International Journal of Foundations of Computer Science
© World Scientific Publishing Company

A SIMILARITY MEASURE FOR CYCLIC UNARY REGULAR LANGUAGES

JÜRGEN DASSOW
Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg
PSF 4120; D-39016 Magdeburg; Germany
dassow@iws.cs.uni-magdeburg.de

GEMA M. MARTÍN and FRANCISCO J. VICO
Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga
Severo Ochoa, 4, Parque Tecnológico de Andalucía,
E-29590 Campanillas - Málaga, Spain
gema,fjv@geb.uma.es

Received (Day Month Year)
Accepted (Day Month Year)
Communicated by (xxxxxxxxxx)

A cyclic unary regular language is a regular language over a unary alphabet that is represented by a cyclic automaton. We propose a similarity measure for cyclic unary regular languages by modifying the Jaccard similarity coefficient and the Sørensen coefficient to measure the level of overlap between such languages. This measure computes the proportion of strings that are shared by two or more cyclic unary regular languages and is an upper bound of the Jaccard coefficient and the Sørensen coefficient. By using such similarity measure, we define a dissimilarity measure for cyclic unary regular languages that is a semimetric distance. Moreover, it can be used for the non-cyclic case.

Keywords: Similarity measure; cyclic unary regular language; Jaccard coefficient; Sørensen coefficient.

1. Introduction

Unary regular languages (for short, URLs) are regular languages over a unary alphabet. Due to their relation to many number-theoretic results, as well as their difference from the general case (non-unary regular languages), they are of particular interest in the study of state complexity [11, 12], i.e., which is defined as the size of the minimal finite automaton accepting the language. Some papers on state complexity of URLs have been published. For example, in [7], deterministic unary automata, nondeterministic unary automata and probabilistic unary automata accepting the same languages are compared with respect to their size, and in [10] the nondeterministic state complexity of URLs and of their complements are also compared. A cyclic URL (for short, CURLs) is a URL that can be represented by

a cyclic automaton. Some properties of CURLs and unary nondeterministic finite automata have been investigated in [5, 8].

However, there is a lack of results that compare two (neither cyclic nor non-cyclic) URLs based on their strings. As far as we know, there does not exist a measure of the overlap between two URLs or between two CURLs, as it is the case with other types of languages. For example, in [2], an iterative procedure to compute the relative entropy between two stochastic deterministic regular grammars is proposed and in [3], a general approach to compute a similarity measure between distributions generated by probabilistic tree automata is defined.

On the other hand, if one considers dynamic systems or genetic algorithms, where the populations are presented by unary regular sets (see e.g. [4]), then the selection process requires a comparison of such sets. Thus we are interested in a similarity measure for unary regular sets.

In the case of finite sets A and B the Jaccard coefficient and the Sørensen (or Dice) coefficient defined by

$$JC_{A,B} = \frac{|A \cap B|}{|A \cup B|} \text{ and } SC_{A,B} = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{1}$$

are well-known measures of similarity (see [15], [13], [16]). Obviously, the intuitive idea behind these measures is that sets are more similar if they have more elements in common. These measures cannot directly be used for infinite sets. Since we are interested in infinite regular sets, the Jaccard and Sørensen coefficients cannot directly be used for CURLs.

In this paper, we introduced modified variants of these coefficients. But their computations cannot directly be performed using a given representation of the CURLs by their minimal automata; it needs a transformation to another representation of the CURLs.

Thus, in this work, we propose a similarity measure for CURLs that computes the overlap between two or more CURLs directly from the given representations by minimal automata. Moreover, we prove that the similarity measure for CURLs proposed in this work is an upper bound of the Jaccard coefficient and the Sørensen coefficient for CURLs. Furthermore, if a sequence of CURLs approaches a certain CURL with respect to one of the considered similarities, then this also holds for the other ones. Thus by the relation between the measures it seems that a tendency can be seen earlier by using the newly introduced measure.

Using the similarity measure, we also define a dissimilarity measure for CURLs. That will be done in the same way as the Jaccard distance is defined by using the Jaccard coefficient (in the case of finite sets). In contrast to the Jaccard distance, such a dissimilarity measure for CURLs is not a metric distance, since the triangle inequality is not satisfied. We prove that it is a semimetric distance.

Finally, we mention that we can also use the dissimilarity measure proposed in this work in the case of non-cyclic URLs. Therefore, in general, we have a dissimilarity measure for URLs (cyclic and non-cyclic). We show that the dissimilarity

measure for URLs is a symmetric distance (it does not satisfy the identity of indiscernibles) and not a semimetric distance.

