String Distances

D. Luzeaux

ETCA/CREA/SP, 16 av. Prieur de la Côte d'Or
F-94114 ARCUEIL, France
E-mail: luzeaux@etca.fr

Abstract

In this paper we make first a short survey of usual distances between strings and state some results on the induced topologies; then we introduce a new family of distances generated by arbitrary sequences and give finally a completeness result.

1 Introduction

We will be dealing in the next sections with words written on a finite alphabet and will study metrics between words. This is an interesting theoretical problem and has furthermore many applications in computer science and related fields: pattern recognition [Fu82], control strategy problem of interconnection networks [BT91], automatic learning of rule-based controllers [Luz91], fractal geometry [Edg90, BT92].

In order to avoid confusion, let us recall our main notations: let T be a finite set (the *alphabet*) whose elements are called *letters* and let us define an operation $\bullet$ (the *concatenation*) which, given two letters a and b, builds the *word* (or *string*) $a \bullet b$. The set of such words is called T^2. We define recursively in the same way the sets T^n for each integer n; let us call λ (other classical notations are 1 or ϵ) the neutral element for concatenation (the *empty word*) and T^* the union of all T^n: $T^* = \{\lambda\} \cup T \cup T^2 \cup \cdots \cup T^n \cup \cdots$. T^* is the set of all words that may be written on the alphabet T with the concatenation operation. Usually T has at least two elements, and we will always consider that is the case. $(T^*, \bullet)$ has a monoid structure and is therefore called the *free monoid*. In the next sections, we will write ab instead of $a \bullet b$ for sake of simplicity. If E is a set of words, we define the sets $a \bullet E = \{aw \mid w \in E\}$ and $E \bullet a = \{wa \mid w \in E\}$. Next we define T^ω as the set of countably infinite strings from the alphabet T.

Instead of considering T^* in its whole, it is possible to consider only subsets (*languages*), some of which are easily characterized: for instance, if the subset is recursively enumerable, *grammars* are an appropriate tool. This leads to very interesting problems [Pin88] (how to characterize topologically sets described by given grammars), but we will keep here a general overview and will only consider T^*. Actually we will be interested in defining metrics on T^* which will be called *string distances*. We will briefly introduce some metrics on T^ω, which can be seen as an extension of T^*, and should thus not be ignored.

In the whole paper, we will deal with metric spaces; thus being compact, sequentially compact and countably compact are equivalent properties. We will use any of these properties to claim compacity.

2 Survey of some familiar string distances

2.1 Levenstein distance

The *Levenstein distance* shows a local similarity between two words; it is the minimum number of three operations that transform one of the two words into the other. Let us define more precisely these three applications from T^* into $\mathcal{P}(T^*)$, mapping a string into a set of strings: they are respectively destruction (des), insertion (ins) and substitution ($subs$) defined by:

$\forall a \in T,\ des(a) = \{\lambda\}$
$\forall u \in T^* \setminus \{\lambda\}, \forall v \in T^* \setminus \{\lambda\},\ des(uv) = des(u) \bullet v \cup u \bullet des(v)$
$\forall a \in T,\ ins(a) = T \bullet a \cup a \bullet T$
$\forall u \in T^* \setminus \{\lambda\}, \forall v \in T^* \setminus \{\lambda\},\ ins(uv) = ins(u) \bullet v \cup u \bullet ins(v)$
$\forall a \in T,\ subs(a) = (T \setminus \{a\})$
$\forall u \in T^* \setminus \{\lambda\}, \forall v \in T^* \setminus \{\lambda\},\ subs(uv) = subs(u) \bullet v \cup u \bullet subs(v)$

Let s be a string. Starting from the three previous operations, we may build recursively the sets $(C_n(s))_{n \in \mathbf{N}}$:

$$C_0(s) = \{s\}$$

$$\forall n \geq 1, C_n(s) = des(C_{n-1}(s)) \cup ins(C_{n-1}(s)) \cup subs(C_{n-1}(s))$$

These sets are the strings that may be created, starting from a given string and applying several times any of the three operations. The *Levenstein distance* between the strings s_1 and s_2 is the integer defined by:

$$d(s_1, s_2) = \inf\{n \in \mathbb{N} |\ s_2 \in C_n(s_1)\}$$

It is obvious that $d : T^* \times T^* \to \mathbb{R}_+$ defines a distance. d has in fact only integer values. Let us consider the topology on T^* induced by this distance. For sake of simplicity we will identify the word w and the singleton $\{w\}$.

