Some Common Mistakes In IR Evaluation, And How They Can Be Avoided

Norbert Fuhr
University of Duisburg-Essen, Germany
norbert.fuhr@uni-due.de

Abstract

This paper points out some mistakes that can be frequently found in IR publications: MRR and ERR violate basic requirements for a metric, MAP is based on unrealistic assumptions, the numbers shown overstate the precision of the result, relative improvements of arithmetic means are inappropriate, the simple holdout method yields unreliable results, hypotheses are often formulated after the experiment, significance tests frequently ignore the multiple comparisons problem, effect sizes are ignored, reproducibility of the experiments might be nearly impossible, and sometimes authors claim proof by experimentation.

1 Introduction

The field of IR is very proud of its long history of experimentation-oriented research. However, taking a closer look at current evaluation practices, one notices that some of them are in conflict with basic theoretic findings. Since most of them have a long tradition, many researchers regard them as established procedures which they also apply. Occasionally reviewers flag some of them as errors to be corrected, but this usually only affects the paper under consideration.

Thus, the goal of this paper is to give a list of frequent bad practices, along with concrete advice on how to avoid them.

2 The mistakes and how to avoid them

2.1 Thou shalt not compute MRR nor ERR

The definition of mean reciprocal rank (MRR) assumes that a user scans the ranked list and stops after the first relevant. Let r denote the rank of the first relevant document, then the reciprocal rank is defined as RR(r)= $1/r$. Intuitively, this measure looks ok: it returns values between 0 and 1, where the perfect result yields 1.0.

However, MRR shows some strange behavior, as the following example illustrates: Assume that we have three queries, and system A returns the first relevant document at ranks 1, 2 and 4, respectively, while system B returns the relevant answers in each case at rank 2. So, with system A, we find the first relevant item on average at rank 2.33 ($= \frac{1}{3}(1 + 2 + 4)$), which is worse than what system B does. However, MRR(A)=$\frac{1}{3}$(1/1 + 1/2 + 1/4) = 0.58, while MRR(B)=0.5. So A outperforms B on MRR, but looking at the average rank, we get the opposite finding! Obviously, this observation is due to the basic property of expected values: $E(1/R) \neq 1/E(R)$.

The major problem with MRR, however, is the following: as observed already in [19], the difference 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 [20] [22, p. 49–51]. However, one cannot compute the mean for an ordinal scale! Only the median would be possible here — which would produce ties in most cases.

As a simple way out, one can regard the rank numbers directly, without any transformation, and then compute the arithmetic mean for a set of queries (like we did in the example from above). Let us call this measure 'mean first relevant' (MFR). Although MFR has the less common property of higher values being worse, its values are more understandable than those of MRR. Moreover, MFR is even a ratio scale: Looking at the user effort for finding the first relevant document, an MFR value of x means the x-fold effort in comparison to the ideal value of 1.

One could think about computing the harmonic mean of the RR values, which would result in the inverse of MFR, and thus be a correct operation.. However, when we want to apply significance tests on these harmonic means, we have the same problem as before: since it is not allowed to compute differences, neither t-test nor Wilcoxon would be possible.

The ERR measure proposed in [5] is a generalization of reciprocal rank to graded relevance scales, and so it suffers from the same problem. ERR is based on the idea that the stopping probability is a function of the relevance grade of the document under consideration (in the binary case, this probability is 1 for relevant documents, and 0 otherwise): $ERR := \sum_{r=1}^{n} \frac{1}{r} \cdot P(\text{user stops at position } r)$. Instead, we propose as corresponding generalization of MFR the "expected first relevant" measure:

$$EFR := \sum_{r=1}^{n} r \cdot P(\text{user stops at position } r)$$

Even for single queries, ERR yields counter-intuitive results: As an example, assume that system A returns a partially relevant document (with stopping probability 0.5) at rank 1 and a fully relevant one (with stopping probability 1) at rank 5, while system B returns the same documents at ranks 2 and 3, respectively. Then we would have $ERR(A) = 0.5 \cdot 1/1 + 0.5 \cdot 1/5 = 0.6$, while $ERR(B) = 0.5 \cdot 1/2 + 0.5 \cdot 1/3 = 0.417$. In contrast we have $EFR(A) = 0.5 \cdot 1 + 0.5 \cdot 5 = 3$, while $EFR(B) = 0.5 \cdot 2 + 0.5 \cdot 3 = 2.5$. So A would be better according to ERR, but is worse in terms of EFR; however, only the latter is directly related to expected user effort.