This paper is organized as follows. In Section 2, we present some notations that are used in the sequel. In Section 3, we define a measure of similarity for CURLs by modifying the Jaccard similarity to infinite sets. In Section 4, we propose a definition of the Jaccard coefficient and the Sørensen coefficient, which can be used for CURLs. and compare them with the similarity measure. In Section 5, we define a measure of dissimilarity for CURLs by using the similarity measure and present some properties of it. We finish with a short discussion concerning URLs.

2. Some Notation

The reader is assumed to be familiar with the basic concepts of formal language theory. Here, only some notations used in the paper will be recalled. For further information the reader is referred to [14].

In the sequel, we will consider that $0 \in \mathbb{N}$. For the cases in which zero is not included, we will write $\mathbb{N}^+$. A periodic sequence of numbers with period y, i.e., x, $x+y$, $x+2y$, $x+3y, \ldots$, will also be called a (natural) succession. The cardinality of a finite set X is designated by $|X|$. For a word $x = x_1x_2 \ldots x_n$ over some alphabet V, $x_i \in V$ for $1 \leq i \leq n$, we denote by $x(i)$ the i-th letter of x, i.e., $x(i) = x_i$.

In this paper we work with languages over a unary alphabet. Let A be a deterministic finite automaton over a unary alphabet (for short, UDFA) that represents a regular language. As the alphabet is unary, each UDFA will have the diagram that is shown in Figure 1. Its states are divided into two groups, the first one, that we call initial part, contains the states from the initial state 0 to the state $i-1$, the second one, that we call loop, contains the remaining states. The initial word can be empty in those automata, whose last state transits to its initial state.

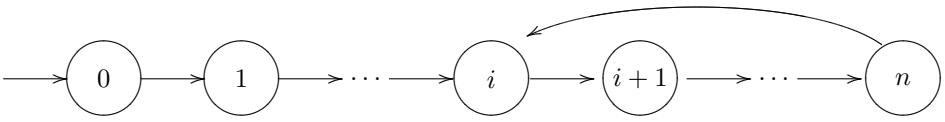

Fig. 1. Diagram of a UDFA

A UDFA will be represented as a vector (v, w) where $v \in \{0,1\}^*$ describes the initial part and $w \in \{0,1\}^+$ describes the loop. The zeros represent the non-accepting states of the automaton, and the ones represent the accepting states of the automaton.

For example, the representation of the automaton in Figure 2 is $(011, 110)$.

For a UDFA (v, w), where the states are numbered starting from zero, let

$$A = \{a \mid 1 \leq a \leq |v|,\ v(a) = 1\} \quad \text{and} \quad B = \{b + |v| \mid 1 \leq b \leq |w|,\ w(b) = 1\}.$$

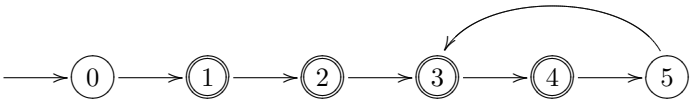

Fig. 2. An example of a UDFA where the states with two circles are the accepting states

Let $n, m \in \mathbb{N}^+$. If $|A| = n$ and $|B| = m$ is assumed, then $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_m\}$.

Since the strings accepted by UDFAs are sequences of the same symbol, we can identify a string with its length. Then, the set of strings accepted by a given UDFA will be represented by a subset of the natural numbers. Any natural number k that belongs to such a subset represents the string of length k. Thus, we say that (v, w) represents the language

$$\{a_1, a_2, \ldots, a_n\} \cup \{b_1 + |w|k, b_2 + |w|k, \ldots, b_m + |w|k \mid k \in \mathbb{N}\} \tag{3}$$

For example, the URL that is given by the automaton in Figure 2 is represented by $\{1, 2\} \cup \{3 + 3k, 4 + 3k \mid k \in \mathbb{N}\}$. In the sequel we use a shorter notation where the set of (3) is given by

$$\{\{a_i, b_j + |w|k\}_{k\in\mathbb{N}}\}_{i=1,2,\ldots,n,\ j=1,2,\ldots,m} \tag{4}$$

Without loss of generality, we will use the minimal UDFA (for short, MUDFA) that represents a given URL to obtain the previous notation for the URLs, in this way, we will have the minimal m, that will simplify the calculus of the similarity measure, and a unique representation for each CURL (since there is a unique MUDFA for any given URL).

A UDFA is cyclic if its initial word is empty. Then, we represent it as a vector (λ, w) where $w \in \{0, 1\}^+$ and λ is the empty string. We say that a URL R is a CURL if its MUDFA is cyclic.