Each word appears then as an open set (a word is the center of an open ball of radius 1/2); it is a closed set too. **The induced topology is the discrete topology.**

Every bounded set has a finite number of elements and conversely. The sets $C_n(s)$ are the closed balls of radius n and center s.

Let $(u_n)_{n \in \mathbb{N}}$ be a sequence of words; as d takes its values on $\mathbb{N}$, every converging sequence is stationary.

If any finite or infinite set of words is called a *language*, it follows that every language is closed (any limit of a sequence is part of the sequence). With this distance and the induced topology, T^* **is complete, but neither compact nor connex** (Proof: as d takes integer values, any Cauchy sequence is stationary, therefore converging; as T^* is infinite, it is not compact for the discrete topology; the non-connexity is of course a consequence of the discrete topology).

Let us consider a string s and a language G; **there is a finite number of strings in G minimizing the distance between s and G** (Proof: take any element t of G and consider the ball with center s and radius $d(s,t)$; this ball is $C_{d(s,t)}(s)$ and has a finite number of elements; the intersection between this ball and G is finite; an element of G minimizing the distance between s and G is obtained by taking the minimum of all distances between s and the finite points of the previous intersection).

2.2 Discrete distances

Let u be a word and $u = u_1 u_2 u_3$ a factorization of this word; u_2 is a *factor*, u_1 is a *prefix* and u_3 is a *suffix*. A word v is a *subword* of u if it may be obtained from u by deleting letters of u. The length of u is written $|u|$. Let us call $L_1(u,v)$ (respectively $L_2(u,v)$) the greatest length of a factor (respectively subword) belonging to u and v.

Two distances can then be defined: $d_i(u,v) = |u| + |v| - 2L_i(u,v)$ with $i = 1, 2$.

d_1 and d_2 are distances called respectively *factor distance* and *subword distance*. They are part of a more general family: the *discrete distances*. These may be defined [Cho87] on a free monoid: they take their values in $\mathbb{N}$ and verify: $d(u, \lambda) = |u|$.

The following inequalities are then true: (d is any discrete distance)

- $|u| \leq d(u,v) + |v|$
- $d(u,v) \leq |u| + |v|$
- $d_2(u,v) \leq d_1(u,v)$
- $d_2(xy, uv) \leq d_2(x,u) + d_2(y,v)$
- $d_2(xu, xv) = d_2(ux, vx) = d_2(u,v)$
- $d_2(xuy, zvt) - |xyzt| \leq d_2(u,v) \leq d_2(xuy, zvt) + |xyzt|$

It is obvious that **the induced topology is the discrete topology**.

2.3 An ultrametric distance

Instead of defining distances on $\mathbb{N}$, it may be more interesting to define them on the real set.

Let u and v be two words: $u = u_1 \cdots u_{|u|}$ and $v = v_1 \cdots v_{|v|}$. Let us call $k_{(u,v)}$ the lower bound of the set $\{k|\ u_k \neq v_k\}$. We may define the application:

$$\begin{array}{rlcl} d: & T^* \times T^* & \longrightarrow & \mathbb{R} \\ & (u,v) & \longmapsto & \frac{1}{k_{(u,v)}} \text{if } u \neq v \\ & (u,u) & \longmapsto & 0 \end{array}$$

d is a distance (the triangular inequality follows from: $k(u,w) \geq \min(k(u,v), k(v,w))$ which implies $d(u,w) \leq \max(d(u,v), d(v,w))$).

(T^*, d) **is an ultrametric space**, as can be seen with the last previous inequality. The induced topology is the discrete topology (for a given $u \in T^*$, we have: $\forall v, u \neq v, d(u,v) \geq (|u|+1)^{-1}$).