2.2 Thou shalt not use MAP

MAP was developed originally as a refinement of earlier methods of representing the recall-precision-curve by a single number (e.g. 11-point average precision at 0, 0.1, …, 1 recall [12, ch. 8]). Later Robertson [16] was able to show that it is possible to define a user model for this metric, by making certain assumptions about user behavior: When scanning documents in rank order

1. users stop only after a relevant document, and
2. the probability of stopping is the same for all relevant ranks.

While the first assumption might be approximately true in some applications (but how would users know when they have reached the last relevant document?), it is hard to imagine any use case where the second one holds. Most users will stop early on, and only a small fraction will go further down the ranked list; e.g., [10] points out that the distribution follows the power law, and web search behavior shows similar patterns[1] [9].

Therefore, it is more realistic to assume a skewed stopping distribution over the ranks, and also allow for stopping after a nonrelevant item. This is exactly what rank-biased precision (RBP) [13] does, where the user-defined parameter p gives the probability that the user continues to the next rank. Thus, the fraction of users stopping at rank k is $s(k) = (1 - p)p^{k-1}$.

[17] mentions two weaknesses of this measure: First, the maximum possible value depends on the parameter p; this seems to be a minor issue, since hardly anyone understands what e.g. a MAP value of 0.234 means (see also Section 2.8). Second, RBP has less discriminative power; this finding may be due to the fact that RBP puts more weight on the events that most users observe (i.e. documents in the top ranks), while MAP gives equal weight to items that hardly anyone looks at. Presumably, many systems yield very similar user experiences, whereas MAP tries to differentiate on aspects hardly any user is interested in.

Just like RBP, discounted cumulative gain [11] also makes more realistic assumptions about users' stopping behavior than MAP; as an additional advantage, it can deal with multivalued relevance scales. Most important, this is one of the few measures where we have empirical evidence on the agreement between user preferences and evaluation metric [18].

While most of the other bad practices described here are clear errors, using MAP might be theoretically correct. However, as it is based on a superficial user model, it yields misleading results, and thus should be avoided.

2.3 Thou shalt not overstate the precision of your results

Standard IR evaluation software typically prints results with four decimal digits, and so most authors copy these numbers directly into their paper. However, these four digits create the illusion of a precision that hardly ever exists. In experiments, when we have only a few hundred observations (e.g. relevance of documents), then four decimal places are inappropriate. So, as a minimum requirement, a single positive observation more or less in the raw data should affect at most the last digit shown.

When measuring precise quantities like e.g. in physics, it is common to denote the standard deviation in case there are repeated measurements. This would also be possible in IR. On the other hand, given that we are dealing with stochastic experiments, the most reasonable approach is to specify confidence intervals. For example, if we have 50 queries and measure the P@10 value, then we have 500 observations; assuming these events as being independent of each other, we could approximate the underlying binomial distribution by a Gaussian, where the Wilson Score Interval[2] (see also [22, p. 175]) yields for P@10=0.7 a 95% confidence interval of [0.658, 0.739]. However, as we have only 50 independent events (i.e. queries) here, the actual confidence interval is even wider – the variance in performance

[1] `https://searchenginewatch.com/sew/study/2276184/no-1-position-in-google-gets-33-of-\search-traffic-study`, *last checked: October 31, 2017*

[2] `https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval`, *last checked: October 31, 2017*

of different topics of a test collection is well-known. When there is no closed formula for the confidence interval, one can apply bootstrap methods [7].

