Unbiased Similarity Estimators Using Samples

Conrado Martínez[1(✉)], Alfredo Viola[2], and Jun Wang[1]

[1] Department of Computer Science, Universitat Politècnica de Catalunya, Barcelona, Spain
conrado@cs.upc.edu, jun.wang@estudiantat.upc.edu

[2] Instituto de Computación, Universidad de la República, Montevideo, Uruguay
viola@fing.edu.uy

Abstract. Computing a similarity measure (or a distance) between two complex objects is a fundamental building block for a huge number of applications in a wide variety of domains. Since many tasks involve computing such similarities among many pairs of objects, many algorithmic techniques and data structures have been devised in the past to reduce the number of similarity computations and to reduce the complexity of computing the similarity (e.g., dimension-reduction techniques). In this paper, we focus on computing the similarity of two sets and show that computing the similarity of two random samples drawn from the respective sets leads to an (asymptotically) unbiased estimator of the true similarity, with relative standard error going to zero as the size of the involved sets grows, and of course at a much lower computational cost as we compute the similarity of the significantly smaller samples. While this result has been known for a long time since Broder's seminal paper (Broder, 1997) for the Jaccard similarity index, we show here that the result also holds for many other similarity measures, such as the well-known cosine similarity, Sørensen-Dice, the first and second Kulczynski coefficients, etc.

1 Introduction

There are numerous applications where we need to evaluate the similarity (or the distance) among many pairs of complex objects, for example to locate the best matches to some target or to group a collection of objects into clusters of similar objects. In order to reduce the complexity of such tasks, several approaches have been proposed in the literature, and we can classify them, roughly speaking, in two families: 1) one is trying to reduce the total number of similarity/distance evaluations exploiting properties of the metric space (most notably the triangle inequality) and organizing the information into a suitable data structure, such as

This work has been supported by funds from the MOTION Project (Project PID2020-112581GB-C21) of the Spanish Ministry of Science & Innovation MCIN/AEI/10.13039/501100011033.

O. Pedreira and V. Estivill-Castro (Eds.): SISAP 2023, LNCS 14289, pp. 56–63, 2023.
https://doi.org/10.1007/978-3-031-46994-7_5

vantage-point trees, Burkhard-Keller trees, GHTs or GNATs (see for instance [2, 3,9]); 2) a second approach is to use a different similarity or distance measure, one that is much simpler to evaluate while approximating the real similarity well enough, effectively reducing the dimensionality of the search space; this is, for example, the approach in the paper by Broder [1] and the approach taken here. There are many more techniques for dimensionality reduction and even those which combine both approaches, like *Locality-Sensitive Hashing* (LSH) (see for example [8,9]).

In this work, we focus on the problem of estimating the similarity $\sigma(A, B)$ of two sets A and B. It is often the case that we will need to sort the two sets and scan them once sorted to compute their intersection, their union or their symmetric difference in order to compute their similarity $\sigma(A, B)$. Moreover, this has often to be repeated many times between different pairs of sets (the sorting step can be done just once for each set). In this scenario, when the similarity of some set A with many others has to be computed, it makes sense instead to preprocess A in linear time to extract a sample S_A of significantly smaller size ($|\mathsf{S}_A| \ll |A|$) which will be used to do the similarity evaluation. This is also the idea behind *minhashing* (see [8] and references therein), which guarantees that the probability that two sets have the same *minhash* is equal to their Jaccard similarity. In our case, we propose computing $\sigma(\mathsf{S}_A, \mathsf{S}_B)$[1], and we provide a formal proof that $\sigma(\mathsf{S}_A, \mathsf{S}_B)$ is an unbiased estimator of $\sigma(A, B)$, for several different similarity measures σ, including Jaccard and the cosine similarity, but many others as well.

