Modelling String Structure in Vector Spaces

Richard Connor[1][0000−0003−4734−8103], Alan Dearle[2][0000−0002−1157−2421], and Lucia Vadicamo[3][0000−0001−7182−7038]

[1] Division of Mathematics and Computing Science, University of Stirling, Scotland
richard.connor@stir.ac.uk
[2] School of Computer Science, University of St Andrews, St Andrews, Scotland
alan.dearle@st-andrews.ac.uk
[3] Institute of Information Science and Technologies (ISTI), CNR, Pisa, Italy
lucia.vadicamo@isti.cnr.it

Abstract. Searching for similar strings is an important and frequent database task both in terms of human interactions and in absolute world-wide CPU utilisation. A wealth of metric functions for string comparison exist. However, with respect to the wide range of classification and other techniques known within vector spaces, such metrics allow only a very restricted range of techniques. To counter this restriction, various strategies have been used for mapping string spaces into vector spaces, approximating the string distances within the mapped space and therefore allowing vector space techniques to be used.
In previous work we have developed a novel technique for mapping metric spaces into vector spaces, which can therefore be applied for this purpose. In this paper we evaluate this technique in the context of string spaces, and compare it to other published techniques for mapping strings to vectors. We use a publicly available English lexicon as our experimental data set, and test two different string metrics over it for each vector mapping. We find that our novel technique considerably outperforms previously used technique in preserving the actual distance.

Keywords: Metric Mapping · n-Simplex projection · Pivoted embedding · String · Jensen-Shannon distance · Levenshtein distance

1 Introduction

Searching over strings has been a central activity of database systems since their inception. In addition to their use in documents composed of character sequences, strings are commonly used to encode the structure of molecules, DNA sequences, product descriptions and many other objects drawn from the real world. The need to search for similar strings arises due to differences in encoding and spelling, textual descriptions being slightly different, and the need to find partial matches. In addition to common tasks such as Web searching, string

Copyright © 2019 for the individual papers by the papers authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors. SEBD 2019, June 16-19, 2019, Castiglione della Pescaia, Italy.

CEUR-WS.org/Vol-2400/paper-45.pdf

similarity also plays an important role in many real-world database tasks including deduplication and cleaning. Similarity search is also important in the context of record linkage where similar representations of real word entities require unification. Searching for similar strings is therefore an important and frequent database task both in terms of human interactions and in absolute worldwide CPU utilisation.

A wealth of metric functions for string comparison exist. In previous work, we have identified the *n-Simplex projection* as a way of mapping a class of metric space into lower-dimensionality Euclidean spaces with better indexing properties [1, 5, 6]. Previously we have applied this mapping only to the problem of similarity search: that is, finding those objects which are most similar to a given query object from a very large space. As part of this work we have identified a subclass of proper metrics, called *supermetric*, which are amenable to our techniques. This class includes those spaces which have the so-called four-point property [3], which in turn includes all metric spaces which can be isometrically embedded in Hilbert space. Furthermore, any Hilbert-embeddable space has a deeper property, in that any finite set of n points can be isometrically embedded in an $(n-1)$-dimensional Euclidean space.

So far, we have examined only metrics and data within $\mathbb{R}^n$ spaces. Here, we extend the work beyond metric indexing, and to non-numeric spaces for the first time. Our techniques allow certain string spaces to be represented in vector spaces in such a way that the distances are maintained with various degrees of approximation. Our main contribution is to explain the mapping from the string space to the vector space. Based on a sound mathematical framework, we demonstrate that this mapping into vector spaces is significantly better than those previously known.

2 Background and Related Work

Our domain of interest is string spaces, by which we refer to a (typically large) finite set of character strings $\mathbb{S}$, and a distance function $d : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}^+$ which compares them. With only a little loss of generality we restrict our interest further to proper metrics which are common over strings. Such spaces are very important in many domains; not just where strings themselves are the objects of interest (e.g., natural language analysis), but also where another semantic domain is represented in string form (e.g., genome analysis).

