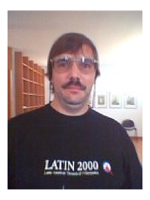
Unbiased Similarity Estimators Using Samples

A Coruña, Spain, October 9–11, 2023


Conrado Martínez
U. Politècnica de Catalunya
Barcelona, Spain
conrado@cs.upc.edu

Alfredo Viola
Univ. de la República
Montevideo, Uruguay
viola@fing.edu.uy

Jun Wang
U. Politècnica de Catalunya
Barcelona, Spain
jun.wang@estudiantat.upc.edu

Abstract

Many fundamental applications in Computer Science, Data Science and others require a large amount of distance or similarity evaluations among pairs of objects; when the objects are complex (e.g., audio, video, images, large texts, genome or protein sequences, . . .) each similarity evaluation is already computationally expensive, and if you have to do many of them the computational cost becomes very significant. Many authors have proposed the use of *sketches* (a.k.a. *surrogates*, *fingerprints*) of the objects to speed up the similarity evaluations, for example, if the sketches of two objects are dissimilar then the objects themselves are dissimilar too. In this work, we consider similarity between sets A and B (in many applications the objects that we need to work with are already sets or can be viewed as such, e.g., we can assimilate texts with the sets of distinct words that they contain). Quite intuitively, if we pick random samples S_A and S_B from A and B, the similarity of S_A and S_B should be a good estimate of the true similarity of A and B. And this should hold for many different similarity measures, including the well known Jaccard index or the cosine similarity. We prove that $\sigma(\mathsf{S}_A, \mathsf{S}_B)$ is an (asymptotically) unbiased estimator for $\sigma(A, B)$ for many different similarity measures σ, giving a general framework to prove such result; we also establish some "post-processing" that needs to be applied to random samples in order to remove undesirable biases in the estimations, and provide a detailed mathematical analysis of the standard deviation of the estimations.

Introduction

Given sets A and B from some domain $\mathcal{U}$, random samples S_A and S_B from A and B, respectively, and a *similarity* measure σ:

- Is $\sigma(\mathsf{S}_A, \mathsf{S}_B)$ a good estimator of $\sigma(A, B)$?
- If so, is it unbiased? Asymptotically unbiased?
- How does the accuracy of the estimation relate to the size of the samples?
- For which similarity measures do we have good estimators based on random samples?
- Will any random samples do?

The pioneer work of Broder [1] gave some answers for several of these questions in the case of the Jaccard similarity and also for the so-called *containment index* $c(A, B)$ that measures how much $A \subset B$. Using random samples of fixed size, one gets unbiased estimators of both measures; the accuracy was not considered.

Applications

Good estimators of similarity using samples will be quite useful in contexts in which we have a large collection of sets $A_1, A_2, \ldots, A_N$ and we must perform many similarity evaluations $\sigma(A_i, A_j)$ or $\sigma(A_i, B)$, such as in classification tasks or in proximity searches. Substituting complex similarity evaluations by something simpler has been successfully used in many approximate search schemes, like Locality-Sensitive Hashing [2, 4].

- Extracting a random sample of fixed size k from A_i has cost $\Theta(|A_i| \log k)$
- If $|A_i|$ is known then we can set $k = k(|A_i|)$; quite intuitively, if $|A_i|$ is large we should work with larger samples
- Even if $|A|$ is not known, one can use a scheme such as Affirmative Sampling [3] which will produce a random sample of (expected) size $\log |A_i|$ or $|A_i|^c$ $(0 < c < 1)$, without prior knowledge of $|A_i|$ and using extra memory proportional to the size of the sample (not of $|A_i|$). In order to speed up later computations it is often useful to sort the samples with cost $\Theta(|\mathsf{S}_{A_i}| \log |\mathsf{S}_{A_i}|)$
- Once we have random samples S_{A_i} for each set A_i in the collection we can evaluate $\sigma(\mathsf{S}_{A_i}, \mathsf{S}_B) \approx \sigma(A_i, B)$ with cost $\Theta(|\mathsf{S}_{A_i}| + |\mathsf{S}_B|)$ vs $\Theta(|A_i| + |B|)$–assuming that we have also sorted all A_i and B
- For example, suppose we apply K-means to our collection, where K is the number of clusters. The cost of the algorithm is roughly $\mathcal{O}(N\overline{|A|} \log \overline{|A|} + N \cdot K \cdot \overline{|A|} \cdot \ell)$, where $\overline{|A|}$ is the average size of the A_i's and ℓ the number of iterations until convergence (or we stop the algorithm). Using samples the cost will be $\mathcal{O}(N\overline{|A|} + N\overline{|\mathsf{S}_A|} \log \overline{|\mathsf{S}_A|} + N \cdot K \cdot \overline{|\mathsf{S}_A|} \cdot \ell)$, now with a factor $\overline{|\mathsf{S}_A|} \ll \overline{|A|}$ in the main term.