Suppose we have two random samples S_A and S_B of A and B, respectively. Our goal is to prove that $\sigma(\mathsf{S}_A, \mathsf{S}_B)$ is an unbiased estimator or otherwise show how to correct the bias. Moreover, the accuracy (as measured by the standard relative error $\sqrt{\mathbb{V}\{X\}}/\mathbb{E}\{X\}$) of the estimation will depend on the size of the samples, and our goal is to quantify it in precise terms. A detailed knowledge of how the accuracy depends on the size of the samples is henceforth fundamental to obtain the desired compromise between accuracy and computational efficiency.

We will consider in this work many different similarity measures, presented in Table 1. For more detailed information about these measures, see for instance [4].

Table 1. Several similarity measures between two sets A and B

Jaccard	$J(A,B) = \frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	Cosine	$\cos(A,B) = \frac{\lvert A\cap B\rvert}{\sqrt{\lvert A\rvert\cdot\lvert B\rvert}}$
Sørensen-Dice	$\mathrm{SD(A,B)} = 2\frac{\lvert \mathrm{A}\cap\mathrm{B}\rvert}{\lvert \mathrm{A}\rvert+\lvert \mathrm{B}\rvert}$	Correlation	$\mathrm{corr(A,B)} = \cos^2(\mathrm{A,B}) = \frac{\lvert \mathrm{A}\cap\mathrm{B}\rvert^2}{\lvert \mathrm{A}\rvert\cdot\lvert \mathrm{B}\rvert}$
Kulczynski 1	$K_1(A,B) = \frac{\lvert A\cap B\rvert}{\lvert A\triangle B\rvert}$	Kulczynski 2	$K_2(A,B) = \frac{1}{2}\left(\frac{\lvert A\cap B\rvert}{\lvert A\rvert} + \frac{\lvert A\cap B\rvert}{\lvert B\rvert}\right)$
Simpson	$\mathrm{Simpson(A,B)} = \frac{\lvert \mathrm{A}\cap\mathrm{B}\rvert}{\min(\lvert \mathrm{A}\rvert,\lvert \mathrm{B}\rvert)}$	Braun-Blanquet	$\mathrm{BB(A,B)} = \frac{\lvert \mathrm{A}\cap\mathrm{B}\rvert}{\max(\lvert \mathrm{A}\rvert,\lvert \mathrm{B}\rvert)}$
Containment	$c(A,B) = \frac{\lvert A\cap B\rvert}{\lvert A\rvert}$		

[1] The random samples S_A and S_B need a final "filtering" phase before they can be used to compute the similarity $\sigma(\mathsf{S}_A, \mathsf{S}_B)$.

The main results of this paper can be summarized as follows: 1) For all the similarity measures in Table 1, the similarity of the random samples (after some appropriate "filtering") is an asymptotically unbiased estimator of the similarity of the corresponding sets. We have extended the results of [1] by giving general tools which help to establish the same result for many measures, eventually they could be used for some not considered here; 2) The standard relative error of the estimations goes to 0 as the size of the sets grows; something that cannot happen unless the size of the samples grows with the size of the sets. While the result is far from surprising, no previous work in the literature addressed the quantification of the standard error of the estimations, in particular, this hadn't been studied when the size of the samples is a function of the size of the sets.

Structure of the Paper. In Sect. 2 we will briefly review Affirmative Sampling, the random sampling algorithm which we assume will be used to draw random samples from the sets; it is our choice because the size of the returned samples grows with the (unknown) size of the set from which we sample, without prior knowledge of the set. This is useful in contexts in which we are presented the sets in an on-line fashion or when we have actually multi-sets (for example, text documents) and we measure their similarity in terms of the underlying sets of distinct elements (in the example, the *vocabularies* of the text documents).

In Sect. 3 we present the main results of this paper, namely, that the similarity of samples is an unbiased estimator of the similarity of the sets, for each one of the similarity measures shown in Table 1. The formal proofs of our results appear in the full version [7], not here, due to space constraints.

