor	\$750
P3	associate professor	\$850
P4	full professor	\$1000

(b) Professorial ranks

Label	Class	$N(\cdot)$
SR	strong reject	–3
R	reject	–2
WR	weak reject	–1
WA	weak accept	+1
A	accept	+2
SA	strong accept	+3

(c) Referee evaluations during peer review

Label	Class	$N(\cdot)$
G0	not relevant	0.00
G1	somewhat relevant	0.25
G2	relevant	0.75
G3	highly relevant	1.00

(d) Grades for document relevance to a topic

"average country of birth" and "median work category" of the ten professors mentioned above. Moreover, the inappropriateness of computing medians and means remains even if the class labels "look" like numbers. For example, suppose that the international phone dialing prefixes were used as the class labels, rather than the two letter acronyms: +61 for Australia, +56 for Chile, and so on (ignoring the fact that +1 actually covers several countries). The same dataset of ten professors would now be described as:

{+61, +86, +61, +39, +56, +56, +86, +61, +39, +81},

but it is still *not* permissible to compute the median or mean.

When categorical data is combined into ordered tuples – for example, the pairing "⟨country of birth, work category⟩" – the set of tuples is also categorical data.

B. ORDINAL DATA

Ordinal data classes result if strict inequality as well as equality are permissible operators for comparing data items, that is, if all of $=$, $\neq$, $<$, and $>$ are operational. Table 2 gives four instances of ordinal class labels and class descriptions; the final column headed $N(\cdot)$ will be discussed shortly. The first section of Table 2 is taken from an online "alcohol risk screening" assessment, and is one of a suite of questions that collectively ask about frequency (the one shown), intensity (the amount of alcohol consumed at each session), and impact (physical or emotional damage to self or to relationships with family/friends). There is a clear ordering, with category A0 involving less frequent alcohol consumption than categories A1, A2, and so on.

The fact that the classes are ordered means that cumulative statistics are permissible. For example, a university might report (in reference to Table 2(b)) that 70% of its professors are at level P2 or higher. The same university might also ask students to rate courses via a five-point Likert scale, using the class labels "strongly disagree", "disagree", "neutral", "agree", and "strongly agree" in response to a statement "this course was well taught". It would then be permissible to compute a dissatisfaction score for each course by summing the percentage of "strongly disagree" and "disagree" responses.

The ordered classes mean that it is valid to identify the smallest (*min*) and largest (*max*) item in any dataset, and also permissible to sort a dataset into "order". Medians of ordinal-scale datasets may also be calculated, albeit with a degree of caution. For example, in the dataset {P1, P2, P3, P4} it is unclear what median value should be reported, but inferring it to be "P2.5" via the "arithmetic" (P2+P3)/2 over the two middle points in the sorted arrangement is clearly absurd. The best that can be said in this example is that the median lies in the interval [P2, P3].[2]

In the absence of specific additional information, tuples based on ordinal data must be regarded as being categorical data. Suppose, for example, that a university asked an alcohol screening question of its professors, to create a set of tuples "⟨alcohol consumption, professorial rank⟩". We might feel justified in concluding that ⟨A1, P2⟩ comes "before" the pair ⟨A2, P4⟩, since both dimensions agree on that ordering. But we would have no ability to put ⟨A3, P3⟩ and ⟨A2, P4⟩ into "order"; therefore, the pairs must be categories. This important point will be returned to in Section III when we consider IR evaluation.

C. NUMERICAL TRANSFORMATIONS

When the number of ordinal classes is small (for example, radio-button surveys and Likert scales), the median is a relatively blunt and non-discriminating tool; and an *ordinal to numeric mapping* is often used to transform the class labels into numbers that can be processed arithmetically

[2] In detail: a class label m is a median of a dataset X if half (or more) of the members of X are $\leq m$ and half (or more) of the members of X are $\geq m$. For the dataset {P1, P2, P3, P4} that means that both P2 and P3 are medians; and in the dataset $X = \{$P1, P1, P4, P4$\}$ all four class labels P1, P2, P3, and P4 are medians. It also means that in the numeric dataset $X = \{1, 1, 4, 4\}$ *every* value $1 \leq m \leq 4$ is a median. The use of 2.5 in this numeric case is then a convention that isolates a single value amongst the infinite number of possibilities.

(and perhaps statistically). For example, the five Likert dis/agreement class labels from "strongly disagree" to "strongly agree" might be converted to the numeric values "1" to "5" so that "average agreement" can be computed.

Each section in Table 2 shows one such possible mapping, denoted $N(\cdot)$. For example, in Table 2(a) each answer (to this and each other question in the survey instrument) is assigned a "risk points" value, and a sum is computed over the answers across the set of screening questions, to indicate the extent to which the survey respondent is likely to be affected by alcohol-induced health and social problems. Similarly, in Table 2(c), the sum of the $N(\cdot)$ values over the pool of referees assigned to each paper might be computed, and used as an overall assessment, with negative sums reflecting net "rejection", and positive sums indicating net "acceptance". Despite the apparent ease with which such mappings can be constructed, care is required, and each of those four example mappings is just one instance of an infinite variety of possibilities that could be devised and then argued for.

D. PROCESSING THE PROFESSORIAL PAYROLL

We now turn to a more detailed example involving an ordinal to numeric mapping. Suppose that, in the context of Table 2(b), some university has a total of ten professors, with academic positions given by the dataset:

$$\{\mathsf{P3}, \mathsf{P2}, \mathsf{P4}, \mathsf{P2}, \mathsf{P2}, \mathsf{P2}, \mathsf{P1}, \mathsf{P3}, \mathsf{P4}, \mathsf{P2}\}\,;$$

and that the university wants to include that information into its annual report via a small table:

Class	P1	P2	P3	P4
Count	1	5	2	2

This is legitimate, since it is permitted to tabulate occurrence frequencies in both categorical and ordinal datasets. There is no sense of there being an "average" professor (although many of our students would regard us as being "mean"!); but the median is a permissible statistic, and in this dataset there is no ambiguity, the median is P2. Suppose further that the next table of the annual report lists the current weekly salaries for those four professorial ranks:

Class	P1	P2	P3	P4
Salary	\$700	\$750	\$850	\$1000

This table is thus an ordinal to numeric mapping that allows professorial ranks to be converted to numbers, and hence for the dataset of ten professorial ranks to be mapped to a dataset of ten weekly salaries (all in units of dollars):

$$\{850, 750, 1000, 750, 750, 750, 700, 850, 1000, 750\}\,.$$

Calculation of the median salary over the professors is a legitimate and correct operation; it is clearly \$750 per week. The university could add that statistic to its annual report with a clear conscience.

Now suppose that one of the P2 professors is promoted to level P3 before the annual report is finalized. What becomes of the median salary? For a even-sized set of numbers the usual convention is to take the mid-point between the two middle values (see, for example, Hays [18, Section 4.2]); after the promotion, that computation yields a median of $(\$750 + \$850)/2 = \$800$ per week, or \$50 per week higher than it was previously.

Finally, the university also decides to include the total salary being paid across the set of professors. In a separate work area the annual report's editor prepares this table, to reflect the situation after the successful promotion:

Class	P1	P2	P3	P4
Count	1	4	3	2
Salary	\$700	\$750	\$850	\$1000
Payment	\$700	\$3000	\$2550	\$2000

The editor then sums the bottom row to get a total weekly salary cost of \$8250, finalizes the annual report, and sends it to the printer.

The very first copy arrives on the Provost's desk just a few days later. Worried about the budget, the Provost looks at these various statistics, including the fact that there are ten professors and a total salary cost of \$8250 per week, and concludes that the average weekly salary (since even Provosts can divide by ten in their heads) per professor is \$825. It never crosses the Provost's mind to ponder the fact that the original data was ordinal, and that it was converted to numeric data via a mapping. To the Provost the current professorial salary scale is simply a set of facts that have, at this instant in time, certain fixed values amongst a vast sea of possibilities. In other words, the Provost sees the amounts being paid as an accurate measurement in regard to the attribute "professorial salary payments".