Various machine learning techniques have shown great success in the analysis and classification of natural language terms, based on a very large supervised or unsupervised input, which gives a vector space mapping reflecting the semantics of the original text [15, 16]. In such cases, the vector space acts as a semantic proxy space, in that terms with similar meaning will be close when measured in the vector space. In our context, we are interested only the in *structural* similarity of strings, as measured by metrics such as edit distance. The intent is to map into a vector space while preserving the original string distances as far as possible. There are some known techniques for mapping string spaces into vector

Table 1: The nearest few terms from the SISAP lexicon for a selection of four tokens, showing both Levenshtein and Jensen-Shannon Shingle distances

"string"		"metric"		"simplex"		"error"	
lev	jsd	lev	jsd	lev	jsd	lev	jsd
string	string	metric	metric	simples	simple	error	error
strings	stringing	matric	meteoric	dimpled	simples	terror	terror
surfing	strings	eric	metricize	triplex	simpler	mirror	err
tin	striking	dearie	metrical	bingley	simplest	frog	borrower
ringing	stirring	petted	symmetric	names	pimples	racy	emperor
briefing	staring	makeup	mimetic	jugular	simply	petted	horror

spaces, whilst attempting to preserve the integrity of the distance function, but none of these appear to be particularly effective.

This paper presents a mechanism for transforming metric string spaces to vector spaces, which can be applied whenever the distance function d is a proper metric; this includes most edit distances.

2.1 String Metrics

In the literature, many string distance functions have been proposed to measure the dissimilarity between two text strings. In this work we restrict our attention to the *Levenshtein Distance* and *Jensen-Shannon distance* as explained below. These two metrics are representative of the wide range of more or less sophisticated metrics available. For our purposes, we also note a deep mathematical property which is important to our context: the Jensen-Shannon distance is isometrically Hilbert embeddable, whereas the Levenshtein distance is not. The importance of this will be made clear later in the text.

Levenshtein Distance: the best known edit distance between strings defined as the minimum number of single-character edit operations (insertions, deletions or substitutions) needed to transform one string into the other. For example, the distance between "error" and "horror" is 2 (one insertion and one substitution), the distance between "writing" and "writer" is 3 (one deletion, two substitutions). Note that the Levenshtein distance d_{Lev} between two strings of length m and n is at least $|m - n|$, and at most $max(m, n)$.

Jensen-Shannon Distance over Strings: is derived from a distance metric defined over labelled trees [8]. In outline, the technique of digram shingling [13] is used to transform each character string into a very high-dimensional probability space, where each individual character and each character digram is represented as a different event; thus a string represented in a 127-character set will be represented in a space comprising $127^2 + 127$ different events. Each such event is assigned a frequency according to its appearance in the token, including notional start and finish characters denoted ϕ and

ω in the following example. For example, the string "ann" is composed of the characters and diagrams "ϕa","an", "nn" "nω","a", and "n" with corresponding frequencies $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$, and $\frac{1}{3}$.
Jensen-Shannon distance is then applied over these probability ensembles[4].We recall that the Jensen-Shannon distance between two probability vectors x, y is defined as $d_{JS}(x, y) = \sqrt{JSD(x, y)}$, where $JSD(x, y)$ is the Jenson-Shannon divergence.

$$JSD(x, y) = 1 - \frac{1}{2} \sum_i x_i \log_2 x_i + y_i \log_2 y_i - (x_i + y_i) \log_2 (x_i + y_i). \quad (1)$$

To show a characterisation of these two metrics, Table 1 shows the nearest few strings within the SISAP lexicon to an appropriate selection of tokens.

2.2 String space to Vector space mappings

The only information available within our context derives from the distance function over the string domain; this is treated as a "black box" function and therefore any approach to map the string space into a vector space must rely only upon these distances. Usually, the distances between each string and a fixed set of *pivots* (i.e., reference objects) is exploited to map the string space to a vector space. There are however various different ways of using this information, five of are compared in this paper.

In this section we explain three string-to-vector mappings which are variously described in the literature, and indeed are the only methods that we know of other than our own. To introduce a unifying terminology, we refer to these mechanisms by the following terms: *Pivoted Embedding - ℓ_∞*, *Pivoted Embedding - ℓ_2*, *LMDS - ℓ_2*, which we will continue to compare experimentally with two of our own invention, referred to here as *nSimplex - lwb* and *nSimplex - zen*.

In terms of the construction of a pivot set, it should be noted that in all cases other than LMDS, the selection of n pivot objects leads to the construction of an n-dimensional vector space; in the case of LMDS, the dimension of the constructed space equals the number of positive eigenvalues of the matrix of the squared distance between the pivots, which is at most $n - 1$.

