# Taking Risks with Confidence

Rodger Benham
RMIT University
Melbourne, Australia

Ben Carterette
Spotify
New York, USA

Alistair Moffat
University of Melbourne
Melbourne, Australia

J. Shane Culpepper
RMIT University
Melbourne, Australia

## ABSTRACT

Risk-based evaluation is a failure analysis tool that can be combined with traditional effectiveness metrics to ensure that the improvements observed are consistent across topics when comparing systems. Here we explore the stability of confidence intervals in inference-based risk measurement, extending previous work to five different commonly used inference testing techniques. Using the Robust04 and TREC Core 2017 NYT corpora, we show that risk inferences using parametric methods appear to disagree with their non-parametric counterparts, warranting further investigation. Additionally, we explore how the number of topics being evaluated affects confidence interval stability, and find that more than 50 topics appear to be required before risk-sensitive comparison results are consistent across different inference testing frameworks.

## KEYWORDS

Risk-biased evaluation; effectiveness metric; confidence interval

**ACM Reference Format:**
Rodger Benham, Ben Carterette, Alistair Moffat, and J. Shane Culpepper. 2019. Taking Risks with Confidence. In *Australasian Document Computing Symposium (ADCS '19), December 5–6, 2019, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3372124.3372125

## 1 INTRODUCTION

Several risk measures have been proposed in IR, with the goal of measuring overall system improvement, while limiting effectiveness loss on individual topics. The most popular measure is URisk by Wang et al. [16],

$$URisk_{\alpha} = (1/n) \cdot \left[\sum Wins - (1+\alpha) \cdot \sum Losses\right], \qquad (1)$$

where wins reflect positive differences in scores on the experimental system; where losses are the reverse; and where $n$ is the number of paired comparisons.

Dinçer et al. [7] studentized URisk, calling the measure TRisk, providing (to date) the only inferential measure of risk. When a proposed system has a TRisk score above 2 compared to a baseline, it is regarded as providing risk-free improvement. Conversely, a score less than $-2$ indicates that the proposed system is statistically risky, with degraded performance when risk is taken into account. While there is clearly value in inferential approaches to risk-based comparisons, Benham et al. [2] have recently shown that the asymmetric weighting induced by $\alpha$ in URisk can lead to unexpected deviations from normality. For TRisk, and other parametric statistical tests that assume a normal distribution, an asymmetric distribution will distort the inferences, but whether the degradation is important remains to be known. In this work, we seek to quantify how much that distributional shift affects those inferences, by comparing the confidence intervals produced by parametric and nonparametric statistical tests. A follow-up experiment explores how topic sample sizes affect confidence intervals in risk inference.

Null hypothesis statistical testing involves checking to see if an experiment's statistic falls outside a specified range in a control sample, with the goal of establishing that the observed outcome did not occur by sampling error (or random chance). This dichotomy comes with the possibility of falsely rejecting the *null hypothesis* $H_0$ (a type I error), or falsely accepting $H_0$ and ignoring a true effect (a type II error). To address the tension between type I and II errors, the parameter $\mathcal{A}$ (changed from $\alpha$ to avoid confusion with risk measures) modulates the sensitivity of the test (how frequently true and false positives are accepted), while $\beta$ controls test specificity (how frequently true and false negatives are identified). IR practitioners conventionally use $\mathcal{A} = 95\%$ with a power of $\beta = 80\%$ [17], corresponding to the significance levels explored in early work [5]. The parameters $\mathcal{A}$ and $\beta$ are not the only factors determining the outcome of statistical inference; the sample size $n$ of the experimental group, and the *effect size* $\delta$ (the magnitude of the difference between the two samples), are also important. A too-small $n$ for a given $\delta$ yields an *underpowered* study unlikely to measure statistically interesting effects; and a too-large $n$ for $\delta$ is *overpowered*, and may reject $H_0$ even though the measured effect $\delta$ is negligible [11].