2.4 Metrics on T^ω

In this section, we will define string distances on the set of infinite strings (for further properties of T^ω see for instance [BN80]). Let s be an element of T^ω and $n \in \mathbb{N}$; we write $s[n]$ for the initial segment of s of length n ($s[0]$ is defined as λ). For any α in T^*, we define $[\alpha]$ as the set of all infinite strings that begin with α:

$$[\alpha] = \{s \in T^\omega \mid \exists n \in \mathbb{N}, \alpha = s[n]\}$$

Let r be a real number satisfying $0 < r < 1$. Let u and v be two infinite strings and $u = \alpha u'$, $v = \alpha v'$ where $\alpha \in T^*$ and $u'[1] \neq v'[1]$. In other words, α is the longest common prefix; let k be the length of α.

Let us define ρ_r by: $\rho_r(u, v) = r^k$.

ρ_r is **a distance** on T^ω (Proof of the triangular inequality: if u, v and w are elements of T^ω, let k be the length of the longest common prefix of u and v, and let m be the length of the longest common prefix of v and w. If $n = \min(k, m)$, it is obvious that at least the first n letters of u agree with the first letters of w. Therefore $\rho_r(u, w) \leq r^{\min(k,m)} = \max(r^k, r^m) = \max(\rho_r(u, v), \rho_r(v, w))$).

The last inequality shows that (T^ω, ρ_r) **is an ultrametric space.**

With this distance, T^ω is **complete** (Proof: let (s_n) be a Cauchy sequence; for each k, there is $n_k \in \mathbb{N}$ such that: $\forall n, m \geq n_k$, $\rho_r(s_n, s_m) < r^k$. This means that $s_{n_k}[k] = s_m[k]$ for all $m \geq n_k$. Let s be defined as: the k-th letter of s is the k-th letter of s_{n_k}. Then (s_n) converges to s: for all $\epsilon > 0$, choose k with $r^k < \epsilon$. Then for $m > n_k$, we have $s_m[k] = s[k]$, thus $\rho_r(s_m, s) \leq r^k < \epsilon$).

(T^ω, ρ_r) **is separable**, as the set $\{[\alpha] | \alpha \in T^*\}$ is a countable base for the open sets (actually, every open ball is some $[\alpha]$).

(T^ω, ρ_r) **is compact** (Proof: let A be an infinite set of T^ω; then at least one of the sets $A \cap [a_1]$, $\cdots$, $A \cap [a_p]$ is infinite, where $a_1, \cdots, a_p$ are the letters of the alphabet T. Assume it is $A \cap a_1$; then at least one of the sets $A \cap [a_1 a_1]$, $\cdots$, $A \cap [a_1 a_p]$ is infinite. Going on with this reasoning, we build at each step i a string growing in length, with one additional element at each step, so at step i, this string is some $s[i]$ where $s \in T^\omega$. It is obvious that when i grows, the corresponding string converges to s. For any $\epsilon > 0$, there exists n such that the ball of center s and radius ϵ contains $[s[n]]$, and by construction, we know that $A \cap [s[n]]$ is infinite. Thus s is a limit point of A, and (T^ω, ρ_r) is countably compact).

As we have seen, any value of r yields the same topology; however, depending on the applications, some values of r may be preferred (computing fractal dimensions for instance in [Edg90].

3 Distances generated by positive sequences

These distances will take their values on a countable subset of $\mathbb{R}$ and induce different metrical structures depending on their support; they are generated by positive sequences which introduce a cost on the differences between two strings; it is then possible to obtain distances such that two strings are arbitrary near if they have almost similar prefixes, or on the contrary similar suffixes.

Let us call $\delta_{x,y}$ the map which takes 1 as a value if $x \neq y$ and 0 in the other case. Let us consider a sequence $(a_i)_{i \in \mathbb{N}}$ of strictly positive reals. u_i is the i-th letter of u. We define:

$$d_{[a_i]} : \begin{array}{ccc} T^* \times T^* & \longrightarrow & \mathbb{R} \\ (u, v) & \longmapsto & \sum_i a_i \delta_{u_i, v_i} \end{array}$$

$d_{[a_i]}$ **is a distance** (Proof: the triangular inequality follows from the fact that δ verifies it and $(a_i)_{i \in \mathbb{N}}$ has positive terms). T^* with the topology induced by $d_{[a_i]}$ has different structures depending on the generating sequence.