Nor is the Provost concerned by the fact that the average salary value cannot be mapped back to a professorial rank (indeed, perhaps the Provost is a demographer, and hence equally comfortable with the fact that the average couple have 1.93 offspring). If the Provost did want to compute the inverse mapping, the best that can be said is that the average salary-weighted professorial rank lies in the interval [P2, P3]; that is, exactly the same as can be said for the median professorial rank once the promotion has taken place.

It is also perfectly appropriate to apply a *categorical to numeric* mapping to categorical datasets. The university might have an agreed mapping that, for each of the class labels listed in Table 1(b), specifies the workload fraction available for teaching duties:

Class	AM	RO	TO	TR
Teaching fraction	0.20	0.10	0.80	0.35

The dataset of ten "academic work type" categories associated with the same ten professors could thus be mapped to a dataset of ten numeric "teaching fractions"; and then those ten fractions could be summed and averaged to determine, respectively, the total teaching capacity of the university, and the average teaching fraction per employed professor.

E. INTERVAL SCALES

The third level in Stevens' hierarchy corresponds to numeric data for which the difference between any pair of values has meaning, but the values themselves do not necessarily have direct interpretation (or may, but it is somewhat arbitrary). As an example, consider the timestamps employed in the Unix operating system, which are measured in seconds since 1 January 1970, UTC. At the time this sentence was being planned, the "time" was indicated by 1636099886; now that the sentence has been (nearly) typed, the measurement is 1636100081.[3] Each of those two large numbers is, in isolation, somewhat meaningless; but the difference between them has a clear interpretation – the two time measurements were 195 seconds apart. Planning and then composing that one sentence took over three minutes.

Another example is given by the Celsius and Fahrenheit temperature scales. According to one, water freezes at 0°; according to the other, at 32°. Nevertheless, the two scales measure the same underlying attribute: each one degree rise in temperature corresponds to the addition of a fixed amount of thermal energy to a specified volume of water. That fact holds regardless of whether the one degree increase is between 40° and 41°, or between 73° and 74°, provided that either Celsius or Fahrenheit is used for both components of the comparison.

On the other hand when time is measured in "years", it is not an interval scale measurement relative to the underlying attribute of "orbits of the sun". It is a good approximation, and most of us would be willing to say that 2021 is "ten years after" 2011 in the same manner as 2011 is "ten years after" 2001; and are also willing to accept that the cultural basis for selecting the reference year – the beginning of the current monarch's rule; or the birth or death of some historical religious figure – is arbitrary. But the span from 2001 to 2010 inclusive contains 3652 days and 315,532,800 seconds, whereas the span from 2011 to 2020 contains 3653 days and 315,619,200 seconds. More importantly, the span from 2001 to 2010 inclusive contains 9.9988 solar orbits, whereas the period from 2011 to 2020 contains 10.0016 solar orbits. That is, the interpretation attached to intervals measured in years as a surrogate for "solar orbits" differs according to whereabouts in the scale those intervals are taken. Nor are days or seconds linearly translatable into years.[4]

The measurement points of an interval scale that are used in any particular set of measurements or observations are *not* required to be equi-distant on the number line. With the exception already noted, Unix "seconds" are always one second apart, and "days" are normally one "rotation of the earth" apart. But "first day of the month" dates when expressed as Unix timestamps are not at fixed intervals (at least, not in the current Gregorian calendar); similarly, "business days" is a valid interval-based measurement that bypasses two days every seven. In general, measurement points and measured values can be as close to or far apart from each other as is consistent with accurate representation of the underlying attribute that is being recorded and in accordance with the purpose for which it is being measured.

The same flexibility extends to categoric to numeric mappings, and to ordinal to numeric mappings. It is perfectly acceptable for the salary increments between professorial ranks to be of different sizes in Table 2(b); and for the ordinal to numeric mapping $N(\cdot)$ shown in Table 2(c) to make the interval between WA and WR twice as large as the interval between WR and R. In the first case the decision on target values would have been made by the university as a reflection of the cost of attracting and retaining staff of the required caliber; in the second, the decision on those relative intervals would have been made by the PC Chairs for the conference in question, based on their experience of referee behavior and the outcomes they sought via the paper review process. Once established, any such mapping allows the class labels to be converted to numbers, and for differences to be computed and compared. The fact that in Table 2(c) the target value 0 is not generated by any of the six label options is of no concern; -0.5 and $+5$ are not amongst the mapped targets either. Nor is \$925 a salary level that is available in Table 2(b).

Datasets based on interval scales allow translation operators ($mx + c$, where $m > 0$ and c are constants) to be applied without affecting relativities, and, as noted, for differences between measurements, and ratios of differences between measurements, to be compared; but not ratios of measurements themselves. Consider the ordinal to numeric payroll mapping shown in Table 2(b). That mapping means that it is valid to both compute differences and also to attribute meaning to the ratios between differences. For example, $N(\text{P4}) - N(\text{P3}) = 1.5 \times (N(\text{P3}) - N(\text{P2}))$, and it is evident that promotion to P4 from P3 results in a pay-rise that is 1.5 times larger than the pay-rise that our friend received earlier when they were awarded their promotion to P3.

Ratios between intervals defined by one scale might not correspond to comparable intervals on a different scale that represents a different underlying attribute. For example, the P1 professor who gets promoted to P2 might gain the same added *utility* from their modest \$50 pay-rise as a P2 professor gains when promoted to P3 and receives a \$100 increase in their weekly pay. If we wish to map professorial classes to perceived utility of income we are measuring a different underlying attribute, and should use a different ordinal to numeric mapping. On the other hand, Celsius and Fahrenheit do measure the same underlying attribute, and one scale is thus a translation of the other.

When the measurements are made on a continuous scale the values in the dataset might have varying degrees of precision. We can count weekly salaries down to the cent or even sub-cent level if we wish to, or stick to whole dollars,

[3] Both obtained from https://www.unixtimestamp.com/index.php.

[4] Strictly speaking, nor is Unix time an interval scale, because international time-keepers insert the occasional "leap second" too, most recently making 31 December 2016 one second longer than the 86,400 = 24 × 60 × 60 seconds of a standard day, see https://en.wikipedia.org/wiki/Leap_second. Unix timestamps assume that there are always exactly 86,400 seconds every day; and at any leap-second boundaries, that extra second is achieved by subtracting one from the operating system's internal time variable, and observing the same second a second time.

or have a mixture. Similarly, temperatures might be expressed as integers sometimes, or to three decimal places at others; and a landscape gardener planning a paling fence measures their 50 meters to less precision than does the builder of the swimming pool for an upcoming Olympics. This is not an issue. The requirement for an interval scale is *solely* that taking differences must always yield values that can be compared to each other as ratios and hence be assigned meaning relative to the underlying attribute; and that those interpretations must be invariant with respect to *where* in the scale the differences arise.

Monetary amounts, distances, weights, and so on, all result in datasets that have interval scale properties. A 40 kilogram weight is heavier than a 30 kilogram weight by exactly the same 10 kilogram difference as a 25 kilogram weight is heavier than a 15 kilogram weight. And the last kilometer of a cycling race is exactly the same length as the first kilometer, regardless of how long the race is. That final kilometer might require more mental resilience than the first, and it might require more muscle energy production too – both of which are underlying attributes that are not ‘‘distance’’, and hence cannot be measured in units of kilometers – but it will certainly be one kilometer long.

When a dataset is presented on an interval scale (or following the process of mapping a categorical or ordinal dataset to obtain a derived interval-scale dataset), all of the operations permissible on ordinal-scale measurements are again permissible. In addition it is also permissible to compute the arithmetic mean (average). As a geometric interpretation, the arithmetic mean is the point $\bar{p}$ at which the sum of the signed differences $p_i - \bar{p}$ for the elements p_i in the dataset is zero, confirming that the relationship between the arithmetic mean and the elements in the dataset is invariant to the possible arbitrariness of the origin point and multiplicative scale of the measurements.