Therefore, the notation for a CURL M, that is represented by the UDFA (λ, w), will be

$$M = \{\{b_i + |w|k\}_{k\in\mathbb{N}}\}_{i=1,2,\ldots,m}. \tag{5}$$

Thus a CURL is given by an infinite set of natural numbers, more precisely, by the union of a finite number of periodic sequences (with a fixed period). It is clear that $b_i < |w|$ for any $i \in \{1, 2, \ldots, m\}$.

3. A Similarity Measure for CURLs

3.1. *Similarity between two successions*

It is natural to say that two successions have the similarity 0 if they have no numbers in common. Therefore we are interested in the cases where the intersection of the two successions of natural numbers is not empty.

Lemma 1. *For each two natural successions* $A = \{a+bn\}_{n\in\mathbb{N}}$ *and* $B = \{c+dk\}_{k\in\mathbb{N}}$, $A\cap B \neq \emptyset$ *if and only if* $c-a$ *is a multiple of* $gcd(b,d)$ *(where* $gcd(b,d)$ *is the greatest common divisor of* b *and* d*).*

Proof. $a + bn = c + dk$ if and only if $c - a = bn - dk$. By the Main Theorem on *gcd*, there is a solution in $\mathbb{Z}$ of this equation, if and only if $c - a$ is a multiple of $gcd(b, d)$. □

Lemma 2. *Let* M *be a CURL. Given* $A, B \in M$, $A \cap B \neq \emptyset$ *if and only if* $A = B$.

Proof. Let us suppose that $A \cap B \neq \emptyset$. If we assume that $A = \{a_i + bk\}_{k\in\mathbb{N}}$ and $B = \{a_j + bk\}_{k\in\mathbb{N}}$, by Theorem 1, $A \cap B \neq \emptyset$ if and only if $|a_i - a_j|$ is a multiple of $gcd(b, b) = b$. Since $a_i, a_j < b$, we have $0 \leq |a_i - a_j| < b$. Then $|a_i - a_j|$ is a multiple of $gcd(b, b) = b$ if and only if $|a_i - a_j| = 0$, that is, $A = B$. □

Let $A = \{a + bn\}_{n\in\mathbb{N}}$ and $B = \{c + dk\}_{k\in\mathbb{N}}$ be two natural successions. We use the frequency in which the overlapped elements, i.e., elements which are in A as well as in B, appear in A as the measure of the overlap (thus it reflects the portion of elements of B in A).

Let

$$T = \{k \in \mathbb{N} \mid \frac{c-a}{b} + \frac{d}{b}k \in \mathbb{N}\}$$

be the set of natural numbers such that the element $c + dk$ of B is contained in A. Furthermore, let t be the minimal number in T. We determine the amount that has to be added to t in order to obtain another element of the set T. Thus, $x \in \mathbb{N}$ with $\frac{c-a}{b} + \frac{d}{b}(t + x) \in \mathbb{N}$ is looked for. Since

$$\frac{c-a}{b} + \frac{d}{b}(t + x) \in \mathbb{N} \text{ if and only if } \frac{c-a}{b} + \frac{d}{b}t + \frac{d}{b}x \in \mathbb{N}$$

and $\frac{c-a}{b} + \frac{d}{b}t \in \mathbb{N}$, we have $\frac{c-a}{b} + \frac{d}{b}t + \frac{d}{b}x \in \mathbb{N}$ if and only if $\frac{d}{b}x \in \mathbb{N}$ if and only if $x = \frac{b}{gcd(b,d)}m$ for some $m \in \mathbb{N}$.

So, if $T = \{t + \frac{b}{gcd(b,d)}m \mid m \in \mathbb{N}\}$, then the overlapped terms belong to

$$T' = \{\frac{c-a}{b} + \frac{d}{b}(t + \frac{b}{gcd(b,d)}m) \mid m \in \mathbb{N}\}.$$

Therefore the distance of two successive elements of T' is given by

$$\frac{c-a}{b} + \frac{d}{b}(t + \frac{b}{gcd(b,d)}m) - [\frac{c-a}{b} + \frac{d}{b}(t + \frac{b}{gcd(b,d)}(m-1))] = \frac{d}{gcd(b,d)}.$$

So, starting from t can be affirmed that a natural number that belongs to T' will be found in A every $\frac{d}{gcd(b,d)}$ terms.

Therefore, $\frac{gcd(b,d)}{d}$ can be considered as the overlap of A with B. That is, we have done a partition of the succession A into d disjoint subsets and $gcd(b,d)$ words of them belong to B.