We continue by giving explanations of each of these mechanisms; the first three in this section, the two nSimplex mappings in Section 3.

Pivoted Embedding with ℓ_∞. Pivoted Embedding is a metric space transformation based on the distances between data objects and a set of *pivots*. Given a metric space (U, d), and a set of pivots $\{p_1, \ldots, p_n\}$, the Pivoted Embedding f_P maps an object $o \in U$ to a n-dimensional vector $f_P(o) =$

[4] We appreciate this outline description is insufficient to accurately describe the function used in experiments, however it does succinctly give its essence. The interested reader is referred to the publicly available code base for this paper, which contains full source code for this metric: `bitbucket.org/richardconnor/it_db_conference_2019`.

$[d(o,p_1),\ldots,d(o,p_n)]$. Using the triangle inequality it can be proven that for any $o,q \in U$

$$\max_i |d(o,p_i) - d(q,p_i)| \leq d(o,q) \leq \min_i |d(o,p_i) + d(q,p_i)|. \quad (2)$$

The maximum distance at any dimension is captured by the Chevyshev[5] (ℓ_∞) distance, and therefore the distance $\ell_\infty(f_P(o), f_P(q))$ is guaranteed to be a lower-bound of $d(o,q)$. As the number of pivots increases, this lower-bound distance can reasonably be expected to provide a better approximation to the true distance, and this technique has been used to effect in some metric space contexts, e.g. [7, 19].

Pivoted Embedding with ℓ_2. Other distances over the same pivoted embedding spaces have also been used. For example, in the context of string spaces (i.e. $U = \mathbb{S}$), Spillmann et al [17] used the family of Minkowski metrics, which encompasses the well-known Euclidean distance (ℓ_2). It should be noted that there is no underlying theory for this technique, and in particular the ℓ_2 metric applied over the pivoted embedding space has no formal guarantees that we know of with respect to the string space it is intended to model.

The authors primarily investigate the selection of pivots to best effect within several different contexts, but make no particular justification for their choice of vector representation - except, perhaps, that it is the most straightforward possible mapping.

LMDS. In the general metric space context, perhaps the best known technique is *metric Multidimensional Scaling* (MDS) [12]. MDS aims to preserve *inter-point distances*. Given m objects and the distances $\delta_{i,j}$ between those points, it finds a set of m points $\{x_1,\ldots,x_m\}$ in a Euclidean space $\mathbb{R}^k$ such that $\ell_2(x_i,x_j)$ is close as possible to $\delta_{i,j}$. Those coordinates are computed using spectral analysis of the matrix of the squared interpoint distances.

For any metric which is isometrically Hilbert embeddable, it is possible to find a perfect ℓ_2 embedding for any n objects within $n-1$ Euclidean dimensions [11]. However, when the number m of data points is large the classical MDS is too expensive in practice due to a requirement for $O(m^2)$ distance computations and spectral decomposition of an $m \times m$ matrix.

The *Landmark MDS* (LMDS) [9] is a fast approximation of MDS. LMDS uses a set of k *landmark* points (i.e. pivots) to compute $k \times m$ distances of the data points from the landmark points. It applies classical MSD to the landmark points and uses a distance-based triangulation procedure to project the remaining data points. Specifically if $X_k \in \mathbb{R}^{k\times k}$ is the output of MDS on the pivot set, the embedding of a new point s into $\mathbb{R}^k$ is computed as $x_s = -\frac{1}{2}X_k^+(\delta_s^2 - \delta_\mu^2)$, where $(\delta_s^2)_i = d(s,p_i)$, $(\delta_\mu^2)_j = \frac{1}{n}\sum_{i=1}^n d(p_i,p_j)$ and X_k^+ is the pseudo inverse of X_k.

[5] The Chebyshev distance (ℓ_∞) is a true metric defined on a vector space to be the greatest difference along any coordinate dimension.

3 The n-Simplex Projection

The n-Simplex Projection projection is described in full in [6]; here we give a sketch of the approach. Although it requires a little mathematical background, the outcome is that there exists a relatively simple and efficient mapping from any space with the correct properties into an n-dimensional Euclidean space for an arbitrary choice of n. The mapping is contractive, and the associated error asymptotically approaches zero as n increases.