After that, we present in a short section (Sect. 4) the results of a small empirical study that we have conducted, showing significant accordance with the theoretical results of the previous section. We close in Sect. 5 with some final remarks and a discussion about future developments of this line of research.

2 Sampling

Let $A \subseteq \mathcal{U}$ be a finite subset of the domain $\mathcal{U}$ (also finite, but potentially extremely large). Assume that to every element $x \in \mathcal{U}$ we have assigned a random number $h(x) \in [0,1]$, the outcome of an independent draw from a uniformly distributed random variable in $[0,1]$. Then for any τ, $0 < \tau < 1$, the subset $A^{\geq\tau} = \{x \in A \,|\, h(x) \geq \tau\} \subseteq A$ is a random sample of A. Any of the n elements of A has exactly the same probability of belonging to $A^{\geq\tau}$ as any other element in A. Likewise, if we consider the subset A_k of A with the k elements of A with larger (smaller) value of h then A_k is also a random sample. In particular, if $\tau = \min\{h(x) \,|\, x \in A_k\}$ is the minimum h value in A_k then $A_k = A^{\geq\tau}$.

In practice, to obtain such random samples we can use a hash value $h(x)$ for each element, as presented in [1]; under pragmatic assumptions we can safely neglect the probability of collisions. Looking at the hash values as real numbers in $[0,1]$, for any $x \in \mathcal{U}$ and any value $z \in [0,1]$ we assume that $\Pr\{h(x) \leq z\} = z$, for a reasonably well-chosen hash function h; that is, the hash values are uniformly distributed.

If the size n of the set A from which we want to draw a sample is known beforehand, then we can take $k = k(n)$ and draw a sample of size k as described above. But even if the size of the set A were not known in advance (and we do not want to incur the costs in time and space to compute it) we can still draw random samples of variable size, growing with the (unknown) size n of the set. This can be accomplished thanks to Affirmative Sampling (AS) [6], an easy and practical alternative to just keeping the k elements with largest (smallest) hash values in A. AS also uses a fixed parameter k but produces samples of size $\geq k$ (unless $n < k$). The expected size $\mathcal{S} = |\mathsf{S}|$ of the random sample S produced by AS is $\mathbb{E}\{\mathcal{S}\} = k \ln(n/k) + \text{l.o.t.}$ with $n = |A|$. Easy variants of AS will produce random samples of size $\Theta(n^\alpha)$ for a fixed given α, $0 < \alpha < 1$ (see [6]).

Whether we use fixed-size samples or variable-size samples, as long as they are random we can make inferences about the "population" (the set A) from the sample S, for instance, about the proportion of elements in A that satisfy a certain property P. Let $n_P = |A_P|$, with A_P the subset of elements satisfying P, and n the number of elements in A. Denote $\vartheta_P := \frac{n_P}{n}$ the fraction of elements that satisfy P. Let us assume in the computations below and for the remaining of the paper that $n \geq \mathcal{S} \geq k \geq 2$, that is, that the sampling algorithm will return at least $k \geq 2$ elements. Otherwise, if the set contains less than k distinct elements, the sample contains **all** elements in A and their relevant statistics, and we can answer queries exactly.

If we have a random sample S of A of size $\mathcal{S}$ and $\mathsf{S}_P = \mathsf{S} \cap A_P$ then it is well known that the proportion $\hat{\vartheta}_P = \mathcal{S}_P/\mathcal{S}$ is an unbiased estimator of ϑ_P, even when $\mathcal{S}$ is a random variable and not a fixed value (see, for example, [6] and references therein). Quite intuitively, the accuracy of the estimator will depend on the size of the sample. The result giving its variance can be found in many places, however the size $\mathcal{S}$ of the sample is assumed fixed there. The more general statement given here, when $\mathcal{S}$ is a random variable, can be found, together with its proof, in [6].