Although confidence intervals seem to be large, they highlight the actual precision of the measurement as well as the effect size, and also avoid some of the problems with statistical testing [6] (see also below).

2.4 Thou shalt not compute relative improvements of arithmetic means

When comparing arithmetic means of measurements (e.g. over a set of queries), authors often denote the relative improvements of their preferred method over competing ones. However, this way of comparison is not appropriate, as the following example shows: assume that we have two queries, where system A yields a value of 0.2 for query 1 and 0.5 for query 2, then system B yielding 0.18 and 0.54 would be regarded as being better, based on the arithmetic mean. On the other hand, a query-wise comparison shows that B looses 10% on the first query, and gains only 8% on the second one, thus its average change would be -1%. More generally, suppose that the random variable X denotes the performance of system A, and K is the relative improvement of B over A. Then we have for the corresponding expected values $E(X \cdot K) \neq E(X) \cdot E(K)$, and thus, we cannot compute $E(K)$ as $E(X \cdot K)/E(X)$. Thus, for arithmetic mean, only absolute improvements are meaningful (since $E(X-Y) = E(X)-E(Y)$).

If one is interested in relative changes, one should use the geometric mean μ_G instead (see e.g. the definition of GMAP [15]). This kind of mean can be applied to ratio scales only, but it has the nice property that for two series $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$, $\mu_G(X/Y) = \mu_G(X)/\mu_G(Y)$.

The question of preferring either arithmetic or geometric mean depends on the nature of the performance differences between the systems being compared (in face of variations in the query/task-wise results): Assuming that the absolute differences are independent of the actual values, then arithmetic mean would be the most appropriate averaging method. If, however, we think that the differences are roughly proportional to the measured values, then we should go with the geometric mean.

Sometimes, it is claimed that the geometric mean puts more emphasis on the weaker results; again, this is true only if we assume that system differences are absolute rather than proportional. Otherwise, arithmetic mean puts too much emphasis on the stronger results.

Overall, the best way to choose between arithmetic and geometric mean would be a correlation analysis[3] [21, ch. 5, p. 99ff] of system differences vs. absolute values.

2.5 Thou shalt not apply the simple holdout method

In classical system-oriented IR evaluation, the so-called holdout method of separating the available data into a training set and a (disjoint) test set has been standard for many years. However, this method yields quite unreliable outcomes, since the results may be heavily dependent on the split (e.g. the nice reproducibility study [14] shows that the performance on the two halves of the collection differed by 20%, and, for other splits, system differences sometimes were significant, and sometimes not).

[3] `https://en.wikipedia.org/wiki/Correlation_and_dependence`, *last checked: October 31, 2017*

In order to avoid these problems, cross-validation (see e.g. [22, p. 152-6]) should be a minimum requirement for experimentation.

Some settings require not only a training set, but also a tuning set, e.g. for determining the best parameter setting, or choosing the best one from a variety of methods (see also the multiple hypotheses problem described below). Again, this tuning set should be disjoint from the test set, so that we have to split the available data into three different subsets. For k-fold cross-validation, after putting aside one fold after the other for testing, we can either use just one fold for tuning and the remaining $k - 2$ sets for training, or we can perform an inner loop where we rotate the tuning fold over the $k - 1$ folds.

In evaluation initiatives, the setting usually corresponds to the simple holdout method, where participants are initially given some training data, and get access to the testing set only for performing their runs, while the ground truth (e.g., relevance judgments) for this set is disclosed later. Although there are reasons for this procedure, more reliable results could be achieved (at a later stage) if one would also regard the results after switching the role of training and test set, or apply cross-validation on the union of both.

When authors compare their results to the best published run for a test collection, then the selection of the latter can also be regarded as a tuning process. Thus, such a comparison cannot be fair, it should rather be performed on a separate test collection. While one would be tempted to say that this disadvantage makes the new, better method even more valuable, there are additional problems that need to be considered, as described in the following.