F. RATIO SCALES

When data is measured using a ratio scale, the data elements themselves have meaning, as well as their differences; and the ratio between data elements is a permissible computation that has the same interpretation across the measurement scale. For example, weight measured in kilograms is a ratio scale, with 20 kilograms twice as heavy as 10 kilograms in the same way that 50 kilograms is twice as heavy as 25 kilograms, and in the same way that 44.092 pounds (that is, 20 kilograms) is twice as heavy as 22.046 pounds. Consistency of ratios means that the zero point of the scale is no longer arbitrary, and that it must be in a single unique location for all ways of measuring that underlying attribute.

All of weight (in kilograms), distance (in kilometers), money paid as salary (in dollars), and temperature (in degrees Kelvin, but not in degrees Celsius or Fahrenheit) are ratio scales. Plus, if for some reason we are specifically interested in time since 1 January 1970, then Unix timestamps are a ratio scale, with 50000000 being twice as distant from 1 January 1970 as is 25000000. On the other hand, if we have no reason to attach significance to 1 January 1970, then Unix timestamps are (only) an interval scale. Similarly, the referee score mapping shown in Table 2(c) is not a ratio scale.

G. ABSOLUTE MEASUREMENTS

If the attribute that is being measured is one that can be directly quantified, then that value can be used as the measurement without further transformation. For example, ‘‘number of children’’ is an absolute attribute (taking on values zero, one, two, and so on) that does not require that ‘‘units’’ be specified – compare with length measured in centimeters or inches or yards or meters (or light-years or parsecs). This category is not included in Stevens’ taxonomy, but for completeness it makes sense to note it here.

H. INTERPRETABLE OUTCOMES

If the observer has complete freedom to choose an ordinal to numeric mapping, then little interpretation can be placed on any computed attributes, such as the mean. We could get a different outcome by choosing a different mapping. But if the ordinal to numeric mapping is defined by the context in which the dataset was created, and is bound to a set of target values by some external reality – as was the case, for example, with the professors’ salaries – then that factual relationship makes the mapping’s values meaningful, and hence interpretable in terms of the attribute from which the measurement was derived. In the main example of this section, the Provost knew the total cost of the ten professors because their salaries were defined via an agreed and published mapping that could be summed, a real-world consequence of the professorial ranks.

Values that are derived via some mapping can only be interpreted *in the context of that mapping*, and if the mapping changes, so too will the derived values, and perhaps even the relativities. If the ten professors decamp and move en masse to another university (while retaining their current ranks), they are likely to be subject to a different set of salaries. If so, a different ordinal to numerical mapping will apply, and after calculating their average salary relative to that university’s pay scales, their new Provost might reach a different conclusion about the average salary-weighted professorial rank.

Similarly, if the conference PC chairs used a different mapping from referee acceptance grades to numbers then the submitted papers might get sorted into a different overall ‘‘average paper score’’ ordering, and a different set of papers might be accepted. But provided the mapping is defined by the PC chairs in advance, and is based on principles that they believe can be successfully argued, then calculating average referee scores is defensible, even though the relationship between the mean of a mapping-derived dataset and the mapping’s set of target values is not required to be invariant to mapping changes. This cause-and-effect nexus between mapping and conclusions is both normal and acceptable, and provided the mapping has its basis in the real world attribute that is the focus of the measurement, should not be regarded as being proscribed in any way.

I. NUMBERS DON'T REMEMBER

In a parable involving "football numbers", Lord [21] observes (giving an opinion via the voice of the "statistician" in the story) that: "*Since the numbers don't remember where they came from, they always behave just the same way, regardless*". This statement has provoked extensive commentary, both in support and in opposition; with Scholten and Borsboom [37] giving one of the more recent – and also more insightful – analyses.

The complementary argument made in this section is that if you *do* know where the numbers came from and why they have the values that they do, and are confident that those values can be justified in reference to the real world attribute that the mapping is designed to represent, then those numbers may be used in your analysis and interpretation of that real world equivalence. The next section applies that principle to effectiveness measurement in information retrieval.

III. MEASUREMENT IN INFORMATION RETRIEVAL

A. SEARCH ENGINE RESULT PAGES

Evaluation in *batch information retrieval* (also sometimes referred to as *offline evaluation*) centers on *search engine result pages*, or SERPs, see Sanderson [35] for an overview. In simplest (and highly stylized) form a SERP is an ordered permutation of the n documents in the collection managed by the search service, or a k-element prefix of such a permutation; and is the visible output that is presented to a user in response to a query. Each item in the SERP is either a document, or a surrogate summary of a document referred to as a *snippet*. In most IR batch evaluation methodologies (but not all) users who examine a snippet in a SERP are regarded as having also examined the document behind the snippet, and we will continue with that assumption here.

The interfaces provided by commercial search systems present much richer interfaces, containing a complex amalgam of elements including query suggestions, images, knowledge panels, extracted answers, and so on; and require equally sophisticated assessment techniques. But stylized SERPs of the form we consider here continue to be the mainstay of IR evaluation when retrieval similarity formulations and ranking models are being developed, and are the focus of both the arguments put by Ferrante et al. [14], [15], [16] and our response to those arguments.

The primary underlying attribute in IR evaluation is the *usefulness* of a SERP in terms of how well it addresses the information need that provoked the user's query. To that end, "usefulness" can be defined either as a combination of *correctness*, *coverage*, *comprehensiveness* and *cost of consumption* of the information conveyed by the SERP's documents, or in terms of the user's *satisfaction* after they have consumed the SERP. Comparative evaluations based on usefulness are then used to determine which search service, or parameter settings within a single service, give rise to the SERPs with the greatest usefulness.

Search result pages are categorical data, because they are ordered n-tuples (or ordered k-tuples) of documents, which are themselves categorical. A further standard assumption is that the individual documents making up the SERP can each be assigned a per-document value known as *relevance*, corresponding to their in-isolation usefulness in response to the query. In most experimental contexts document relevance is represented on an ordinal scale, based on *relevance grades*. The simplest possible measurement scale is a binary one, with labels "G0" meaning "non-relevant" and "G1" meaning "relevant". Graded relevance scales make use of more classes, see the example already shown in Table 2(d). Relevance scales based on arbitrary numeric values are also possible, and do not alter the arguments presented here.

Given that each document in a SERP can be assigned a relevance grade, SERPs can be assigned to categorical classes based on their ordered sequences of n (or $k < n$) relevance grades. For example, a five item SERP provided in response to a query might be a member of the class $\langle \mathsf{G1}, \mathsf{G3}, \mathsf{G0}, \mathsf{G0}, \mathsf{G2} \rangle$, with the five ordinal document relevance classes as defined in Table 2(d).

B. COUNTING AND ORDERING SERPs

Even in the simplest case, with binary document relevance classes, the number of SERP classes is huge. If n_0 is the number of G0 non-relevant labels across the n documents, and n_1 is the number of G1 relevant labels, then there are a total of $n!/(n_0!n_1!)$ different SERP classes. Even when a k-element prefix of the SERP is taken, with $k \leq \min\{n_0, n_1\}$, there are still 2^k different SERPs.

It was noted above that SERPs are categorical data; nevertheless, some SERP relativities can be derived from the ordering embedded in the document relevance scale. For example, when the five-element SERP

$$\langle \mathsf{G1}, \mathsf{G3}, \mathsf{G0}, \mathsf{G0}, \mathsf{G2} \rangle \,,$$

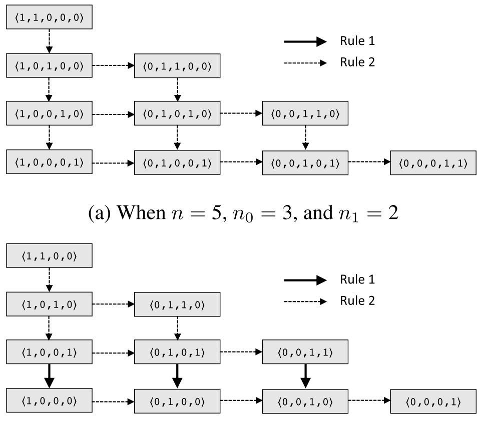
is compared with the SERP

$$\langle \mathsf{G1}, \mathsf{G2}, \mathsf{G0}, \mathsf{G0}, \mathsf{G1} \rangle \,,$$

it is apparent that the second one cannot possess more usefulness than the first one, as it is less relevant in two document positions, and equal in the other three dimensions. More generally, we can be confident that SERP S1 is *non-inferior* to SERP S2 (denoted S1 $\succeq$ S2) by considering two monotonicity relationships:

- *Rule 1*: SERP S1 is non-inferior to SERP S2 if every element of S1 is greater than or equal to the corresponding element of S2 in terms of their ordinal document relevance labels;
- *Rule 2*: SERP S1 is non-inferior to SERP S2 if S2 can be formed as a transformation of S1 in which one or more elements are swapped rightwards and exchanged with elements of strictly lower document relevance that move leftward.