1. Any metric space (U, d) which is isometrically embeddable in a Hilbert space has the so-called *n-point property*. This means that, for any finite selection of n objects, an isometric embedding of just those objects exists in $(n-1)$-dimensional Euclidean space. That is, for any set $\{u_1, \ldots, u_n\} \subset U$ there exists a mapping $f : \{u_1, \ldots, u_n\} \rightarrow \mathbb{R}^{(n-1)}$ such that $d(u_i, u_j) = \ell_2(f(u_i), f(u_j))$ for any two objects u_i, u_j. This background work is attributed to mathematicians such as Blumenthal, Menger, Wilson etc. [3, 14, 18]
2. In practical terms, we have identified that metrics with this property include: Euclidean, Cosine, Jensen-Shannon, Triangular, and Quadratic Form distances [4]. In this context we also note that other proper metrics such as Chebyshev, Levenshtein, and Hamming do *not* possess the property. However, for any metric space (U, d) and any $n \geq 4$ there exist a constant $\alpha_n \leq 1/2$ such that for all the $0 < \alpha \leq \alpha_n$ the space (U, d^α) has the n-point property [10]. Specifically, Blumenthal proved that $\alpha_4 = 1/2$, and thus for any proper metric space (U, d), the space $(U, \sqrt{d})$ has the 4-point property. For $n > 4$, Deza and Maehara [10] proved that $\alpha_n \geq 0.72/n$, thus for any $0 < \alpha \leq 0.72/n$ the space (U, d^α) has the n-point property. Moreover, it is worth noting that for any finite metric space (U, d) of *negative type* [10] (e.g. hypermetric, ℓ_1-embeddedable metric, ultrametric) then $(U, \sqrt{d})$ has the n-point property.
3. In [6], we give the simple corollary of the n-point property that any n objects in the space can be used to form the vertices of a *simplex* in $(n-1)$-dimensional space, such that the edge lengths of the simplex are in one-to-one correspondence with the distances measured in the original space, and furthermore we give an algorithm to construct the Euclidean coordinates of its vertices. This therefore gives a concrete instantiation of the existential function f mentioned above.
4. This function is defined inductively, and at each step constructs a new apex in k dimensions based on a new object and a simplex previously formed in $(k-1)$dimensions, called *base simplex*, according to the distances between the new object and each object represented by the vertices of the base simplex.
5. Finally, we observe that once a base simplex is formed, this may be used to create any number of apexes based on it, using objects from the original space. For a choice of n reference objects, each of these apexes corresponds to a point in n-dimensional Euclidean space.

3.1 Dissimilarity functions over the n-Simplex projected points

For a fixed set of pivots, $\{p_1, \ldots, p_n\}$, let $f_S : U \to \mathbb{R}^n$ be the n-Simplex projection obtained using the pivots to form the vertices of the base simplex. We have shown [6] that for any two objects $o, q \in (U, d)$

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \leq d(o,q) \leq \sqrt{\sum_{i=1}^{n-1}(x_i - y_i)^2 + (x_i + y_i)^2}, \tag{3}$$

where $f_S(o) = [x_1, \ldots, x_n]$, $f_S(q) = [y_1, \ldots y_n]$. Thus the Euclidean distance $\ell_2(f_S(o), f_S(q))$ between any two apexes is a lower-bound of the distance between the corresponding objects within the original space, and an upper-bound could be computed as well. Furthermore, as more dimensions are chosen for the base simplex, these distance bounds asymptotically approach the true distance. The lower-bound (d'_{Lwb}) is a proper metric function on $\mathbb{R}^n$, while the upper-bound (d'_{Upb}) is a dissimilarity function but not a proper metric, because it does not satisfy the identity postulate (i.e. $d'_{Upb}(x, x)$ may not be equal to zero).

Other dissimilarity functions over the apex vectors may be considered as well, in particular, any function that is always between the lower-bound and the upper-bound will asymptotically approach the true distance when increasing the number of pivots. An example is the *mean* between the lower- and upper-bound, which as been proved to be particularly effective for similarity search task [1]. Here we introduce the *Zenith* dissimilarity function

$$d'_{Zen}(x,y) = \sqrt{\sum_{i=1}^{n-1}(x_i - y_i)^2 + x_i^2 + y_i^2}, \quad \forall\, x, y \in \mathbb{R}^n \tag{4}$$

which always satisfies $d'_{Lwb}(x,y) \leq d'_{Zen}(x,y) \leq d'_{Upb}(x,y)$, but has a geometrical interpretation stronger than the arithmetic mean between the upper- and the lower-bound. In fact, the Zenith distance, as well as the lower-bound and the upper-bound, can be interpreted as the Euclidean distance between two points in $\mathbb{R}^{n+1}$. Figure 2 shows the case $n = 2$.