Lemma 1. *The random variable $\hat{\vartheta}_P := \mathcal{S}_P/\mathcal{S}$, where $\mathcal{S}$ is the size of the random sample S and $\mathcal{S}_P$ is the number of elements in S that satisfy P, is an unbiased estimator of $\vartheta_P := n_P/n$, that is, $\mathbb{E}\left\{\hat{\vartheta}_P\right\} = \vartheta_P$, assuming that $n \geq \mathcal{S} > 0$. Moreover,*

$$\mathbb{V}\left\{\hat{\vartheta}_P\right\} = \frac{n_P(n - n_P)}{n(n-1)} \cdot \left(\mathbb{E}\left\{\frac{1}{\mathcal{S}}\right\} - \frac{1}{n}\right).$$

If the behavior of the random variable $\mathcal{S} = |\mathsf{S}|$ is smooth enough[2] and $\mathbb{E}\{\mathcal{S}\} \to \infty$ when $n \to \infty$ then the accuracy of the estimator $\hat{\vartheta}_P$ will improve, as the variance will decrease and tend to 0 as $n \to \infty$.

[2] One can prove that in many random sampling schemes, in particular for AS, we have $\mathbb{E}\{1/\mathcal{S}\} = \mathcal{O}(1/\mathbb{E}\{\mathcal{S}\})$.

3 Estimating Similarity

Consider now two sets A and B and two random samples S_A and S_B, respectively, such that $\mathsf{S}_A = A^{\geq \tau_{\mathsf{S}_A}}$ and $\mathsf{S}_B = B^{\geq \tau_{\mathsf{S}_B}}$. Looking at Table 1 we quickly notice that many of the measures are of the form $|C_P|/|C|$, for some set C built from A and B, and some subset C_P of C of elements that satisfy a certain property P. For example, for the Jaccard similarity we have $C = A \cup B$ and the property P is "x belongs to both A and B" ($C_P = A \cap B$). To apply Lemma 1 we need to figure out how to obtain a random sample S_C of C out of the given random samples S_A and S_B, and how to find the elements in S_C which satisfy P. Our random sampling scheme guarantees that $\mathsf{S}_A = \{x \in A \,|\, h(x) \geq \tau_{\mathsf{S}_A}\}$ and $\mathsf{S}_B = \{x \in B \,|\, h(x) \geq \tau_{\mathsf{S}_B}\}$. Given two sets X and Y, let $\tau^*(X, Y) := \max\{\tau_{X \setminus Y}, \tau_{X \cap Y}, \tau_{Y \setminus X}\}$; we will take the convention that $\tau_\emptyset = 0$. Let $\tau = \tau^*(\mathsf{S}_A, \mathsf{S}_B)$. Then we can show that

$$\mathsf{S}_A^{\geq \tau} \cup \mathsf{S}_B^{\geq \tau} = (\mathsf{S}_A \cup \mathsf{S}_B)^{\geq \tau} = (A^{\geq \tau} \cup B^{\geq \tau}) = (A \cup B)^{\geq \tau},$$

that is: "filtering" S_A and S_B according to the largest threshold $\tau = \tau^*(\mathsf{S}_A, \mathsf{S}_B)$ and taking the union of these filtered samples we get a random sample of $A \cup B$. The filtering trick can also be used to produce a random sample of $A + B$ from random samples S_A and S_B, for $A \times B$, etc. This allows us to prove our first main result (its proof appears in the full version [7] of this paper).