Definition 3. *The overlap of an infinite succession $A = \{a+bn\}_{n\in\mathbb{N}}$ with another one $B = \{c+dk\}_{k\in\mathbb{N}}$, that we will call $ISO_{A,B}$ (for Infinite Successions Overlap), is defined as:*

$$ISO_{A,B} = \begin{cases} \frac{gcd(b,d)}{d} & \text{if } A \cap B \neq \emptyset \\ 0 & \text{in other case} \end{cases}$$

Lemma 4. *Let $A = \{a+bn\}_{n\in\mathbb{N}}$ and $B = \{c+dk\}_{k\in\mathbb{N}}$ be two natural successions. Then $A \subseteq B$ if and only if $ISO_{A,B} = 1$.*

Proof. $ISO_{A,B} = 1$ if and only if $gcd(b,d) = d$ if and only if $b = du$ for some $u \in \mathbb{N}$. Since $ISO_{A,B} = 1$ we have $A \cap B \neq \emptyset$ and thus, by Lemma 1, $c - a = t \cdot gcd(b,d) = td$ for some $t \in \mathbb{N}$. For $n \in \mathbb{N}$, we get $a + bn = c + dt + dun = c + d(t + un)$, which proves that any element of A is contained in B or equivalently, $A \subseteq B$.

If $A \subseteq B$, then for any $n \in \mathbb{N}$, there exists $m \in \mathbb{N}$ such that $a + bn = c + dm$. In particular, it holds for $n = 1$, that is, there exists $m \in \mathbb{N}$ such that $a + b = c + dm$. Since $c - a = td$, we have $b = d(t + m)$. Thus $gcd(b,d) = d$ and $ISO_{A,B} = 1$. □

The similarity of two successions combines $ISO_{A,B}$ and $ISO_{B,A}$.

Definition 5. *The similarity measure between two infinite successions $A = \{a + bn\}_{n\in\mathbb{N}}$ and $B = \{c + dk\}_{k\in\mathbb{N}}$, that we will call $ISS_{A,B}$ (for Infinite Successions Similarity), is defined as:*

$$ISS_{A,B} = \begin{cases} \frac{ISO_{A,B} + ISO_{A,B}}{2} & \text{if } A \cap B \neq \emptyset \\ 0 & \text{in other case} \end{cases}$$

Given two infinite successions A and B, $0 \leq ISS_{A,B} \leq 1$, since $0 \leq ISO_{A,B} \leq 1$ and $0 \leq ISO_{B,A} \leq 1$ for any infinite successions A and B.

3.2. *The proposed similarity measure for CURLs*

In this section, we define the similarity measure for CURLs by using the similarity measure between two successions that has been defined in the previous section.

Given two CURLs M and N, we have that $M \cap N \neq \emptyset$ if and only if there exist at least $A \in M$ and $B \in N$ such that $A \cap B \neq \emptyset$.

Definition 6. *Let M and N be two CURLs. We define the overlap of M with N,*

that we will call $URLO_{M,N}$ *(for Unary Regular Languages Overlap), as:*

$$URLO_{M,N} = \begin{cases} \frac{1}{m} \sum\limits_{\substack{A \in M \\ B \in N}} ISO_{A,B} & \text{if } M \cap N \neq \emptyset \\ 0 & \text{in other case} \end{cases}$$

where $m \in \mathbb{N}$ *is the number of successions of* M*.*

By following the same reasoning as in the previous section, we define the similarity measure between two CURLs as follows:

Definition 7. *Let* M *and* N *be two CURLs. We define the similarity measure between* M *and* N*, that we will call* $URLS_{M,N}$ *(for Unary Regular Languages Similarity), as:*

$$URLS_{M,N} = \begin{cases} \frac{URLO_{M,N} + URLO_{N,M}}{2} & \text{if } M \cap N \neq \emptyset \\ 0 & \text{in other case} \end{cases}$$

Our definitions require that the regular sets R and S are given as sets M and N of periodic sequences which are induced by the minimal automata of R and S. We now prove that any other description as sets M' and N' of periodic sequences which are induced by DFAs accepting R and S give the same similarity.