In general, given the pivots $\{p_1, \ldots, p_n\}$ and two objects o, q we know that there exists an *isometric* embedding of those $n + 2$ points into $\mathbb{R}^{n+1}$.

Let $v_o, v_q, v_{p_1}, \ldots, v_{p_n}$ be the mapped points, i.e. vectors such that $d(o,q) = \ell_2(v_o, v_q)$, $d(o,p_i) = \ell_2(v_o, v_{p_i})$, $d(q,p_i) = \ell_2(v_q, v_{p_i})$, $d(p_i,p_j) = \ell_2(v_{p_1}, v_{p_j})$, for all $i, j = 1 \ldots, n$. The points $v_{p_1}, \ldots, v_{p_n}$ are the vertices of the so-called base simplex, which lies in a $(n-1)$dimensional subspace. The points $v_{p_1}, \ldots, v_{p_n}$ and v_o lies in a hyperplane of $\mathbb{R}^{n+1}$, which we refer to as H. By rotating v_q around the base simplex (i.e. rotate the point while preserving the distances to the pivots) we have two possible projections onto the hyperplane H (one above and one below the base simplex). Let v_q^+ and v_q^- be those points, which coincide with the intersections of hyperplane H and the n balls of centre p_i and radius $d(p_i, q)$. In general, we don't explicitly know the distance $d(o,q)$, so we don't know the true altitude of the vertex v_q over the hyperplane H. The vertices

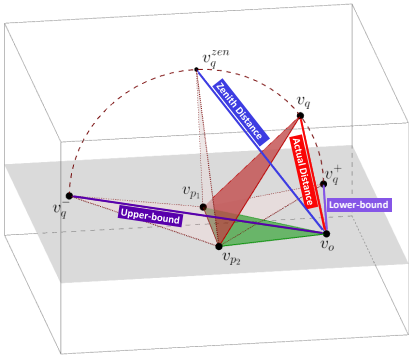


Fig. 1: Example of n-Simplex projection for $n = 2$, where the two pivots p_1, p_2 and two data objects o, q are first isometrically embedded in 3D Euclidean space and then projected in 2D Euclidean plane. The mapped points are indicated with $v_{p_1}, v_{p_2}, v_o, v_q$ in the graph. By rotating v_q around the edge $\overline{v_{p_1} v_{p_2}}$ until it is coplanar with v_{p_1}, v_{p_2}, v_o, we obtain two possible apices (v_q^+ and v_q^-), that are the intersections of the hyperplane containing v_{p_1}, v_{p_2}, v_o, and the the two balls centred in v_{p_1}, v_{p_2} and radii $d(q, p_1)$, $d(q, p_2)$, respectively. The apex v_q^{zen} is that at the highest altitude over the hyperplane that still preserves the distance to the two pivots.

v_q^+ and v_q^- are those obtained by approximating v_q with points at zero altitude such that the distances to the pivots are preserved. Another possible choice is considering the *zenith* point v_q^{zen} which is the point at the highest altitude over H that still preserves the distances to the pivots. In [6] we show that the quantities $\ell_2(v_o, v_q^+)$ and $\ell_2(v_o, v_q^-)$ provide a lower-bound and an upper-bound of $d(o, q)$. By construction, the quantity $\ell_2(v_o, v_q^{zen})$ will always been between those bounds, and it corresponds to the *quadratic mean* of the lower- and upper-bounds, i.e. $d'_{Zen} = \sqrt{\frac{(d'_{Lwb})^2 + (d'_{Upb})^2}{2}}$.

3.2 Application of n-Simplex to Levenshtein

As previously mentioned, with respect to our techniques there is one essential difference between the two metrics we are testing: Jensen-Shannon distance is Hilbert-embeddable, and therefore has the n-point property; Levenshtein does not. This means that, in the form defined above, it cannot be used to build a simplex in n-dimensional space for some arbitrary n.