The induced topology is the discrete topology (Proof: let u be a word with length $|u|$. For any v different from u, we have: $d(u, v) \geq \min_{1 \leq i \leq |u|+1}(a_i)$, because v has either a greater length and differs then from u at least at the rank $|u| + 1$, or v is shorter and differs at a rank between 1 and $|u|$).

The main interest of this family of distances is the next result on the **completeness** of $(T^*, d_{[a_i]})$.

Proposition: $(T^*, d_{[a_i]})$ *complete* $\iff (\sum_{i=1}^n a_i)_{n\in\mathbb{N}}$ *is not a Cauchy sequence.*

Proof: let us call $S_n = \sum_{i=1}^n a_i$

$\bullet(\sum_{i=1}^n a_i)_{n\in\mathbb{N}}$ Cauchy sequence $\Longrightarrow (T^*, d_{[a_i]})$ non complete.

Let $(u^{(n)})_{n\in\mathbb{N}}$ be the sequence with general term $u^{(n)} = b^n$, where b is a letter of the alphabet T. Then, $d(u^{(n+p)}, u^{(n)}) = S_{n+p} - S_n$, so $(u^{(n)})_{n\in\mathbb{N}}$ is a Cauchy sequence and does not however converge in $(T^*, d_{[a_i]})$.

$\bullet(\sum_{i=1}^n a_i)_{n\in\mathbb{N}}$ is not a Cauchy sequence $\Longrightarrow (T^*, d_{[a_i]})$ complete.

Let $(u^{(n)})_{n\in\mathbb{N}}$ be a sequence.

If for every n, $|u_n|$ is bounded by a real L, we have: $\forall p, d(u_{n+p}, u_n) \geq \min_{i\leq|L|} a_i$ and the sequence (u_n) cannot be a Cauchy sequence if it is not stationary.

If the sequence of lengths of the u_n is not bounded, it is not possible to extract from the sequence (u_n) a subsequence (v_n) such that: $\forall n, |v_{n+1}| > |v_n|$. Let N be an integer; for any p, we have $d(v_{N+p}, v_N) = S_{N+p} - S_N$, therefore if (S_n) is not a Cauchy sequence, (v_n) cannot be one either, and neither can be (u_n). It follows that if (S_n) is not a Cauchy sequence, the only Cauchy sequences of elements of T^* are stationary sequences, which obviously converge.

4 Concluding remarks

We have seen some distances on finite and infinite strings, some of which are useful in pattern recognition applications. We have introduced a family of distances which is very interesting as the metric induced structure has different properties depending on the generating sequence. Furthermore the sequence introduces the notion of arbitrary weights between letters which is practical when studying robustness for instance in computer applications: in pattern recognition, objects may be represented by strings and geometrical transformations on objects are translated into operations between strings; the sequence (a_i) can be seen as the cost of these transformations. These considerations lead to robust grammatical inference [Fu82, Luz91]. Further work on string distances can provide a useful tool to describe fractal curves with grammars [BT92].

References

[BN80] L. Boasson and M. Nivat. Adherences of languages. *Journal of computer and system sciences*, 20, 1980.

[BT91] J. Blanc-Talon. *Approche structurelle du contrôle du réseau d'interconnexion d'une machine multi-processeurs pour les algorithmes de reconnaissance des formes, et applications*. PhD thesis, Paris XI, 1991.

[BT92] J. Blanc-Talon. How can grammars generate fractal curves? Technical report, CSIRO-CSIS-DIT, Canberra ACT, Australia, 1992.

[Cho87] C. Choffrut. Fonctions rationnelles et distances discrètes. Technical report, LITP, 1987.

[Edg90] G.A. Edgar. *Measure, topology and fractal geometry*. UTM, Springer Verlag, 1990.

[Fu82] K.S. Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, 1982.

[Luz91] D. Luzeaux. *Pour l'apprentissage, une nouvelle approche du contrôle: théorie et pratique*. PhD thesis, Paris XI, 1991.

[Pin88] J.E. Pin. Topologies for the free monoid. Technical Report 17, LITP, 1988.