Nayak and Hazra [10] describe the necessary steps for choosing the right statistical test. Of interest is whether to use tests that assume normality (*parametric* tests) in the context of risk-sensitive evaluation. Kitchen [9] comments that while parametric tests are slightly more powerful when their assumptions have been met, nonparametric tests are substantially more powerful when normality is violated. A common defense of using parametric tests is the Central Limit Theorem (CLT). The CLT roughly states that for a set of independent random variables of any shape, as sample size $n \to \infty$, the difference distribution between the sample statistic and population statistic is described by a normal distribution. Fifty topics had been found to be a large enough $n$ on classic IR measures [13], but Smucker et al. [14] also note that smaller sample sizes might still be acceptable. The CLT can play an important role in determining whether parametric inferences are valid or not, and so the number of topics evaluated will affect how normally distributed the data is. Given this context, we consider two research questions:

- **RQ1:** *How do parametric and non-parametric confidence intervals differ when performing risk-based evaluation?*
- **RQ2:** *How does evaluating larger topic samples affect confidence intervals when performing risk-based evaluation?*

*ADCS '19, December 5–6, 2019, Sydney, NSW, Australia*

ACM ISBN 978-1-4503-7766-9/19/12…$15.00
https://doi.org/10.1145/3372124.3372125

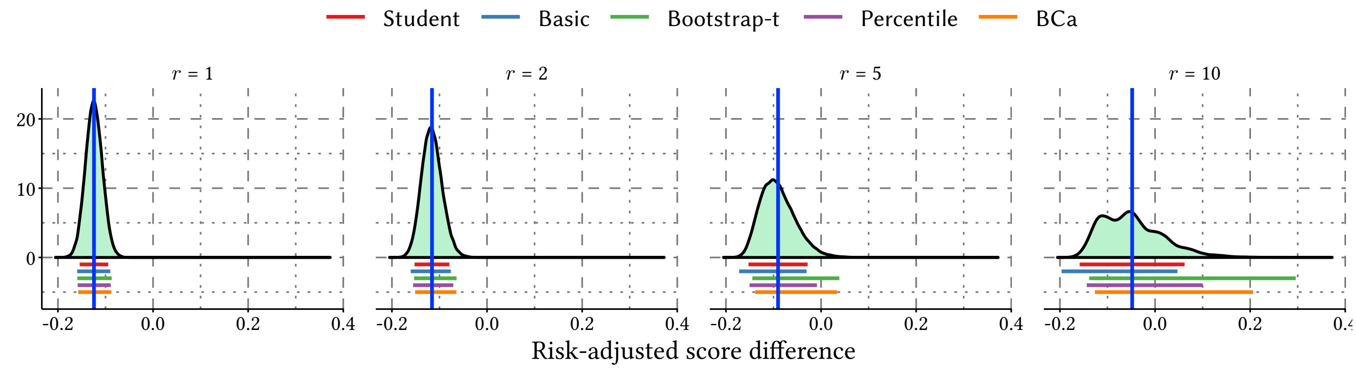


**Figure 1:** An example of 95% confidence intervals of URisk⁻ changing as $r$ is varied, with the density of the bootstrap sampling distribution shown for each $r$. A BM25 baseline run is compared against the best run `pircRB04td2` submitted to Robust04 as measured by AP, where the topics are sub-sampled to a traditional evaluation of 50 topics. The blue line indicates the arithmetic mean of the distribution.

## 2 APPROACH

**Notation**. Benham et al. [2] observe that the positive and negative directions of URisk and TRisk are inverted given conventional wisdom, and propose that URisk$^-$ and TRisk$^-$ be used instead, with a reversal of signs (URisk$^-$ = $-1 \cdot$ URisk), and with the $(1+\alpha)$ loss scalar component replaced by a single variable, denoted here as $r = (1+\alpha)$. To motivate the risk parameter $r$, if $r = 2$, the magnitude of each topic score difference less than zero between the baseline and the challenger system is doubled, and so on.

**Corpora and Risk Baselines**. We investigate risk score distributions using the TREC *New York Times* CORE 2017 corpus [1], as well as the Robust04 corpus [15], where NYT had the usual 50 topics and Robust04 had 250 (bar topic 672 which had no relevant documents). Both tracks had strong participation. Since our goal is to broadly study the inferential behavior of risk measures and the typical count of topics in a collection is 50, we sample from the Robust04 set by taking every fifth element, for example 301, 306, . . . , 696. Other samples of 50 were taken with similar results. Then, we evaluate the full Robust04 topic set and measure the shape of the distribution, and comment on how that impacts the support for the CIs. For each of the Robust04 and NYT corpora, we take as the baseline an untuned Okapi BM25 run computed using Indri 5.11.