However, as observed above, for any $n > 4$ there exists some exponent α_n with $0.72/n \leq \alpha_n \leq 0.5$, such that for any $\alpha \in (0, \alpha_n]$ the space $(\mathbb{S}, d_{lev}^{\alpha})$ satisfies the n-point property [10]. Therefore, for a fixed α in the desired range we can use the n-simplex projection to transform the space $(\mathbb{S}, d_{lev}^{\alpha})$ into $(\mathbb{R}^n, d')$, where d' is a dissimilarity function over the projected points (e.g., the n-simplex lower-bound or the zenith measure). If $v_i, v_j \in \mathbb{R}^n$ are the projections of the strings $s_i, s_j \in \mathbb{S}$,

then we can approximate the actual distance $d_{lev}(s_i, s_j)$ with $(d'(v_i, v_j))^{1/\alpha}$. It is worth noting that using the value $\tilde{\alpha}_n = 0.72/n$ guarantees that $(U, d^{\tilde{\alpha}_n})$ has the n-point property for *any* metric space (U, d); however for a specified metric space it is possible to find a suitable exponent that is greater than $\tilde{\alpha}_n$. We experimentally observed that, in our context, $\alpha = 0.5$ works well for $(\mathbb{S}, d_{lev})$; in fact we have found no counter-examples where it is impossible to embed any n objects from our domain into an $(n-1)$-dimensional Euclidean space. Thus we applied the n-Simplex projection on $(\mathbb{S}, \sqrt{d_{lev}})$. Note we do not claim this property for $\mathbb{S}$ in general, although we are interested in pursuing this further as we suspect it to be the case.

Given the above, when comparing the n-Simplex techniques in the context of the Levenshtein distance, we first cast the space $(\mathbb{S}, d_{lev})$ into $(\mathbb{S}, \sqrt{d_{lev}})$ to create the vector space, and then square the results before comparison. In terms of intrinsic dimensionality, or discriminability, this is very undesirable; the application of this metric-preserving transformation decreases the variance of measured distances, which makes a proximity query more unstable as the the square-rooted distance has less power in discriminating between nearest and furthest neighbour [2]. Thus, the techniques which do not require to perform this mapping should benefit considerably in terms of relative error. This effect shows in some of our results, as mentioned later.

4 Experimental Analysis

To test the five techniques outlined above, we use a single collection of strings drawn from a public English lexicon of 69,069 words[6] and consider Levenshtein distance (d_{Lev}) and Jensen-Shannon distance (d_{JS}) over this collection. For varying numbers of pivots, we apply the five different methods and measure distortion, relative error, and absolute squared error.

For each measured number n of pivots, we repeatedly select a subset random subset of n strings to act as pivots, and a further random selection of 10,000 strings to act as the finite data set. We consider all $\binom{n}{2}$ pairwise distances within the set, i.e. nearly 50 million distances, and for each pair measure both the original distance and the mapped distance to calculate the properties listed. We repeat this process until the standard error of the mean for each measurement is within an acceptable limit, and then report the mean values obtained.

Notationally, we refer to an original metric space $(\mathbb{S}, d)$ which is mapped into a new space $(\bar{\mathbb{S}}, \bar{d})$, where $s_i, s_j \in \mathbb{S}$ map to $s'_i, s'_j \in \bar{\mathbb{S}}$ and so on. We measure the following properties of the mapped data:

Distortion For each mapping, the distortion is the minimum $R \geq 1$ such that there exists a multiplying factor $r > 0$ such that, for all $s_i, s_j \in \mathbb{S}$,

$$r \cdot \bar{d}(s'_i, s'_j) \leq d(s_i, s_j) \leq rR \cdot \bar{d}(s'_i, s'_j).$$

[6] available from `www.sisap.org`

Note that distortion gives a "worst case" analysis and is driven by outliers within the mapping rather than typical errors; the larger the sample of distances, the larger the distortion is likely to be.

Average Relative Error For each pair of distances $d(s_i, s_j)$ and $\bar{d}(s'_i, s'_j)$ we calculate the ratio, and use the mean of these ratios as a scaling factor γ. Then the average relative error is calculated as

$$\frac{1}{\binom{|\mathbb{S}|}{2}} \sum_{i,j} \frac{|d(s_i, s_j) - \gamma \cdot \bar{d}(s'_i, s'_j)|}{d(s_i, s_j)}$$

Mean Squared Error For each pair of distances $d(s_i, s_j)$ and $\bar{d}(s'_i, s'_j)$ we calculate the ratio, and use the mean of these ratios as a scaling factor γ. Then the mean squared error is calculated as

$$\frac{1}{\binom{|\mathbb{S}|}{2}} \sum_{i,j} (d(s_i, s_j) - \gamma \cdot \bar{d}(s'_i, s'_j))^2$$

4.1 Results

The results of these experiments are plotted in Figure 2.