2.6 Thou shalt not formulate hypotheses after the experiment

Although this seems to be basic knowledge, many experimental papers give a different impression: Usually there is a new method (possibly with some variants), for which experiments are carried out, different measures are applied, and then significance tests are performed on some or all of the observed differences. The hypotheses are usually only specified implicitly, and often only after having seen the results; moreover, additional variants may be mentioned only in passing (and the authors also might have tried other methods not described in the text). There are only few papers that specify the hypotheses explicitly (and before the experiment), like saying: ”We want to compare methods A and B using metric X”.

The most crucial point here is that the complete set of hypotheses to be tested must be known before the experiment, as this has consequences for the test setup — see below. For example, it would be incorrect if an author first determines which of the competing methods performs best on his data set, and then compares his own approach only to this one.

In evaluation initiatives, there is often the implicit goal of ranking the participating runs. However, this means that we have to compare all pairs of methods. Looking at the results first and then testing only for some of the observed differences would either require an additional data set, or specific tests (see below).

2.7 Thou shalt not test multiple hypotheses without correction

When reporting experimental results, many authors regard several approaches/variants, and then perform significance tests for each of them. Unfortunately, this procedure leads to the multiple comparisons problem [4] [21, p. 308]. An extensive treatment of this problem from an IR point of view can be found in [4].

As a simple example, assume that we test a drug vs. a placebo on two groups of patients, and then we regard 20 different symptoms after treatment. Applying significance tests with $p = 0.05$ on each of these features, we can expect one 'significant' difference, even if the drug has no effect at all.

In general, when we test more than one feature on the same set of data, we have to do appropriate adjustments. An extensive survey over correction methods can be found e.g. in [1]. A simple (conservative) method is the Bonferroni correction: When testing m hypotheses, we have to divide the desired p-value by m in order to get the significance level to test on (e.g. when considering five metrics, for a p-level of 0.05, we actually have to test with $p = 0.01$).

While the full problem only strikes when the different features are uncorrelated, the problem is less severe when only several measures for a single approach are regarded. On the other hand, if the measures are strongly correlated, then there is no need to test each of them; in case they are less correlated, a correction is necessary. As long as there are no better methods for dealing with this problem, we have to stick with the standard approaches, or regard a single measure only.

[3] points to an even more severe variant of the problem discussed here: When test collections are reused, chances increase that we observe random results. So collection reuse is also a case of multiple hypothesis testing; however, here we perform sequential testing, where we might use knowledge from the outcome of the previous tests, which makes things even worse.

As a minimum requirement for reuse (that still ignores the sequence problem), one should consider the set of tests already published for this collection, add the number of own tests, and then apply a correction for the total number of tests. This approach would still blend out the unreported cases of tests performed on the same collection.

For evaluation initiatives, this problem also has severe consequences, since the number of pairwise tests grows quadratically with the number of submitted runs. The only reasonable method for dealing with this problem is the application of a post-hoc test such as e.g. Tukey's test [5] [21, ch. 10, p. 325ff], which has been used frequently in CLEF [2]; this test checks all pairwise differences between runs. However, when more than one metric is considered, an additional correction is necessary.

2.8 Thou shalt not ignore effect sizes

Even when a (properly exercised) significance test rejects the null hypothesis, we can only infer that it is unlikely that the methods compared yield exactly the same result. However, the test does not tell us anything about the actual difference between the values to be compared [6]. Especially for large data sets (e.g. from popular online services), almost any modification will result in a significant, but tiny difference. Thus, it is essential to report also the size of the effect achieved by the method under investigation.