Rule 1 is an absolute relationship that does not rely on the documents in the SERP being examined in a top-down manner (corresponding to left-to-right in the examples



(a) When $n = 5$, $n_0 = 3$, and $n_1 = 2$

(b) When $n = 5$, $n_0 = 3$, $n_1 = 2$, and $k = 4$

FIGURE 1. Hasse diagrams showing: (a) all SERPs of $n = 5$ documents in which there are $n_0 = 3$ non-relevant and $n_1 = 2$ relevant documents; and (b) the set of $k = 4$ prefixes of those SERPs. The solid arrows indicate "Rule 1" $\succeq$ relationships, and the dotted arrows show "Rule 2" $\succeq$ relationships. Two relevance grades are assumed, with $1 > 0$.

employed here). Rule 2 arises from adding the assumption that the SERP is examined sequentially from left-to-right, but is still not equivalent to lexicographic ordering.

Figure 1(a) shows the set of $\succeq$ relationships when SERPs of $n = 5$ binary document relevance grades with $n_0 = 3$ and $n_1 = 2$ are considered, with "0" and "1" used as shorthand for the document relevance labels G0 and G1 respectively. No Rule 1 relationships are possible when all n documents are included, because the SERPs must be permutations of each other. Figure 1(b) then shows the non-inferiorities that arise when binary SERPs over $n = 5$ documents (still with $n_0 = 3$ and $n_1 = 2$) are truncated at $k = 4$ for the purposes of evaluation. Now Rule 1 relativities also occur.

All of the relationships shown in Figure 1 are transitive, meaning that SERP pairs that are not linked by a directed path of arrows are incomparable. In both the $n = 5$ case and the $k = 4$ version the two axiomatic rules are insufficient to impose a preference ordering between $\langle 1, 0, 0, 1, 0\rangle$ and $\langle 0, 1, 1, 0, 0\rangle$. That is, in the absence of any further information in regard to what it is that users find to be useful, either of these two SERPs might be preferred.

C. IR EFFECTIVENESS METRICS

Given that context, an *effectiveness metric* is a categorical to numeric mapping that assigns a real-valued number to each possible class of SERP. Those derived values are often, but not always, in the range $0 \ldots 1$. For example, assuming binary document relevance categories with class labels G0 and G1, the simple metric "precision at k" (Prec@k) is computed as the number of relevant (G1) documents among the first k in the SERP, divided by k. Both $\langle 1, 0, 0, 1, 0\rangle$ and $\langle 0, 1, 1, 0, 0\rangle$ have Prec@5 scores of 0.4, and both have Prec@4 scores of 0.5. But $\langle 0, 1, 1, 0, 0\rangle$ is deemed to be better than $\langle 1, 0, 0, 1, 0\rangle$ according to Prec@3. Moreover, there is a real-world interpretation of Prec@k that can justify its use as a metric: if we suppose that each user of the search system examines exactly the first k documents in each SERP they receive, then Prec@k measures the fraction of documents viewed by the user that are relevant. That is, there is a clear connection between Prec@k and an aspect of the real-world situation that can be argued as being a way of assessing SERP usefulness, at least for some category of users.

A wide range of other effectiveness metrics have been proposed to augment Prec@k, including top-weighted ones that allocate decreasing importance to documents the further they are from the head of the SERP. For example, the metric RR is defined (again, for binary relevance grades) as the reciprocal of the index of the first position in the SERP that contains a G1 document. The same two SERPs $\langle 1, 0, 0, 1, 0\rangle$ and $\langle 0, 1, 1, 0, 0\rangle$ thus have RR values of 1.0 and 0.5 respectively.

With these definitions, an RR value of 0.25 is possible, but an RR value of 0.75 is not. Similarly, a Prec@3 value of 0.5 not possible, nor a Prec@5 value of 0.35. Should we be concerned by these absences? Fuhr [17] and the sequence of papers noted earlier [14], [15], [16] argue that in the case of RR we most definitely should. In particular, the first of Fuhr's list of ten "*common mistakes*" is "*Thou shalt not compute MRR*" (with MRR being "mean reciprocal rank"), a directive justified by "*... the difference* [in score] *between ranks 1 and 2 is the same as that between ranks 2 and* ∞*. This means that RR is not an interval scale, it is only an ordinal scale.*" We disagree with that conclusion.

D. SALARIES FOR SERPs

To develop the professorial salary levels listed in Table 2(b) the university in question may have engaged remuneration consultants and asked them to undertake a comparative study of current real-world salary expectations for professors of certain specified abilities. The university knows it has to offer competitive salaries if it is to retain staff, but under the ever-watchful eye of the Provost, doesn't want to pay too much. That is, we can assume that the correspondences listed in Table 2(b) have been determined to be "market rates" in some way, and have been drawn from a larger set of initial possibilities that were considered as options. The intervals between the salary points are meaningful; they represent salary differentials that must be paid in a competitive market, measured in dollars.

Suppose that a search engine company – "AmaBaiBinGoo", perhaps – undertakes a similar market rates study. They run surveys, host focus groups, meet with psychologists, and sponsor IR-related conferences; and conclude from their investigations that the great majority of AmaBaiBinGoo users fall into a "shallow-hasty-youthful" demographic that is highly focused on getting a single correct result for each of their queries. The study participants were also asked to estimate the monetary value of example SERPs, and out of an immense amount of data a set of correspondences between SERP classes (that is, SERP categories constructed using binary document relevance grades G0 and G1, as shown in Figure 1) and perceived values emerges:

SERP group	T1	T2	T3	T4	$\cdots$
Value	1.00c	0.50c	0.33c	0.25c	$\cdots$

where the group T1 contains all SERPs that commence with a relevant document, $\langle$G1, ...$\rangle$; group T2 contains all SERPs that have their first relevant document in the second position and commence with $\langle$G0, G1, ...$\rangle$; group T3 contains all SERPs that commence with two non-relevant documents and then have a G1-grade document in third position, $\langle$G0, G0, G1, ...$\rangle$; and so on. These grouped SERP categories – from T1 onward – form an ordinal arrangement, because of the positional references to "first" and "second", but the groups can also still be thought of as categorical labels.

The company's chief financial officer (CFO) takes great interest in this data. To estimate the possible income should AmaBaiBinGoo move to a user-pays income model, the CFO assembles a sample of ten recent queries and the SERPs that were returned for them, and constructs this dataset of SERP groups:

$$\{\text{T1}, \text{T3}, \text{T1}, \text{T4}, \text{T3}, \text{T1}, \text{T1}, \text{T1}, \text{T2}, \text{T3}\}.$$

The mode of this dataset is T1; in an ordinal sense the median is either T1 or T2; and it is meaningless to ask about the "average SERP category". But the AmaBaiBinGoo CFO continues, joining the per-SERP revenue estimates from the market research to the sample SERP distribution:

SERP group	T1	T2	T3	T4
Count	5	1	3	1
Revenue each	1.00c	0.50c	0.33c	0.25c
Income	5.00c	0.50c	1.00c	0.25c

Summing the bottom row tells them that their current revenue expectation from a user-pays model would be 6.75c for this sample of ten queries, or 0.675c per query on average.

Now the critical question arises: is the computation of the average payment per query using this framework a valid computation? We argue that it is, and that it is meaningful in exactly the same way that the average professorial salary is a meaningful value. Both the professorial salaries and the per-SERP payments are on the interval scale of "money", and by design (and expenditure on consultant fees) reflect their respective real-world situations. Hence, "mean value per SERP" is a valid measurement of search according to its underlying attribute – the usefulness of SERPs to users – provided only that we are willing to equate the attribute of usefulness and the attribute of value. But that equivalence is one of the underpinning assumptions of economics: that the price that someone is willing to pay for a service reflects the utility (that is, usefulness) that they expect to derive from it.