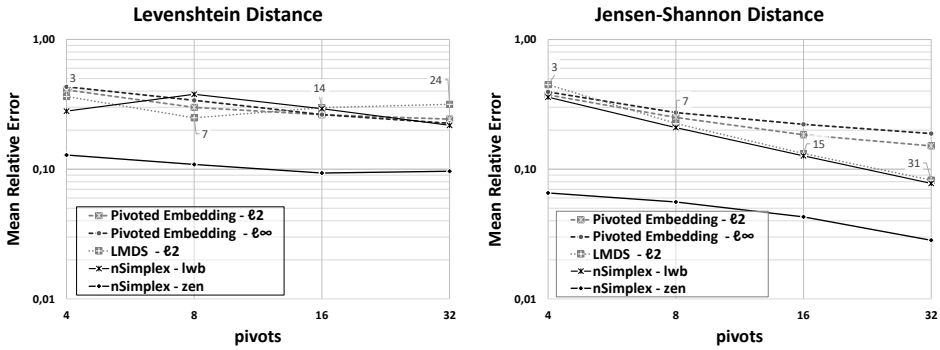
The clearest observation is that, for almost all measures, the nSimplex-Zenith projection is the best. It is clearly by far the best estimator for all measures over Jensen-Shannon distance, and also for Levenshtein in all aspects other than distortion. Note the scale of the improvement is quite dramatic, the Y axis being given in log scale for all charts. It is also notable how little the performance degrades when less pivots are used compared with the other projections.

The n-Simplex lower-bound estimator also performs well, and our conclusion is that, at least for the Jensen-Shannon space, the n-Simplex projection is clearly preferable to any of the other established methods.

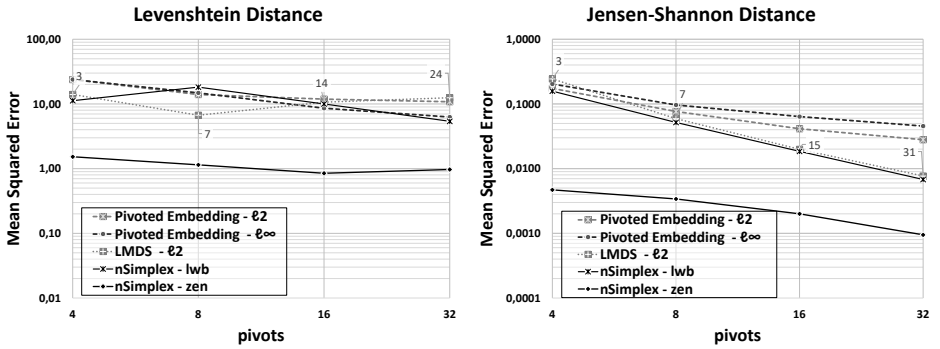
The difference between the two metrics is, we believe, due to the presence of the Hilbert embedding property of the Jensen-Shannon metric, and thus we expect this behaviour to extend to other such metrics. The problem with Levenshtein, which again we would expect to extend to other non-embeddable metrics, is that in order to apply the projection it is necessary to scale the original distance into a much less discriminable form; this we believe in particular is the reason for the relatively poor figures for distortion. This effect deserves further attention to give a better understanding.

5 Conclusions

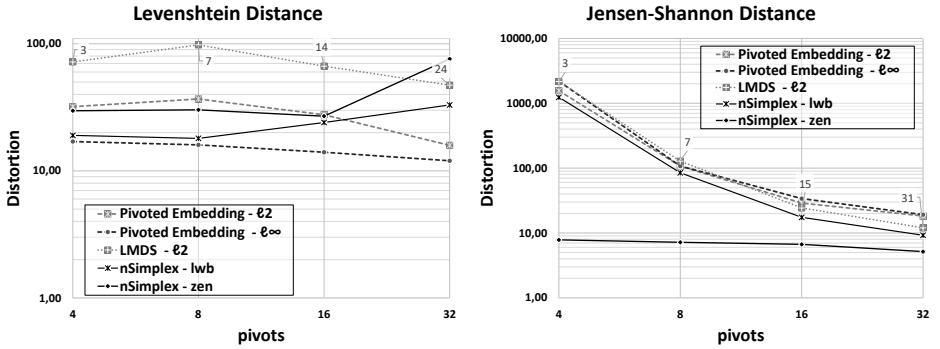
We have described the use of the n-Simplex projection to create mappings from strings into vector spaces, and compared properties of these mappings with three others. This is the first time n-Simplex has been demonstrated over string functions, and in particular we have introduced for the first time the Zenith estimator function. This has been shown to be significantly better in terms of the accuracy of estimations in the mapped space than any previously published mapping.