Theorem 1. *Let σ be any of the similarity measures: Jaccard, Sørensen-Dice, containment coefficient, Kulczynski 2 (second Kulczynski coefficient) or correlation coefficient. Let S_A and S_B be random samples of A and B such that $\mathsf{S}_A = A^{\geq \tau_{\mathsf{S}_A}}$ and $\mathsf{S}_B = B^{\geq \tau_{\mathsf{S}_B}}$, and let $\tau = \tau^*(\mathsf{S}_A, \mathsf{S}_B) = \max(\tau_{\mathsf{S}_A \setminus \mathsf{S}_B}, \tau_{\mathsf{S}_B \setminus \mathsf{S}_A}, \tau_{\mathsf{S}_A \cap \mathsf{S}_B})$. Then $\hat{\sigma} = \sigma(\mathsf{S}_A^{\geq \tau}, \mathsf{S}_B^{\geq \tau})$ is an unbiased estimator of $\sigma(A, B)$, that is,*

$$\mathbb{E}\left\{\sigma(\mathsf{S}_A^{\geq \tau}, \mathsf{S}_B^{\geq \tau})\right\} = \sigma(A, B).$$

Moreover,

$$\mathbb{V}\left\{\sigma(\mathsf{S}_A^{\geq \tau}, \mathsf{S}_B^{\geq \tau})\right\} \sim \sigma(A, B) \cdot (1 - \sigma(A, B)) \cdot \mathcal{O}\left(\mathbb{E}\left\{\frac{1}{\min(|\mathsf{S}_A|, |\mathsf{S}_B|)}\right\}\right),$$

which implies that $\mathbb{V}\{\hat{\sigma}\} \to 0$ if Affirmative Sampling is used to draw the samples, since then $\mathbb{E}\{1/\min(|\mathsf{S}_A|, |\mathsf{S}_B|)\} \to 0$, if $|A|, |B| \to \infty$.

The same result holds for Simpson and Braun-Blanquet measures, provided that we know which of $|A|$ and $|B|$ is smaller (larger). If the sizes of A and B are unknown, we cannot assume that $|\mathsf{S}_A|$ and $|\mathsf{S}_B|$ have the same relation since these sizes are, in general, random variables. If we put $\min(|\mathsf{S}_A|, |\mathsf{S}_B|)$ ($\max(|\mathsf{S}_A|, |\mathsf{S}_B|)$, resp.) in the denominator of the Simpson estimator (Braun-Blanquet, resp.), we will have a bias, albeit the experiments suggest it is not very significant (see Sect. 4). The formal result, as well as a detailed discussion about the bias that we get when not knowing the sizes of A and B, can be found in the full version of the paper [7].

Last, but not least, neither the cosine similarity nor Kulczynksi 1 fit into the framework of Theorem 1. However, both can be written as functions of other similarity measures for which we have unbiased estimators. Namely, for the cosine similarity, we have $\cos(A,B) = \sqrt{\text{corr}(A,B)}$. For Kulczynksi 1 we can write

$$K_1(A,B) = \frac{|A \cap B|}{|A \triangle B|} = \frac{1}{\frac{|A \cup B|}{|A \cap B|} - 1} = \frac{J(A,B)}{1 - J(A,B)}.$$

That is, these measures σ' are such that $\sigma'(A,B) = f(\sigma(\ A,B))$ for a similarity measure which can be estimated without bias: $\mathbb{E}\left\{\sigma(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} = \sigma(A,B)$. In general, given a random variable X, $\mathbb{E}\{f(X)\} \neq f(\mathbb{E}\{X\})$. However, if f is a smooth function in $(0,1)$, that is, $f \in \mathcal{C}^\infty(0,1)$, and the size of the samples grows with the size of the sets then we will have[3]

$$\begin{aligned}\mathbb{E}\left\{\sigma'(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} = \mathbb{E}\left\{f(\sigma(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} &\sim f\left(\mathbb{E}\left\{\sigma(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\}\right) \\ &= f\left(\sigma(A,B)\right) = \sigma'(A,B).\end{aligned}$$

We can therefore obtain asymptotically unbiased estimators for cosine and Kulczynski 1 (first Kulczynski coefficient) using $f(x) = \sqrt{x}$ and $\sigma = \text{corr}$ for the former, and $f(x) = x/(1-x)$ and $\sigma = \text{Jaccard}$ for the latter. The proof that $\mathbb{E}\{f(X)\} \sim f(\mathbb{E}\{X\})$ under the appropriate hypotheses is given in [7].