**Confidence Intervals**. Several different ways of setting confidence intervals are considered. The first uses the Student t-distribution, and then another four are found using the Bootstrap distribution of a statistic [6]. The Canty et al. [3] `R` package `boot` was used to compute those four.

*Student-t Distribution CI*. Student-t CIs can be computed by

$$\hat{\theta}_{\mathcal{A}} = t - Q(\mathcal{A}, n-1)\,\frac{s_t}{\sqrt{n}} \text{ and } \hat{\theta}_{1-\mathcal{A}} = t + Q(\mathcal{A}, n-1)\,\frac{s_t}{\sqrt{n}}\,, \qquad (2)$$

where $t$ is the sample statistic, $s_t$ is the standard deviation of the sample, $n$ is the sample size, and $Q(\mathcal{A}, \lambda)$ is the quantile function of the Student-t distribution. We compute quantiles using the `R` function `qt` in the standard stats package. The Student-t distribution corresponds to the confidence limits in TRisk$^-$.

*Basic Bootstrap CI*. The basic bootstrap approach assumes a normal distribution. The distribution of bootstrap replicates $t^*$ reflect the statistic $T$, whose value in the sample is $t$:

$$\hat{\theta}_{\mathcal{A}} = 2t - t^*_{((M+1)(1-\mathcal{A}))} \text{ and } \hat{\theta}_{1-\mathcal{A}} = 2t - t^*_{((M+1)\mathcal{A})}\,. \qquad (3)$$

*Student-t Distribution Bootstrap CI*. Another parametric approach, this takes the basic CI form, except the $N(0,1)$ approximation is replaced with $Z = (T - \theta)/\sqrt{V}$; where $V$ is the variance statistic, and $\theta$ is the true value of $T$ to form an error region. For each of the sets formed in the $M$ bootstrap iterations, the $t^*$ and $v^*$ values are used to compute $z^* = (t^* - t)/\sqrt{v^*}$:

$$\hat{\theta}_{\mathcal{A}} = t - z^*_{((M+1)(1-\mathcal{A}))}\sqrt{v} \text{ and } \hat{\theta}_{1-\mathcal{A}} = t - z^*_{((M+1)\mathcal{A})}\sqrt{v}\,, \qquad (4)$$

where $v$ is the variance of the original sample.