For some transparent metrics like precision or recall, the difference of the arithmetic means bears already some useful information. E.g., for P@10, a difference of 0.02 translates into one more relevant document per five queries — which many users might regard as the

[4] `https://en.wikipedia.org/wiki/Multiple_comparisons_problem`, *last checked: October 31, 2017*

[5] `https://en.wikipedia.org/wiki/Tukey%27s_range_test`, *last checked: October 31, 2017*

minimum noticeable effect; so smaller differences, even when they are significant, might be irrelevant for most users. In case of more complex metrics like e.g. RBP or NDCG, the difference is hard to interpret.

Effect size not only regards the difference, it also sets it in relation to the variance of the test data. As stated in [21, ch. 7, p. 187] "an effect size is a statistic quantifying the extent to which sample statistics diverge from the null hypothesis". For comparing two arithmetic means μ_1 and μ_2, the effect size Δ is usually defined as[6] [21, ch. 7, p. 187ff]

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}$$

Here σ denotes the standard deviation based on either or both populations. In case one of them plays the role of a baseline, then σ is its standard deviation.

2.9 Thou shalt not forget about reproducibility

Reproducibility of experiments is an important concept in research, supporting validation of reported results as well as allowing for later comparison with new approaches. The ACM task force on reproducibility stated: "A scientific result is not fully established until it has been independently reproduced"[7]. Above, we have also pointed out that there is a need for new testing data when the original test collection has been exploited for tuning, and/or been 'exhausted' by testing too many hypotheses on it.

The PRIMAD model of reproducibility [8] distinguishes between the following components of an experiment:

Research Goal characterizes the purpose of a study;

Method is the specific approach proposed or considered by the researcher;

Implementation refers to the actual implementation of the method;

Platform describes the underlying hard- and software;

Data consists of the input data and the specific parameters chosen to carry out the method;

Actor denotes the experimenter.

When another researcher tries to reproduce the experiment, she will change some of the components, in many cases the *implementation*, the *platform* and the *actor*. In order to reproduce the original experiment, she must have access to a sufficiently detailed description of the other three components: the *research goal* usually poses no problem, but the description of the *method* might already lack some details (like e.g. the methods used for tokenization and stemming or the stopword list). When deep learning methods are applied, a precise description of the network used is hardly possible. Thus, in many cases, the only way for supporting reproducibility is via sharing of the implementation, using e.g. publicly accessible code repositories.

As for the *data*, sharing of test collections has a long tradition in IR. However, when this is not feasible (e.g. with proprietary data), reproducibility becomes nearly impossible. For example, the SIGIR 2017 reviewing guidelines[8] specify reproducibility as follows "...other

[6] https://en.wikipedia.org/wiki/Effect_size, *last checked: October 31, 2017*

[7] https://www.acm.org/data-software-reproducibility, *last checked: October 31, 2017*

[8] http://www.informagus.nl/sigir2017/review/guidelines-pc-fp/reviewing.html, *last checked: October 31, 2017*

researchers would be able to reproduce the method and/or results presented in the paper if they had access to the same or similar resources”. The crucial point here is the similarity of resources: How can another researcher check if her test collection is sufficiently similar to the data used in the original study? (See e.g. the 20% difference between subsets of the same collection mentioned above [14].) In other sciences, there may be standards for describing the essential components of a experiment. In IR, however, we just don’t know which of the many possible features characterizing a test collection are relevant for the outcome of an experiment, and if these features describe the data in sufficient detail. In any case, the current standard of mentioning only a few characteristics of the test data is clearly insufficient – researchers should at least share extensive statistics about their data, again via publicly accessible repositories.

2.10 Thou shalt not claim proof by experimentation

Sometimes authors use phrases like ’our experiments prove’. This demonstrates a fundamental misunderstanding of the role of experiments in our field. Proofs are about universally valid statements, while experiments only demonstrate the validity for a single or a few data sets. A good analogy for IR experiments are software tests: Positive outcomes for all tests performed just show that the software worked properly in these cases, but it is never a proof of correctness of the program under consideration.

Thus, even if a researcher has obeyed all the rules stated above while performing extensive experimentation for a new approach, this is only empirical evidence that the new method works as promised on the data used (and hopefully also in similar situations). Universally valid statements, however, can only be derived at the theoretical level.