The estimator for these similarity measures is only asymptotically unbiased, and that's because all the central moments of order $r \geq 2$ of $\sigma(A,B)$ tend to 0 as the size of the sets grows; however f or any of its derivatives might be very large, in particular when $\sigma(A,B) \to 0$ or $\sigma(A,B) \to 1$, and then the bias can be significant when the asymptotic regime hasn't been reached yet.

The variance of $\sigma'(\mathsf{S}_A^{\geq\tau}, fsB) = f(\sigma(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau}))$ can also be found using the same technique that we have used to establish that the estimator is asymptotically unbiased, by considering the Taylor series expansion of f^2. It is easy to show then that we have

$$\mathbb{V}\left\{\sigma'(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} = \frac{1}{2}\left.\frac{d^2}{dx^2} f^2(x)\right|_{x=\sigma(A,B)} \cdot \mathbb{V}\left\{\sigma(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} + \text{l.o.t.},$$

and since $\mathbb{V}\left\{\sigma(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} \to 0$ when $\min(|A|,|B|) \to \infty$, we know that the variance $\mathbb{V}\left\{\sigma'(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} \to 0$ too (and at which rate). However, $(f^2(x))''$ might be very large (actually tend to ∞) for $\sigma(A,B) \to 1$ or $\sigma(A,B) \to 0$, and then the variance of $\sigma'(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})$ will be non-negligible in practical settings when the similarity of the two sets is very close to 0 or to 1. For example, for the cosine similarity we have $f(x) = \sqrt{x}$ and $(f^2)'' = 0$ which entails a very low variance of the estimator $\cos(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})$ even if the similarity of the two sets is close to 0 or to 1. However, for the first Kulczynski coefficient, we have $f(x) = x/(1-x)$ giving $(f^2)'' = (4x+2)/(1-x)^4$. Hence, when the

[3] $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$.

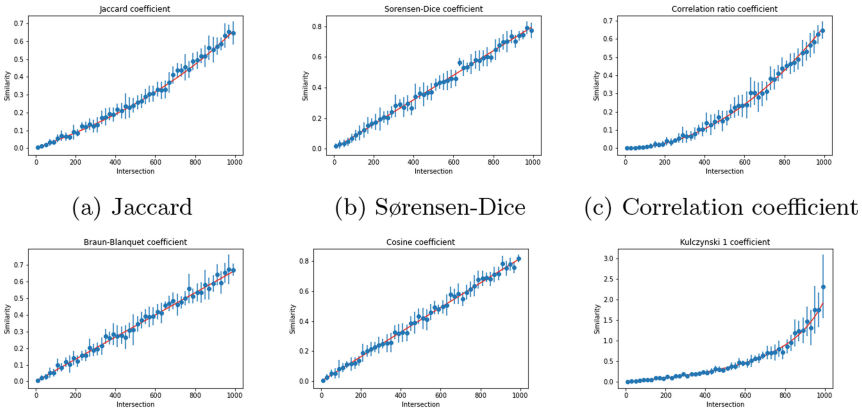
similarity $J(A,B)$ of the two sets is very close to 1 we have $K_1(A,B) \to \infty$ and $\mathbb{V}\left\{K_1(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} \sim \frac{3}{(1-J(A,B))^3} \cdot \mathcal{O}\left(1/\left(k \ln((|A| \cup |B|)/k)\right)\right)$, which will be quite large unless $|A| \cup |B|$ is really huge. This phenomenon shows clearly in the plot Fig. 1f given in next section.

4 Experimental Results

We have conducted a small experimental study with the aim to show representative examples of the good match between our theoretical findings and the estimates obtained in practice.