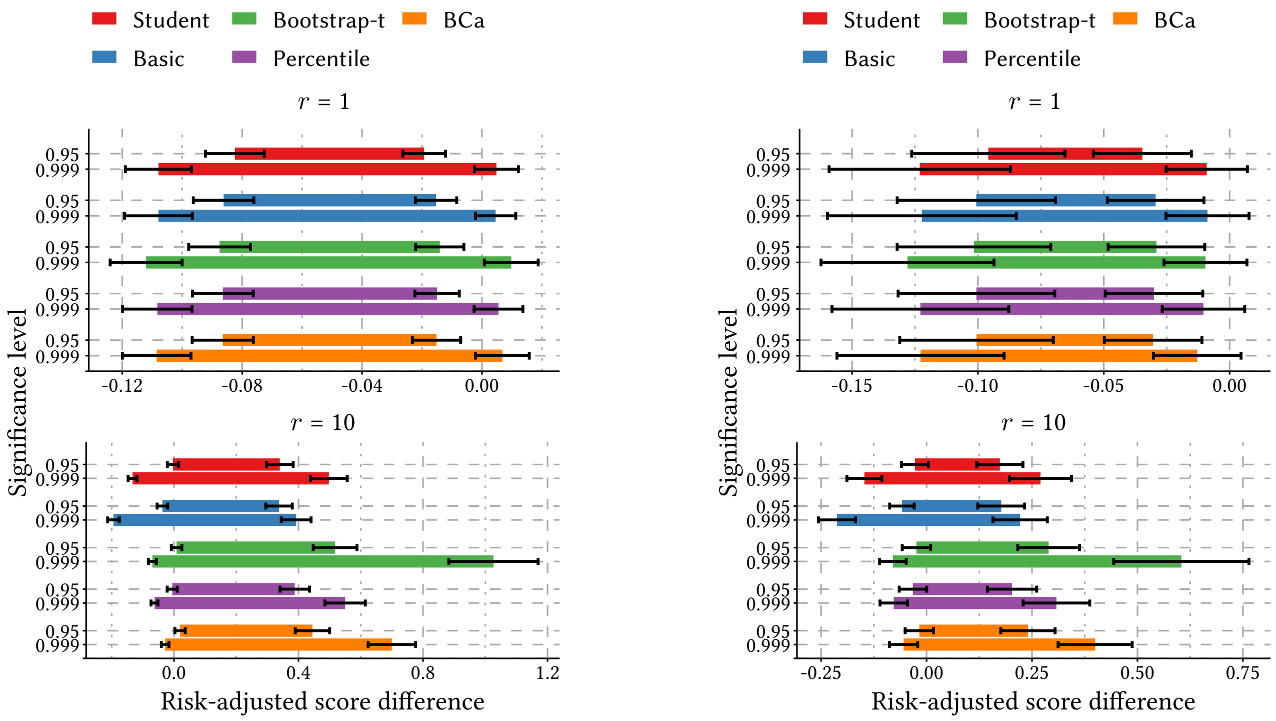
*Percentile Bootstrap CI*. The percentile method can be used on parametric and non-parametric bootstrap samples, supposing the statistic $T$ has a symmetric distribution:

$$\hat{\theta}_{\mathcal{A}} = t^*_{((M+1)\mathcal{A})} \text{ and } \hat{\theta}_{1-\mathcal{A}} = t^*_{((M+1)(1-\mathcal{A}))}\,. \qquad (5)$$

*Bias-Corrected and Accelerated Bootstrap CI (BC$_a$)*. If the swap of quantile estimates come from a biased scale the results will be inaccurate. To address that for the percentile method on non-parametric bootstrap samples, calculating and correcting for the bias in the distribution is needed. Using $t^*$, we compute the bias-correction parameter $w$, which corresponds to the ratio of the number of times a $t^*$ value is less than $t$, to $M$ – which is then transformed into a z-value. The quantile function of the normal distribution is again used to transform the upper and lower $\mathcal{A}$ confidence limits to z-values, denoted $z_{\mathcal{A}}$. Then, the acceleration parameter $a$ is computed. A set of $l^*$ Jackknife estimates of $T$ is formed, where the set $L$ corresponds to the difference in the mean Jackknife estimate, and the corresponding value computed in $l^*$. The acceleration $a$ is calculated using

$$a = \frac{\sum_{l \in L} l^3}{6\left(\sum_{l \in L} l^2\right)^{3/2}}\,. \qquad (6)$$

The adjusted $z_{\mathcal{A}}$ values are now used to compute the adjusted CI $\tilde{\mathcal{A}}$ using $w$ and $a$, that will be used in the basic bootstrap CI formula



**(a)** Robust04 50 sub-sampled topics.

**(b)** NYT corpus.

**Figure 2:** How CIs change with $r$ and significance level. In each graph the top 75% of submitted runs are compared against a BM25 baseline using AP and URisk$^{-}$. The median CI limits (over systems) are plotted as the left-hand and right-hand ends of each solid horizontal bar, and 95% intervals showing the CI limits' ranges are added. Note the different horizontal scales on the panes.

listed in Equation 3:

$$\tilde{\mathcal{A}} = \Phi\left(w + \frac{w + z_{\mathcal{A}}}{1 - a(w + z_{\mathcal{A}})}\right), \tag{7}$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution. After substituting $\tilde{\mathcal{A}}$ for $\mathcal{A}$ in Equation 3, the $BC_a$ CI has been computed.