A Few Technicalities

Many similarity measures are of the form $\frac{|C_P|}{|C|}$, where $C = A \oplus B$ is a set that we obtain operating the sets A and B, and $C_P \subseteq C$ is the subset of elements of C that satisfies a certain property. For example,

$$\mathrm{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

If we are able to construct a random sample S_C of C $\mathsf{S}_C = \mathsf{S}_A \oplus \mathsf{S}_B$ and $\mathsf{S}_C \cap C_P$ could be easily computed as well, then $\frac{|\mathsf{S}_C \cap C_P|}{|\mathsf{S}_C|}$ will be an unbiased estimator of $\frac{|C_P|}{|C|}$. This is a very well known result in Statistics; we have shown that it is the case even if $|\mathsf{S}_C|$ is a random variable, and we have also been able to compute the variance of the estimator.

Another group of similarity measures, like the famous cosine similarity and Kulczynski 1 are of the form $f\left(\frac{|C_P|}{|C|}\right)$ for some smooth function (f is continuously and infinitely differentiable in $(0, 1)$). Despite

$$\mathbb{E}\left\{ f\left(\frac{|\mathsf{S}_C \cap C_P|}{|\mathsf{S}_C|}\right)\right\} \neq f\left(\mathbb{E}\left\{\frac{|\mathsf{S}_C \cap C_P|}{|C|}\right\}\right),$$

we show that it is true asymptotically if variable-size sampling is used, that is, if $|\mathsf{S}_A| \to \infty$ and $|\mathsf{S}_B| \to \infty$ when $|A| \to \infty$ and $|B| \to \infty$, by computing all central moments of the estimator.

Results

Assume we have a hash function $h : \mathcal{U} \to [0, 1]$, and assume that the probability of collision is negligible (provided that h has enough bits, we can safely assume that). Given a set X, we denote τ_X the smallest hash value of any element in X; given $\tau \in [0, 1]$, we denote $X^{\geq\tau} = \{x \in X \mid h(x) \geq \tau\}$ the subset of elements in X with hash value $\geq \tau$.

Under reasonable assumptions about the hash function h, $X^{\geq\tau}$ is a random sample of X. Any subset of size $k = |X^{\geq\tau}|$ from X is equally likely to be $X^{\geq\tau}$. Therefore, to get a random sample of size k from X, it is enough to collect the k elements in X with the largest hash values. There are schemes which allow $k = k(n)$, where $n = |X|$; even the need to know n in advance can be avoided: for example, Affirmative Sampling, will produce samples of expected size $\Theta(\log n)$ or $\Theta(n^c)$ even though n not known.

Theorem 1. *Let σ be any of the similarity measures: Jaccard, Sørensen-Dice, containment coefficient, cosine similarity, Kulczynski 1 (first Kulczynski coefficient), Kulczynski 2 (second) 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 (asymptotically) unbiased estimator of $\sigma(A, B)$, that is,*

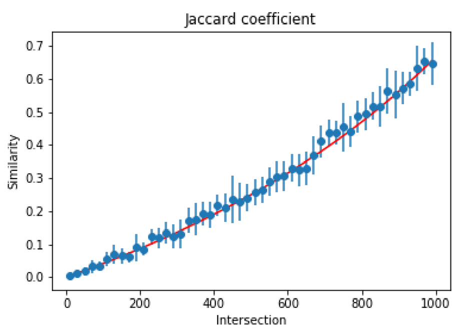
$$\mathbb{E}\left\{\sigma(\mathsf{S}_A^{\geq\tau}, \mathsf{S}_B^{\geq\tau})\right\} \sim \sigma(A, B).$$

Moreover,

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

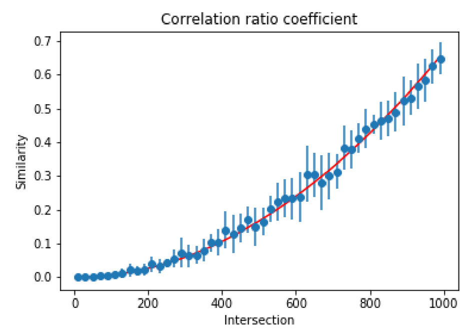
which implies that $\mathbb{V}\{\hat{\sigma}\} \to 0$ if $\min(|\mathsf{S}_A|, |\mathsf{S}_B|) \to \infty$ as $|A|, |B| \to \infty$.