Due to the lack of space, we will report here only one experiment in which we work with two fixed sets $A = \{1, \ldots, m\}$ and $B = \{r, \ldots, r+n-1\}$, for some $r \leq m+1$. Changing the value of r the intersection $|A \cap B|$ will run from 0 (if $r = m+1$) to $\min(m,n)$ (if $r = 1$). In particular, we have chosen $|A| = m = 1000$ and $|B| = n = 1500$. We apply the sampling algorithm $T = 10$ times on each set, with a different randomly chosen hash function each time, thus effectively producing T different estimates of the similarity $\sigma(A,B)$. The plots in Fig. 1 show the true similarity ($\sigma(A,B)$, red line), the estimates (blue dots) and the standard deviation in the T observations (length of the blue bars), as the size of the intersection varies from 0 to $\min(m,n) = 1000$ for some of the similarity measures studied in this paper.

The second experiment, reported in [7], studies how the quality of the estimates impact the application using them, in particular, we have studied how the clusterings produced by k-means change when we use estimates of the similarities instead of the real similarities.



(a) Jaccard (b) Sørensen-Dice (c) Correlation coefficient

(d) Braun-Blanquet (e) Cosine (f) Kulczynski 1

Fig. 1. Empirical estimates of several similarity measures

5 Final Remarks

We show in this paper that many similarity measures between sets can be accurately estimated, going further beyond Broder's results [1] on the Jaccard and containment indices. One fundamental ingredient for (asymptotically) unbiased estimation with a high degree of accuracy is the use of variable-size sampling. Using Affirmative Sampling, the (expected) size of the samples increases with the size of the sets, and the standard relative error in the estimations goes to 0 [6].

Another line of research that we are already working on is the estimation of similarity measures between objects other than sets. For example, multisets, sequences, ... A notable example is partitions (clusterings). For instance, instead of examining all the $\binom{N}{2}$ pair of objects to compute the Rand index (a measure of similarity, see for instance, [5]) of two clusterings of a set of N objects, we can draw two random samples S and S', both containing $\ll N$ elements, form all ordered pairs combining distinct elements from S and S', and estimate the Rand index from those. It is straightforward to show that this is an unbiased estimate of the true Rand index using our techniques, and we think that they can be used to show the same for many other similarity measures between clusterings.

References

1. Broder, A.Z.: On the resemblance and containment of documents. In: Carpentieri, B., De Santis, A., Vaccaro, U., Storer, J.A. (eds.) Proceedings of the Compression and Complexity of SEQUENCES 1997, pp. 21–29. IEEE Computer Society (1997)
2. Chávez, E., Navarro, G., Baeza-Yates, R.A., Marroquín, J.L.: Searching in metric spaces. ACM Comput. Surv. **33**(3), 273–321 (2001)
3. Chen, L., Gao, Y., Song, X., Li, Z., Miao, X., Jensen, C.S.: Indexing metric spaces for exact similarity search. ACM Comput. Surv. **55**(6), 128:1–128:39 (2022). https://doi.org/10.1145/3534963
4. Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer, Berlin (2009). https://doi.org/10.1007/978-3-642-00234-2
5. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
6. Lumbroso, J., Martínez, C.: Affirmative sampling: theory and applications. In: Ward, M.D. (ed.) Proceedings of the 33^rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA), vol. 225 of Leibniz International Proceedings in Informatics (LIPIcs), , Dagstuhl, Germany, pp. 12:1–12:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2022). https://doi.org/10.4230/LIPIcs.AofA.2022.12
7. Martínez, C., Viola, A., Wang, J.: Fast and accurate similarity estimation using samples (2023). https://doi.org/10.13140/RG.2.2.24053.14566
8. Rajaraman, A., Leskovec, J., Ullman, J.D.: Mining Massive Datasets, 3rd edn. Cambridge University Press, Cambridge (2014). https://www.mmds.org/
9. Samet, H.: Foundations of Multidimensional and Metric Data Structures. Morgan Kaufmann, Boston (2006)