E. DIFFERENT USER BEHAVIORS

The reader will doubtless have noted that the SERP "pricing" mechanism used in that previous example corresponds to Reciprocal Rank, RR. What if AmaBaiBinGoo's market research also noted other factors that influence the amount that a user is willing to pay, in addition to the position of the first relevant result in the SERP? For example, suppose that an even more detailed user evaluation (and, who knows, perhaps a deep convoluted neural model as well) reveals that customers are willing to pay 0.50c if the first document in each SERP is relevant; plus (independently) 0.25c if the second is relevant; plus another 0.125c if the third is relevant; and so on; adding up the payments right through the length of the SERP. The "RBP$_{0.5}$" column in Table 3 shows the ten different per-SERP values that can arise when this computation is applied to the set of ten $n = 5$, $n_1 = 2$ SERPs shown earlier in Figure 1, and places those derived scores beside the corresponding RR values. This new mapping function yields the effectiveness metric *rank-biased precision* (RBP) with parameter $\phi = 0.5$ [25].

Both RR and RBP can be applied to the CFO's dataset of ten SERPs, with different average scores emerging. Those two averages must not be compared to each other, because they were computed using different mappings and hence different assumptions about value. That is, the two metrics are incompatible. So, while they will be correlated for certain SERP subsets – because of the axiomatic chains depicted in Figure 1 – they are not convertible, and represent different measurements. Nevertheless, both the RR and RBP$_{0.5}$ averages over a dataset of categorical SERPs are valid computations in the context of the numeric mappings that were employed when computing them. What the CFO must do is decide which mapping best captures the value of each SERP class to the members of their user base, that is, which context they believe is the most realistic assessment of value as a surrogate for the underlying attribute of SERP usefulness.

F. USERS, MODELS, AND METRICS

There are many other possible effectiveness metrics, and more are proposed each year. Table 3 adds two further options to the three that have already been mentioned: *average precision*, AP [5], [28]; and *normalized discounted cumulative gain*, NDCG [19]. Each of the five metrics shown is a distinct categorical to numeric mapping, with the numeric targets representing perceived utility, expressed in units of "willing to pay this many cents for a SERP in this category", and in which "cents" is an imaginary currency that nevertheless has a fixed multiplicative exchange rate that allows conversion to Euros, to USD, to JPY, to RMB, and so on, just as inches can be converted to parsecs.

As a separate comment, AP and NDCG are computed somewhat differently to the three already described. They involve a "normalization" step that adjusts the score (payment) associated with each SERP according the maximum

TABLE 3. All possible SERPs composed of $n = 5$ binary document relevance grades containing $n_1 = 2$ relevant and $n_0 = 3$ non-relevant documents. The metric Prec@5 is 0.4 for all ten SERPs. There are no violations of the $\succeq$ relationships captured by the arrows in Figure 1.

SERP	Evaluated at $n = 5$				$k = 4$
	RR	RBP$_{0.5}$	AP	NDCG	Prec@4
$\langle 1,1,0,0,0 \rangle$	1.000	0.750	1.000	1.000	0.500
$\langle 1,0,1,0,0 \rangle$	1.000	0.625	0.833	0.920	0.500
$\langle 1,0,0,1,0 \rangle$	1.000	0.563	0.750	0.877	0.500
$\langle 1,0,0,0,1 \rangle$	1.000	0.531	0.700	0.850	0.250
$\langle 0,1,1,0,0 \rangle$	0.500	0.375	0.583	0.693	0.500
$\langle 0,1,0,1,0 \rangle$	0.500	0.313	0.500	0.651	0.500
$\langle 0,1,0,0,1 \rangle$	0.500	0.281	0.450	0.624	0.250
$\langle 0,0,1,1,0 \rangle$	0.333	0.188	0.417	0.571	0.500
$\langle 0,0,1,0,1 \rangle$	0.333	0.156	0.367	0.544	0.250
$\langle 0,0,0,1,1 \rangle$	0.250	0.094	0.325	0.501	0.250

amount of relevance available across the collection (in the binary examples used here, expressed by the value n_1), adding an implication that users are willing to pay increased amounts if relevant documents are relatively scarce, but equally implying that users are somehow aware of the scarcity or not of relevant documents in regard to each query they issue. Note also that all of RR, RBP, AP, and NDCG are top-weighted, meaning that if they are evaluated across the whole collection (that is, on full-length SERPs of length n rather than at-k truncated ones), Rule 2, noted above, results in strict superiority ($\succ$), rather than non-inferiority ($\succeq$).

More generally, most IR metrics have a corresponding *user browsing model*, which hypothesizes the way in which users interact with each SERP, and the subconscious process they follow as they consume SERPs and assess usefulness – the attribute that we are trying to measure. Thus, one way in which IR effectiveness metrics have been studied is via the development of user browsing models of increasing sophistication [2], [7], [9], [26], [27], [48], [50]. Each such model maps a categorical SERP to a numeric assessment of that SERP's value on the real number line, usually between 0.0 and 1.0 inclusive, often in units of "expected utility gained per document inspected", using the corresponding browsing model as a guide to the manner in which the user consumes, and ends their consumption of, the SERP.

This is why there have been so many IR metrics proposed – each corresponds to a different interpretation of "SERP usefulness", and hence corresponds to a different category of user, or a different type of search even when being carried out by the same type of user. The AmaBaiBinGoo "shallow-hasty-youthful" demographic was mentioned earlier; similarly, another company's users might belong to a "thoughtful-patient-older" demographic. An IR metric that accurately assesses usefulness for one of these cohorts may not match the other community's interpretation of usefulness, and vice versa. Conversely, if a user model describes the behavior of some demographic group when carrying out some type of information-based search task, then the corresponding dual metric will be well-suited to measuring perceived usefulness in that same specific context of cohort and search task.

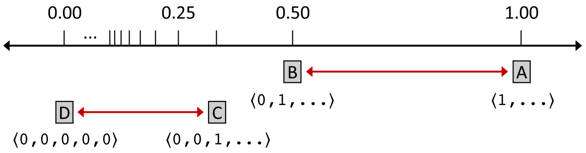


FIGURE 2. Four example SERPs and their RR@5 scores, shown against the real number line. The requirement for an interval scale is that ratios of differences have meaning. In this example that is achieved if users perceive SERP A as improving upon SERP B by 1.5 times as much as they perceive SERP C as improving upon SERP D.

To make this argument concrete, consider again the fundamental basis on which Fuhr [17] and Ferrante et al. [14], [15], [16] criticize RR, headlined by this pattern of scores:

$$\begin{gathered}\text{first relevant document at rank one, RR} = 1.0\\ \downarrow \quad -0.5\\ \text{first relevant document at rank two, RR} = 0.5\\ \downarrow \quad -0.5\\ \text{no relevant document in top } k\text{, RR} = 0.0.\end{gathered}$$

In criticizing RR for this behavior, those authors overlook the possibility that there may be a community of users *wanting those two intervals to be of equal importance*. If some cohort of users share a perception of usefulness that concurs with the relationship between those three classes of SERP, then the non-uniformity of the other available measurement points is of no concern whatsoever, and RR is indeed measuring SERP usefulness.

Figure 2 further explores this relationship between user-perceived SERP usefulness and measured SERP score. In the example four different SERPs are being assessed, and have RR scores (left-to-right, D to A) of 0, 1/3, 1/2, and 1. If we can agree that it is conceivable that a category of users might view SERP A as improving upon SERP B by 1.5 times as much as SERP C improves upon SERP D (and similar for the other ratios of differences that might arise, for example, with A regarded as improving upon D by six times as much as B improves upon C) then there can be no ambiguity: *RR is an interval scale measurement for those users*.