**Significance Levels**. Colquhoun [4] notes that researchers use statistical tests to defend their work, using evidence to support the claim that an effect exists, and where they can only be wrong 5% of the time. However, if there is only a practical effect present 10% of the time, the researcher is wrong 36% of the time due to the false discovery rate. Colquhoun suggests that to avoid making incorrect claims, "*do not regard anything greater than* $p < 0.001$ *as a demonstration that you have discovered something.*" Therefore, we compare the traditional 95% confidence interval with the recommended 99.9% one.

**Bootstrapping**. We used 100,000 iterations, in line with Smucker et al. [13]. For 250 topics, the bootstrap iterations are proportionally increased to 500,000.

## 3 RESULTS

Hesterberg et al. [8] note that the shape of a distribution can be found by interpreting the bootstrap replicates of the difference in the sample statistic. If these values do note conform to a symmetrical bell-curve, a parametric test may lead to false claims. Additionally, bootstrapping can be used to form a "confidence interval" (CI) around the statistic of interest. Figure 1 illustrates the disagreement when varying $r$, comparing the best submitted run in the Robust04 track (`pircRB04td2`) against the fixed BM25 run from Indri 5.11. As $r$ increases, the CIs increasingly disagree. That divergence is

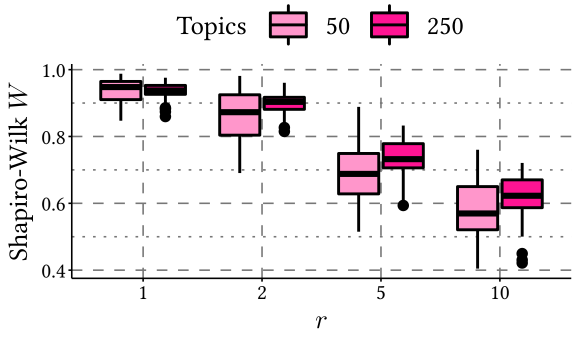


**Figure 3:** Shapiro-Wilk normality test values using 50 or 250 Robust04 topics. Larger $W$ values show increased support for normality. The top 75% of submitted runs are compared against a BM25 baseline for URisk$^{-}$ on the AP measure.

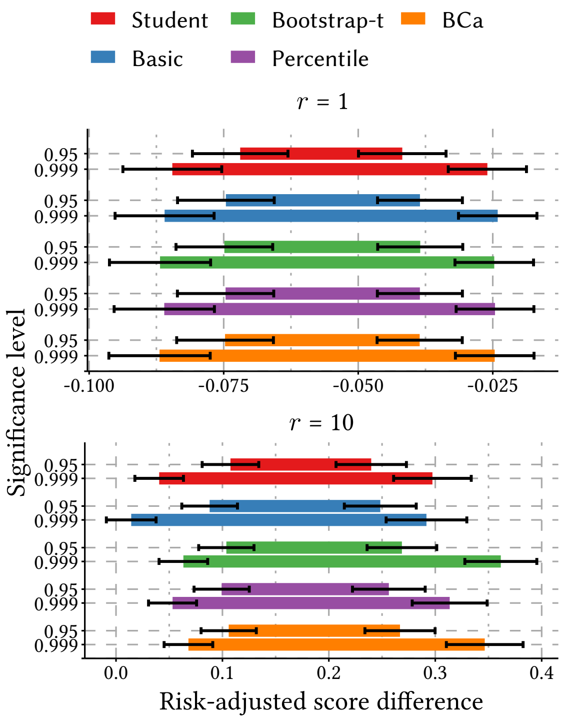


**Figure 4:** All 250 Robust04 topics; corresponding with Figure 2a. There is more agreement in the CIs, suggesting that parametric tests may be used with larger topic samples.

largely due to the asymmetry in the spread of risk values, which erodes the parametric assumption.

Figure 2 takes the majority of the runs submitted to the NYT and Robust04 tracks, as ordered by AP score, with the bottom 25% discarded, and compares them to the baseline BM25 run. For each run the CIs at 95% and 99.9% significance levels are computed, with no risk bias ($r = 1$) and high risk bias ($r = 10$), and the corresponding CI limits computed. For each collection and significance level the CI limit distributions are characterized by plotting the medians (over systems) of the lower and upper ends of the sets of paired system comparisons as a solid bar, and then computing matching "meta-confidence" intervals for those two sets of medians. That is, the thick colored bars show the range between the medians of the lower and upper CI limits for the pairwise system comparisons, and the two horizontal lines superimposed on each bar indicate the 95% confidence intervals on those endpoint values. Those limits are computed using the McGill method which utilizes the CI interquartile ranges from all systems in the comparison.