A similar result holds for other similarity measures like Simpson and Braun-Blanquet.

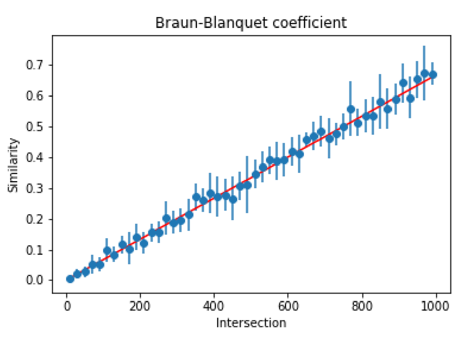


Jaccard

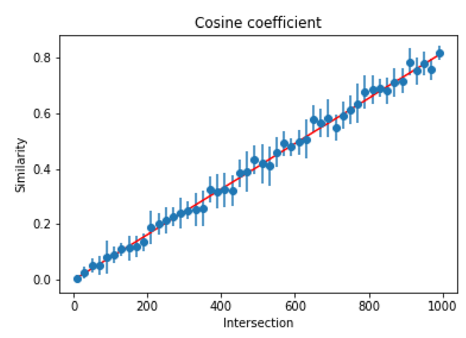
Sørensen-Dice



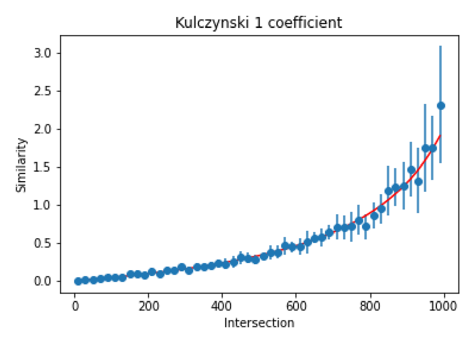
Correlation coeff.



Braun-Blanquet



Cosine



Kulczynski 1

Empirical estimates of several similarity measures

The x-axis in the plots above shows the size of the intersection of two sets $A = \{z_1, \ldots, z_m\}$ and $B = \{z_r, \ldots, z_{r+n-1}\}$, ranging from 0 $(r = m + 1)$ to $\min(m, n)$ $(r = 1)$. In the experiments $m = |A| = 1000$ and $n = |B| = 1500$. The red solid lines show the value of $\sigma(A, B)$. The blue dots show the average of $T = 10$ estimations (sampling T times in each set); the blue bars depict the standard variation.

Conclusions

The similarity of random samples can be used to accurately estimate the similarity of the sets they represent. The samples being of significantly smaller size than the objects, these estimations can be carried out using a tiny fraction of the computational resources one would need to compute the "true" similarity. Some post-processing of the random samples is needed to avoid bias in the estimation, but it does not introduce a serious computational penalty. We have shown that similarity estimation using random samples is possible for many similarity measures between sets, and developed general techniques which might be useful to tackle other new measures not contemplated here. Our careful and solid mathematical analysis (we haven't just conducted an experimental study) should allow a precise quantitative analysis of the impact of using estimations instead of the "true" similarities in applications.

We are also working on the extension of the ideas and techniques here to other kind of objects like multisets or partitions. For example, we have recently proven that the Rand index (a well known measure of similarity) of two partitions of an N-element set can be accurately estimated without checking the $\binom{N}{2}$ possible pairs of distinct elements.

References

[1] Andrei Z. Broder. On the resemblance and containment of documents. In B. Carpentieri, A. De Santis, U. Vaccaro, and J.A. Storer, editors, *Proc. of the Compression and Complexity of SEQUENCES 1997*, pages 21–29. IEEE Computer Society, 1997.

[2] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In Malcolm P. Atkinson, Maria E. Orlowska, Patrick Valduriez, Stanley B. Zdonik, and Michael L. Brodie, editors, *Proc. of the 25th Very Large Database Conference (VLDB)*, pages 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Pub. Inc.

[3] Jérémie Lumbroso and Conrado Martínez. Affirmative sampling: Theory and applications. In Mark Daniel Ward, editor, *Proc. of the 33rd Int. Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA)*, volume 225 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 12:1–12:17, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[4] Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman. *Mining Massive Datasets*. Cambridge University Press, 3rd edition, 2014. Available on-line from http://www.mmds.org/.