This then makes clear our fundamental concern, and brings us to the key point of this work: *any argument that RR – or* *any other metric – is an unsuitable categorical to numeric mapping for measuring IR system effectiveness for some cohort of users or some type of search task must be justified based on rhetoric about user perceptions of SERP usefulness, or on observational data that measures SERP usefulness via some agreed surrogate. Arguments against IR effectiveness metrics cannot be based solely upon statements about the non-uniformity of the intervals between the available measurement points.*

G. INTERVALIZATION OF METRICS

Ferrante et al. [16, page 136193, in connection with their Figure 3] write that "*the real problem with IR evaluation measures is that their scores are not equi-spaced and thus they cannot be interval scales*". This assertion leads to their proposal that existing metrics be *intervalized*, by enumerating all possible metric values over truncated SERPs of some defined length ($k = 10$, or $k = 20$ say, but certainly not $k = 100$, because of combinatorial growth issues) and then mapping the ordering implied by those values to a uniform-interval scale to get new versions of those metrics.

To understand the process of intervalization, consider the metric NDCG, already illustrated in Table 3. If we assume that $n_0, n_1 \geq 3$ then there are $2^k = 8$ different binary-grade SERPs possible of length $k = 3$, with NDCG@3 scores (sorted by score, to three decimal places) of

$$\{0.000, 0.235, 0.296, 0.469, 0.531, 0.704, 0.765, 1.000\}.$$

That set of eight irregularly-spaced NDCG@3 scores would be intervalized to the range $[0, 1]$ via the corresponding uniformly-spaced set of eight target values (all multiples of 1/7, again represented to three decimal places)

$$\{0.000, 0.142, 0.285, 0.428, 0.571, 0.714, 0.857, 1.000\}.$$

The mapped uniform-interval values would then be used to compute means and as a basis for comparing systems, and to undertake statistical tests, as a derived variant of NDCG@3. Similarly, for the metric NDCG@10, a set of 1024 mapped NDCG values would be generated, at uniform intervals of 1/1023.[5]

We believe that intervalization should regarded with scepticism. There is no requirement in Steven's typology that interval scales be restricted to uniform distances between the available measurement points; the requirement is simply that the ratio between pairs of intervals be indicative of the corresponding difference in the underlying attribute. Moreover, altering the categorical to numeric mapping used to assign score to SERPs changes the relativities being measured, and thus affects the outcome of any subsequent arithmetic. This effect is especially notable for the metric RR. If truncated rankings of length k are used, mapping to an equi-spaced scale yields:

$$\begin{gathered}\text{first relevant at rank one}\\ \downarrow\ -1/k\\ \text{first relevant document at rank two}\\ \downarrow\ -1/k\\ \text{first relevant at rank three}\\ \cdots\\ \text{first relevant document at rank } k-1\\ \downarrow\ -1/k\\ \text{first relevant document at rank } k\\ \downarrow\ -1/k\\ \text{no relevant document in the first } k\end{gathered}$$

which is logically equivalent to using the rank of the first relevant document as the assessment of SERP usefulness – let's call it the metric R1, sometimes referred to as "expected search length" [10]. While that is a perfectly valid measure, it probably isn't a plausible way of measuring the underlying attribute of SERP usefulness. Would a user of an IR system really perceive having the first relevant document at rank 100 rather than at rank 99 as being the same amount less useful as is having the first relevant document at rank 2 rather than at rank 1? Indeed, R1 is sufficiently obvious as a possible metric that if it were reflective of user perceptions of usefulness, then it would have been in common use in IR evaluation for the last several decades. There has been a reason why R1 has had almost no use as a measure of SERP quality – because there are no compelling combinations of user cohort and search task for which it reflects SERP usefulness.

Finally, note that even $\text{RBP}_{0.5}$ does not guarantee a uniform-interval scale, adding further confusion to the situation. Fuhr [17] and Ferrante et al. [14], [15], [16] suggest that $\text{RBP}_{0.5}$ is an example of a valid (by their requirements) interval scale IR metric, because if the SERPs being evaluated are of length k, then all multiples of 2^{-k} between 0 and $1 - 2^{-k}$ can be generated as metric scores, and hence the measurement points are at uniform intervals. But that conclusion is only correct when $k \leq \min\{n_0, n_1\}$. When complete SERPs to depth n are scored via $\text{RBP}_{0.5}$, or when truncated SERPs are scored and $n_1 < k$ (a situation that is by no means improbable), the available set of $\text{RBP}_{0.5}$ values is not uniform-interval – as is illustrated in Table 3, comparing the sequence 0.565, 0.531, and 0.375, for example. Indeed, when there is a single relevant document for some query (a navigational query [4], with $n_1 = 1$), $\text{RBP}_{0.5}$ evaluated to depth k gives this pattern of scores:

$$\begin{gathered}\text{relevant at rank one, } \text{RBP}_{0.5} = 0.5\\ \downarrow\ -0.25\\ \text{first relevant document at rank two, } \text{RBP}_{0.5} = 0.25\\ \downarrow\ -0.25\\ \text{no relevant document in the first } k\text{, } \text{RBP}_{0.5} = 0,\end{gathered}$$

[5] Note that here we make use of the "Microsoft" version of NDCG, in which the discount at rank d is $\log_2(1 + d)$ for all $d \geq 1$, whereas the examples provided by Ferrante et al. [16] use the original Kekäläinen [19] parameterized discount in which ranks $d \leq b$ have a discount of 1.0, and ranks $d \geq b$ have a discount of $\log_b d$. In the Kekäläinen [19] implementation, there are 768 distinct NDCG@10 values possible, and hence the intervalized version of this metric would use a uniform interval of 1/767.

exactly matching (with multiplication by two a permissible operation) the RR intervals over the same three k-truncated SERP classes objected to by Fuhr [17].

IV. OTHER CONSIDERATIONS

A. RELATED WORK

Fuhr's exposition [17] addressing experimental protocols in IR has also been commented on by Sakai [32] who, amongst other concerns, writes (with acknowledgment to input from Stephen Robertson): "*it is also not clear to me whether RR really cannot be considered as an interval-scale measure*", and specifically questions the RR example given by Fuhr (first relevant document at rank one versus first relevant document at rank two versus no relevant document at all) and asks why this cannot ever be congruent with the user's perception of SERP usefulness, thereby anticipating our own concerns. Sakai [32] goes on to present agreement rates between human assessors and effectiveness metrics in regard to SERP quality that summarize the results of an experiment by Sakai and Zeng [33] that compared SERPs in a side by side manner and elicited preferences as to overall usefulness (in this case, via the question "Overall, which SERP is more relevant to the query?"). It is experiments such as these that will establish which effectiveness metrics best correlate with user perceptions of SERP usefulness for various search applications and different user cohorts [34]. Similarly, consideration of perceived user experience is what has driven much of the recent development of effectiveness metrics – see Moffat et al. [26], Zhang et al. [50], Azzopardi et al. [2], Thomas et al. [41], and Moffat et al. [27], for example.

There has also been followup commentary in regard to Stevens' original paper [40] about scales of measurement. The contribution by Lord [21] has already been noted; amongst many others the evaluations by Townsend and Ashby [42] and Velleman andWilkinson [47] also help delineate some of the issues that have emerged when considering scales of measurement and their implications. Scholten and Borsboom [37] provide a careful assessment of the role of Lord's claimed "counter example" in regard to Stevens' taxonomy of measurement scales.

In addition to the studies already discussed in Section III, a range of work has considered the underpinning measurements involved in IR. For example, Busin and Mizzaro [6] consider measurement scales and SERP orderings, developing and extending axiomatic relationships akin to the "Rule 1" and "Rule 2" given above; and Ferrante et al. [13] undertake a similar exploration. In a related study, Moffat [24] considers effectiveness metrics in terms of a suite of seven numeric properties that they might possess. Other work–for example, Turpin and Hersh [43], Turpin and Scholer [44], Sanderson et al. [36], Bailey et al. [3], Liu et al. [20], and Zhang et al. [51] - has considered the extent to which whole-of-SERP usefulness is adequately captured by current effectiveness metrics.