In the two $r = 1$ panes, the score distributions indicate that most of the systems outperformed the baseline BM25 run (negative score deltas); however, when $r = 10$ the scaling of losses means that the risk-biased score deltas are positive, and that risk has increased. Moreover, for larger values of $r$, the parametric and non-parametric CI methods disagree, with the discrepancy further emphasized at higher significance levels. We answer **RQ1** by placing more stock in the correctness of non-parametric CIs – when $r = 10$ the parametric assumptions are clearly violated (but also reiterate the caveat there is no ground truth for a right or wrong statistical inference).

Figure 2 explored confidence interval disagreement when the topic sample size is 50. However when a sufficiently large topic set is available, the CIs will converge towards normality due to the central limit theorem. Figure 3 supports that observation. Using Robust04, as the topic sample size increases, the Shapiro-Wilk test [12] shows stronger support for normality with 250 topics than for 50 topics, at each of the four tested risk levels $r$. That is why the confidence intervals computed over the 250 topics show more agreement between parametric and non-parametric tests (shown in Figure 4), answering **RQ2**.

## 4 CONCLUSIONS

We have explored how confidence intervals change based on different risk computations in two different collections – Robust04 and NYT. We observed that CIs disagree for evaluations carried out over 50 topics, but improve when collections reach 250 topics. Therefore, when evaluating typical IR collections of 50 topics using risk-biased evaluations, non-parametric tests should be preferred. In future work, we intend to use these observations to inform a risk inference framework that supports multiple system comparisons.

**Acknowledgments**. The first author was supported by an RMIT VCPS Scholarship. The third and fourth authors were supported by the Australian Research Council (project DP190101113). Christina Knudson assisted by locating some relevant material.

## REFERENCES

[1] J. Allan, D. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. M. Voorhees. TREC 2017 common core track overview. In *Proc. TREC*, 2017.
[2] R. Benham, A. Moffat, and J. S. Culpepper. On the pluses and minuses of risk. In *Proc. AIRS*, 2019. To appear.
[3] A. Canty, B. Ripley, et al. boot: Bootstrap R (S-Plus) functions. *R package version*, 1(7), 2012.
[4] D. Colquhoun. An investigation of the false discovery rate and the misinterpretation of $p$-values. *Royal Society Open Science*, 1(3):140216, 2014.
[5] M. Cowles and C. Davis. On the origins of the .05 level of statistical significance. *Am. Psychologist*, 37(5):553, 1982.
[6] A. C. Davison and D. B. Hinkley. *Bootstrap Methods and Their Application*, volume 1. Cambridge University Press, 1997.
[7] B. T. Dinçer, C. Macdonald, and I. Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proc. SIGIR*, pages 23–32, 2014.
[8] T. Hesterberg, S. Monaghan, D. S. Moore, A. Clipson, and R. Epstein. Bootstrap methods and permutation tests. In *The Practice of Business Statistics*, chapter 18. WH Freeman & Co., 2003.
[9] C. M. R. Kitchen. Nonparametric vs parametric tests of location in biomedical research. *Am. J. Ophthalmology*, 147(4):571–572, 2009.
[10] B. K. Nayak and A. Hazra. How to choose the right statistical test? *Indian J. Ophthalmology*, 59(2):85, 2011.
[11] T. Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015. In *Proc. SIGIR*, pages 5–14, 2016.
[12] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965.
[13] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. SIGIR*, pages 623–632, 2007.
[14] M. D. Smucker, J. Allan, and B. Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proc. SIGIR*, pages 630–631, 2009.
[15] E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proc. TREC*, pages 69–77, 2004.
[16] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proc. SIGIR*, pages 761–770, 2012.
[17] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proc. CIKM*, pages 571–580, 2008.