3 Conclusion and outlook

In this paper, we have discussed some common mistakes in IR evaluation. We have not cited specific papers where these errors can be found, as one can easily spot examples for most of them in a random volume of IR conference proceedings.

The goal of this paper is to improve the current practice of IR evaluation:

- Most important, evaluation initiatives should take a leading role in this effort. Bad practices should be stopped immediately, and proper alternatives be established.
- In a similar way, program committees and editorial boards should develop a clear policy regarding these issues.
- Researchers should avoid the mistakes outlined here, and switch to the better alternatives.

A first step would be the publication of a checklist for evaluation along with CfPs, which refers to the issues described above (and possibly others). This checklist should be used by both authors and reviewers, thus helping to avoid publication of papers with flawed experimentation.

One should bear in mind that the problems discussed here produce misleading or unreliable experimental results, and thus makes their experimental results useless. Only reliable, reproducible results lead to scientific progress.

References

[1] Stefanie R Austin, Isaac Dialsingh, and Naomi Altman. Multiple hypothesis testing: A review. *J. Indian Soc. Of Agricultural Stat*, 68:303–314, 2014.

[2]]Martin Braschler. CLEF 2001 - Overview of Results. In *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers*. pages 9–26.

[3] Ben Carterette. The best published result is random: Sequential testing and its effect on reported effectiveness. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 747–750, New York, NY, USA, 2015. ACM.

[4] Benjamin A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans. Inf. Syst.*, 30(1):4:1–4:34, 2012.

[5] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[6] Jacob Cohen. The earth is round (p ¡ .05). *American Psychologist*, 49(12):997–1003, 1994.

[7] Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statist. Sci.*, 11(3):189–228, 09 1996.

[8] Nicola Ferro, Norbert Fuhr, Kalervo Jarvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. Increasing reproducibility in ir: Findings from the dagstuhl seminar on "reproducibility of data-oriented experiments in e-science". *SIGIR Forum*, 50(1):68–82, 2016. http://sigir.org/files/forum/2016J/p068.pdf.

[9] Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479, New York, NY, USA, 2004. ACM.

[10] R. Islamaj Dogan, G. C. Murray, A. Neveol, and Z. Lu. Understanding pubmed user search behavior through log analysis. *Database: The Journal of Biological Databases and Curation*, 2009.

[11] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, December 2008.

[14] Jinfeng Rao, Jimmy J. Lin, and Miles Efron. Reproducible experiments on lexical and temporal feedback for tweet search. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, volume 9022 of *Lecture Notes in Computer Science*, pages 755–767, 2015.

[15] Stephen Robertson. On gmap: And other transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 78–83, New York, NY, USA, 2006. ACM.

[16] Stephen E. Robertson. A new interpretation of average precision. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *SIGIR*, pages 689–690. ACM, 2008.

[17] Tetsuya Sakai. Metrics, statistics, tests. In Nicola Ferro, editor, *Bridging Between Information Retrieval and Databases*, volume 8173 of *Lecture Notes in Computer Science*, pages 116–163. Springer Berlin Heidelberg, 2014.

[18] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 555–562, New York, NY, USA, 2010. ACM.

[19] Amit Singhal, John Choi, Donald Hindle, and Fernando C. N. Pereira. At&t at TREC-6: SDR track. In D. Harman and E. M. Voorhees, editors, *Proceedings of The Sixth Text REtrieval Conference, TREC 1997, Gaithersburg, Maryland, USA, November 19-21, 1997*, pages 227–232, Gaithersburg, Md. 20899, 1997. National Institute of Standards and Technology.

[20] S.S. Stevens. On the theory of scales of measurement. *Science*, New Series 103(2684):677–680, June 1946.

[21] Bruce Thompson. *Foundations of Behavioral Statistics: An Insight-Based Approach.* The Guilford Press, 2006.

[22] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.