B. USE OF RECALL BASE

Fuhr [17] and Ferrante et al. [14], [15], [16] note the difficulties created by the use in some metrics of what they term the "recall base", the number RB of relevant (assuming only binary relevance grades) documents in the collection for the topic in question, denoted in Section III as the quantity n_1. From the point of view of Fuhr and Ferrante et al., those difficulties arise because normalization by RB means that the set of generable measurement points for any query in a set of topics might not numerically align with the available measurement points for other topics that have different values for RB. But it is worth noting that the recall base affects the available measurements even when RB is not a visible component of the effectiveness metric. To observe this, consider Table 3 again. It lists the ten possible SERP classes for a collection of $n = 5$ documents and for queries with $n_1 = 2$ relevant documents in the collection, together with metric score according to five metrics. If another topic for the same test collection has $n_1 = 1$, the $\text{RBP}_{0.5}$ scores are limited to the set $\{0.500, 0.250, 0.125, 0.063, 0.031\}$, none of which align with the set of available $\text{RBP}_{0.5}$ scores listed in Table 3 for the case $n_1 = 2$.

Moreover, even when truncated rankings are considered, with (say) $k = 2$ used to calculate the scores, a topic for which $n_1 = 1$ is unable to deliver a $\text{RBP}_{0.5}$ score of 0.75. If $\text{RBP}_{0.5}$ is to be used across a collection of topics, and if exactly the same set of measurement points must be available for every topic, then the SERP truncation length k must satisfy $k \leq \min_i n_{1,i}$, where $n_{1,i}$ is the n_1 recall base associated with the ith topic. This places a severe limitation on any experiments making use of that collection. (The same restriction also applies to n_0, but in most retrieval environments n_1 is smaller than n_0 by several orders of magnitude.)

Our contention in this work is that the measurement scale is always the positive real number line, and hence that no question of alignment (or not) of measurement points across sets of topics arises. On the other hand, there are other reasons to eschew metrics that make use of the recall base, based on the desire for effectiveness metrics to reflect plausible user behaviors, and the user's inability to actually know the value n_1 as they consider the SERP [22], [25], [52].

C. GRADED RELEVANCE AND GAIN MAPPINGS

The discussion above focused on binary-level document relevance labels, but the same points apply to multi-level labels of the kind suggested in Table 2(d). Multi-level evaluations normally make use of two mapping stages. The first converts ordinal document relevance classes to numeric *gains* via a *gain mapping* function $N(\cdot)$ that converts ordinal document relevance grades to gain values in $0 \ldots 1$, as shown, for example, in Table 2(d). The second mapping then takes an n- or k-vector of numeric gain values, combines them in a way that discounts gains as ranks increase, and generates a single numeric score. The metrics *discounted cumulative gain* (DCG) and *normalized discounted cumulative gain*

(NDCG) [19] make quite deliberate use of real-valued document gains, as do RBP [25] and *expected reciprocal rank* (ERR) [9], with the goal of providing more nuanced effectiveness measurements, and hence the ability to respond with more sensitivity to perceived differences in SERP usefulness [39]. Average precision can also be broadened to make use of graded document relevance categories [12], [29].

The complex inter-relationships between the range of gain mappings that might be employed, and then the metric mapping itself, further mean that metric scores will not (and as is our firm contention here, need not) result in uniform-interval measurements.

Gain mappings are also measurements, of course, pertaining to the usefulness of individual documents. For example, the ordinal class labels listed in Table 2(d) might be included in a handbook provided to assessors as part of their training, along with detailed descriptions and examples. Document gain labels – the r_i values used to compute effectiveness metrics – can also be more directly measured. For example, magnitude estimation techniques [23], [45], side-by-side preference elicitation [1], [8], [36], [49], and ordinal scales in which the class labels are numbers [30] can all be used to develop numeric document gain labels.

D. STATISTICAL TESTS

Statistical tests are another important component of IR evaluation; see, for example, Smucker et al. [38], Sakai [31], and Urbano et al. [46]. The appropriateness of any particular test depends in part on the distributional conditions required by that test, and it should be noted that our argument here in regard to metric values being numbers that can be averaged is most definitely *not* an argument that all metrics can be tested with any particular statistical test. Thoughtful selection of a statistical test, and, if necessary, verification of any required distributional conditions governing its applicability, must always be a part of IR experimental design. On the other hand, choosing an effectiveness metric because it is amenable to a particular statistical test represents "the tail wagging the dog" (pun intended), and is not a course of action that should be considered. The metric must be chosen first, and only then can the statistical test be selected.

Similarly, we have no concerns with Fuhr's seventh rule [17], covering the need for multiple hypothesis adjustments, but note that in the case of test collection reuse it cannot always be properly achieved. Craswell et al. [11] provide an overview of some of these issues, and Sakai [32] has also voiced opinions in support of Fuhr's comments in regard to statistical testing.

V. CONCLUSION

We have discussed the role of interval scale measures in information retrieval evaluation. Via a sequence of examples we have presented our view that *all* IR effectiveness metrics can be considered to be interval scale measurements, provided only that the mapping from SERP categories to numeric scores has a real-world basis (an external validity) and can be motivated as corresponding to the underlying usefulness of each SERP, as experienced by an identified cohort of users as they carry out some identified search task. That is, while care needs to be exercised when choosing the metric that best fits the user experience for any particular IR application (for example, the "shallow-hasty-youthful" users that form the AmaBaiBinGoo demographic), once that match has been decided, the values calculated by the effectiveness metric may be used as simple numbers "*that don't remember where they came from*" [21]; that is, without regard to their origins in a categorical-scale SERP dataset.

Metric choice is a critically important design decision in any IR experiment, and different metrics might lead to different outcomes from a planned experiment. But the choice between metrics – and hence between possible experimental outcomes – should determined by the projected user cohort and their implicit evaluation of usefulness relative to their search task, and not because of the regularity or otherwise of the gaps between adjacent numeric values generated over the universe of categorical SERP classes, and nor as a consequence of amenability or system separability associated with any particular statistical test.

In addition, we have argued that the proposed intervalization of current IR effectiveness metrics is neither required nor helpful. If the raw metric value is indeed a defensible measurement of SERP usefulness and corresponds to the user's experience when they are presented with a member of that SERP category, then equi-intervalizing those measurements via a different categorical to numeric mapping must of necessity distort and alter any findings that arise, and thus risk masking what would otherwise be valid conclusions. And if the raw metric is not a defensible measurement of SERP usefulness for the search task at hand, then equi-intervalizing its scores is unlikely to improve the situation.

ACKNOWLEDGMENT

Joel Mackenzie, Tetsuya Sakai, Falk Scholer, Paul Thomas, and Justin Zobel provided useful input. Science proceeds via debate and the author is grateful to be part of a community in which debate is not only possible, but actively welcomed.

REFERENCES

[1] N. Arabzadeh, A. Vtyurina, X. Yan, and C. L. A. Clarke, "Shallow pooling for sparse labels," 2021, *arXiv:2109.00062*.

[2] L. Azzopardi, P. Thomas, and N. Craswell, "Measuring the utility of search engine result pages: An information foraging based measure," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 605–614.

[3] P. Bailey, N. Craswell, R. W. White, L. Chen, A. Satyanarayana, and S. M. M. Tahaghoghi, "Evaluating whole-page relevance," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2010, pp. 767–768.

[4] A. Broder, "A taxonomy of web search," *ACM SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.

[5] C. Buckley and E. M. Voorhees, "Retrieval system evaluation," in *TREC: Experiment and Evaluation in Information Retrieval*, E. M. Voorhees and D. K. Harman, Eds. Cambridge, MA, USA: MIT Press, 2005, ch. 3.

[6] L. Busin and S. Mizzaro, "Axiometrics: An axiomatic approach to information retrieval effectiveness metrics," in *Proc. Int. Conf. Theory Inf. Retr. (ICTIR)*, 2013, pp. 1–8.

[7] B. Carterette, "System effectiveness, user models, and user utility: A conceptual framework for investigation," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf.*, 2011, pp. 903–912.

[8] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais, ''Here or there: Preference judgments for relevance,'' in *Proc. Eur. Conf. Inf. Retr. (ECIR)*, 2008, pp. 16–27.

[9] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, ''Expected reciprocal rank for graded relevance,'' in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 621–630.