(a) `Mean Relative Error`



(b) `Mean Squared Error`



(c) `Distortion`

Fig. 2: Results from varying the number n of pivots. Left figures give measurements for Levenshtein Distance, the right for Jensen-Shannon Distance. We recall that both Pivoted Embedding and n-Simplex projection provide a mapping into n dimensions. For the LMDS approach, the dimensionality of the projected space equals the number of positive eigenvalues of the matrix of the squared distance between the pivots. These are indicated by the numbers along the LMDS lines in the graphs.

Acknowledgements

Lucia Vadicamo acknowledges financial support from the VISECH ARCO-CNR project, CUP B56J17001330004, and the Italian National Research Council (CNR) for a Short-Term Mobility Grant (STM) at the University of Stirling.

References

1. Amato, G., Chávez, E., Connor, R., Falchi, F., Gennaro, C., Vadicamo, L.: Re-ranking permutation-based candidate sets with the n-Simplex projection. In: Proceedings of SISAP 2018. pp. 3–17. Springer (2018)
2. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is nearest neighbor meaningful? In: Proceedings of ICDT 1999. pp. 217–235. Springer (1999)
3. Blumenthal, L.M.: Theory and applications of distance geometry. Clarendon Press (1953)
4. Connor, R., Cardillo, F.A., Vadicamo, L., Rabitti, F.: Hilbert exclusion: Improved metric search through finite isometric embeddings. ACM T Inform. Syst. **35**(3), 17:1–17:27 (Dec 2016)
5. Connor, R., Vadicamo, L., Cardillo, F.A., Rabitti, F.: Supermetric search. Inform. Syst. **80**, 108 – 123 (2019)
6. Connor, R., Vadicamo, L., Rabitti, F.: High-dimensional simplexes for supermetric search. In: Proceedings of SISAP 2017. pp. 96–109. Springer (2017)
7. Connor, R.C.H., MacKenzie-Leigh, S., Moss, R.: High dimensional search using polyhedral query. In: Proceedings of SISAP 2014. pp. 176–188. Springer (2014)
8. Connor, R.C.H., Simeoni, F., Iakovos, M., Moss, R.: A bounded distance metric for comparing tree structure. Inform. Syst. **36**(4), 748–764 (2011)
9. deSilva, V., Tenenbaum, J.B.: Global versus local methods in nonlinear dimensionality reduction. In: Proceedings NIPS. vol. 15, p. 721728 (2003)
10. Deza, M., Maehara, H.: Metric transforms and euclidean embeddings. T Am. Math. Soc. **317**(2), 661–671 (1990)
11. Dokmanic, I., Parhizkar, R., Ranieri, J., Vetterli, M.: Euclidean distance matrices: Essential theory, algorithms, and applications. IEEE Signal Proc. Mag. **32**(6), 12–30 (Nov 2015)
12. Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**(1), 1–27 (Mar 1964)
13. Manber, U.: Finding similar files in a large file system. In: USENIX Winter 1994 Technical Conference. pp. 1–10 (1994)
14. Menger, K.: New Foundation of Euclidean Geometry. Amer. J. Math. **53**(4), 721–745 (1931)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS 2013, pp. 3111–3119. Curran Associates, Inc. (2013)
16. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of EMNLP 2014. pp. 1532–1543 (2014)
17. Spillmann, B., Neuhaus, M., Bunke, H., Pekalska, E., Duin, R.: Transforming strings to vector spaces using prototype selection. In: Structural, Syntactic, and Statistical Pattern Recognition. pp. 287–296. Springer (09 2006)
18. Wilson, W.A.: A Relation Between Metric and Euclidean Spaces. Amer. J. Math. **54**(3), 505–517 (1932)
19. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity search: the metric space approach, vol. 32. Springer Science & Business Media (2006)