[10] W. S. Cooper, ''Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems,'' *Amer. Documentation*, vol. 19, no. 1, pp. 30–41, Jan. 1968.

[11] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and J. Lin, ''MS MARCO: Benchmarking ranking models in the large-data regime,'' in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1566–1576.

[12] G. Dupret and B. Piwowarski, ''A user behavior model for average precision and its generalization to graded judgments,'' in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2010, pp. 531–538.

[13] M. Ferrante, N. Ferro, and M. Maistro, ''Towards a formal framework for utility-oriented measurements of retrieval effectiveness,'' in *Proc. Int. Conf. Theory Inf. Retr.*, Sep. 2015, pp. 21–30.

[14] M. Ferrante, N. Ferro, and S. Pontarollo, ''A general theory of IR evaluation measures,'' *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 3, pp. 409–422, Mar. 2019.

[15] M. Ferrante, N. Ferro, and E. Losiouk, ''How do interval scales help us with better understanding IR evaluation measures?'' *Inf. Retr. J.*, vol. 23, no. 3, pp. 289–317, Jun. 2020.

[16] M. Ferrante, N. Ferro, and N. Fuhr, ''Towards meaningful statements in IR evaluation: Mapping evaluation measures to interval scales,'' *IEEE Access*, vol. 9, pp. 136182–136216, 2021.

[17] N. Fuhr, ''Some common mistakes in IR evaluation, and how they can be avoided,'' *ACM SIGIR Forum*, vol. 51, no. 3, pp. 32–41, Feb. 2018.

[18] W. Hays, *Statistics*, 5th ed. New York, NY, USA: Harcourt Brace, 1994.

[19] K. Järvelin and J. Kekäläinen, ''Cumulated gain-based evaluation of IR techniques,'' *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.

[20] M. Liu, Y. Liu, J. Mao, C. Luo, M. Zhang, and S. Ma, '''Satisfaction with failure' or 'unsatisfied success': Investigating the relationship between search success and user satisfaction,'' in *Proc. World Wide Web Conf. World Wide Web*, 2018, pp. 1533–1542.

[21] F. M. Lord, ''On the statistical treatment of football numbers,'' *Amer. Psychol.*, vol. 8, no. 12, pp. 750–751, Dec. 1953.

[22] X. Lu, A. Moffat, and J. S. Culpepper, ''The effect of pooling and evaluation depth on IR metrics,'' *Inf. Retr. J.*, vol. 19, no. 4, pp. 416–445, Aug. 2016.

[23] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin, ''On crowdsourcing relevance magnitudes for information retrieval evaluation,'' *ACM Trans. Inf. Syst.*, vol. 35, no. 3, pp. 19:1–19:32, 2017.

[24] A. Moffat, ''Seven numeric properties of effectiveness metrics,'' in *Proc. Asia Inf. Retr. Societies Conf. (AIRS)*, 2013, pp. 1–12.

[25] A. Moffat and J. Zobel, ''Rank-biased precision for measurement of retrieval effectiveness,'' *ACM Trans. Inf. Syst.*, vol. 27, no. 1, pp. 1–27, Dec. 2008.

[26] A. Moffat, P. Bailey, F. Scholer, and P. Thomas, ''Incorporating user expectations and behavior into the measurement of search effectiveness,'' *ACM Trans. Inf. Syst.*, vol. 35, no. 3, pp. 24:1–24:38, 2017.

[27] A. Moffat, J. Mackenzie, P. Thomas, and L. Azzopardi, ''A flexible framework for offline effectiveness metrics,'' in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 578–587.

[28] S. Robertson, ''A new interpretation of average precision,'' in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2008, pp. 689–690.

[29] S. E. Robertson, E. Kanoulas, and E. Yilmaz, ''Extending average precision to graded relevance judgments,'' in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2010, pp. 603–610.

[30] K. Roitero, E. Maddalena, G. Demartini, and S. Mizzaro, ''On fine-grained relevance scales,'' in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 675–684.

[31] T. Sakai, ''Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015,'' in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 5–14.

[32] T. Sakai, ''On Fuhr's guideline for IR evaluation,'' *SIGIR Forum*, vol. 54, no. 1, pp. 12:1–12:8, 2020.

[33] T. Sakai and Z. Zeng, ''Which diversity evaluation measures are 'good?''' in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 595–604.

[34] T. Sakai and Z. Zeng, ''Retrieval evaluation measures that agree with users' SERP preferences: Traditional, preference-based, and diversity measures,'' *ACM Trans. Inf. Syst.*, vol. 39, no. 2, pp. 14:1–14:35, 2021.

[35] M. Sanderson, ''Test collection based evaluation of information retrieval systems,'' *Found. Trends Inf. Retr.*, vol. 4, no. 4, pp. 247–375, 2010.

[36] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas, ''Do user preferences and evaluation measures line up?'' in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2010, pp. 555–562.

[37] A. Z. Scholten and D. Borsboom, ''A reanalysis of Lord's statistical treatment of football numbers,'' *J. Math. Psychol.*, vol. 53, no. 2, pp. 69–75, Apr. 2009.

[38] M. D. Smucker, J. Allan, and B. Carterette, ''A comparison of statistical significance tests for information retrieval evaluation,'' in *Proc. 16th ACM Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2007, pp. 623–632.

[39] E. Sormunen, ''Liberal relevance criteria of TREC—Counting on negligible documents?'' in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2002, pp. 324–330.

[40] S. Stevens, ''On the theory of scales of measurement,'' *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[41] P. Thomas, A. Moffat, P. Bailey, F. Scholer, and N. Craswell, ''Better effectiveness metrics for SERPs, cards, and rankings,'' in *Proc. 23rd Australas. Document Comput. Symp.*, Dec. 2018, pp. 1–8.

[42] J. T. Townsend and F. G. Ashby, ''Measurement scales and statistics: The misconception misconceived,'' *Psychol. Bull.*, vol. 96, no. 2, pp. 394–401, Sep. 1984.

[43] A. H. Turpin and W. Hersh, ''Why batch and user evaluations do not give the same results,'' in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2001, pp. 225–231.

[44] A. Turpin and F. Scholer, ''User performance versus precision measures for simple search tasks,'' in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2006, pp. 11–18.

[45] A. Turpin, F. Scholer, S. Mizzaro, and E. Maddalena, ''The benefits of magnitude estimation relevance assessments for information retrieval evaluation,'' in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 565–574.

[46] J. Urbano, H. Lima, and A. Hanjalic, ''Statistical significance testing in information retrieval: An empirical analysis of type I, type II and type III errors,'' in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 505–514.

[47] P. F. Velleman and L. Wilkinson, ''Nominal, ordinal, interval, and ratio typologies are misleading,'' *Amer. Statistician*, vol. 47, no. 1, pp. 65–72, Feb. 1993.

[48] A. F. Wicaksono and A. Moffat, ''Metrics, user models, and satisfaction,'' in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 654–662.

[49] Z. Yang, A. Moffat, and A. Turpin, ''Pairwise crowd judgments: Preference, absolute, and ratio,'' in *Proc. 23rd Australas. Document Comput. Symp.*, Dec. 2018, pp. 1–8.

[50] F. Zhang, Y. Liu, X. Li, M. Zhang, Y. Xu, and S. Ma, ''Evaluating web search with a bejeweled player model,'' in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 425–434.

[51] F. Zhang, J. Mao, Y. Liu, X. Xie, W. Ma, M. Zhang, and S. Ma, ''Models versus satisfaction: Towards a better understanding of evaluation metrics,'' in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 379–388.

[52] J. Zobel, A. Moffat, and L. A. F. Park, ''Against recall: Is it persistence, cardinality, density, coverage, or totality?'' *SIGIR Forum*, vol. 43, no. 1, pp. 3–15, 2009.

ALISTAIR MOFFAT received the Ph.D. degree in computer science from the University of Canterbury, New Zealand, in 1986. Since then, he has been a Faculty Member of The University of Melbourne, Australia, with interests in text and index compression and algorithms for string search and information retrieval, including information retrieval evaluation. He was inducted as a member of the SIGIR Academy, in 2021.