Theorem 8. *Let* $M = \{\{a_i + bk\}_{k \in \mathbb{N}}\}_{i=1,2,\ldots,m}$ *and* $\overline{M} = \{\{a'_i + b'k\}_{k \in \mathbb{N}}\}_{i=1,2,\ldots,m'}$ *be two descriptions of the regular set* R*, and let* $N = \{\{c_i + dk\}_{k \in \mathbb{N}}\}_{i=1,2,\ldots,n}$ *and* $\overline{N} = \{\{c'_i + d'k\}_{k \in \mathbb{N}}\}_{i=1,2,\ldots,n'}$ *be two descriptions of the regular set* S*. Then*

$$URLS_{M,N} = URLS_{\overline{M},\overline{N}}.$$

Proof. We first compute $URLS_{M,N}$. Let us assume that there are q pairs (i, j) such that $\{a_i + bk\} \cap \{c_j + dk\} \neq \emptyset$. Then we get

$$URLO_{M,N} = \frac{1}{m} \sum_{i,j} ISO_{\{a_i+bk\},\{c_j+dk\}} = \frac{1}{m} \cdot q \cdot \frac{gcd(b,d)}{d} \tag{3}$$

and an analogous result for $URLO_{N,M}$ taking n and b instead of m and d, respectively. Thus

$$URLS_{M,N} = \frac{\frac{q \cdot gcd(b,d)}{md} + \frac{q \cdot gcd(b,d)}{nb}}{2} = \frac{q \cdot gcd(b,d)(md + nb)}{2nmbd}. \tag{4}$$

Now we prove that other special representations of the sets R and S give the same value. Let z be the lowest common multiple of b and d. We set

$$g = \frac{z}{b} \text{ and } h = \frac{z}{d}.$$

Then we also have

$$g = \frac{d}{gcd(b,d)} \text{ and } h = \frac{b}{gcd(b,d)}. \tag{5}$$

We now construct the successions

$$M' = \{a_i + vb + kz \mid 1 \leq i \leq m,\ 0 \leq v \leq g-1\}_{k \in \mathbb{N}}$$

and

$$N' = \{c_j + v'd + kz \mid 1 \leq j \leq n,\ 0 \leq v' \leq h-1\}_{k \in \mathbb{N}}.$$

Thus we have ng successions in M' and mh successions in N'. Obviously, M and M', as well as N and N', describe the same regular languages. By Lemma 2, all successions of M' and N' are pairwise disjoint. As above we get

$$URLO_{M',N'} = \frac{1}{mg} \sum_{i,j,v,v'} ISO_{\{a_i+vb+kz\},\{c_j+v'd+kz\}} = \frac{q}{mg} = \frac{q \cdot gcd(b,d)}{md}$$

and an analogous result for $URLO_{N',M'}$ which leads to

$$URLS_{M',N'} = \frac{q \cdot gcd(b,d)(md+nb)}{2nmbd} = \frac{q(mg+nh)}{2nmgh} \tag{6}$$

Therefore we have $URLS_{M,N} = URLS_{M',N'}$.

The same argumentation can be used if we consider representations M'_u and N'_u which are based on a multiple u of z.

Let y be the lowest common multiple of b, d, b', d'. Then we get

$$URLS_{M,N} = URLS_{M'_y,N'_y} \text{ and } URLS_{\overline{M},\overline{N}} = URLS_{\overline{M'_y},\overline{N'_y}}. \tag{7}$$

Since M'_y and $(M')'_y$ describe R we get that

$$U = \{a_i + sb \mid 1 \leq i \leq m,\ 0 \leq s \leq \frac{u}{b} - 1\} \text{ and } U' = \{a'_i + tb' \mid 1 \leq i \leq m',\ 0 \leq t \leq \frac{u}{b'} - 1\}$$

describe the set of all words in R of length at most $y-1$. Thus $U = U'$ and consequently $M'_y = (M')'_y$ (since we extend U and U' only by adding multiples of y). Analogously, $N_y = (N')_y$. Therefore, by (7),

$$URLS_{M,N} = URLS_{M'_y,N'_y} = URLS_{\overline{M'_y},\overline{N'_y}} = URLS_{\overline{M},\overline{N}}. \qquad \square$$

Thus, in the sequel, we use the description which is most appropriate for our proofs.

We now present some elementary properties of the similarity measure, we particularly show that it is a value between 0 and 1 (which is a desired property).

Lemma 9. *$0 \leq URLS_{M,N} \leq 1$ for any two CURLs M and N.*

Proof. Let M and N be two CURLs. The relation $0 \leq URLS_{M,N}$ is obvious.

Let $M = \{\{a_i + bk\}_{k \in \mathbb{N}}\}_{i=1,2,\ldots,m}$ and $N = \{\{c_j + dk\}_{k \in \mathbb{N}}\}_{j=1,2,\ldots,n}$ with $n, m \in \mathbb{N}^+$. Let z be the lowest common multiple of b and d. Moreover, let

$$g = \frac{z}{b} = \frac{d}{gcd(b,d)} \text{ and } h = \frac{z}{d} = \frac{b}{gcd(b,d)}.$$

Then we can describe M and N as

$$\begin{aligned} M &= \{A_{1,1}, A_{1,2}, \ldots, A_{1,g}, A_{2,1}, \ldots, A_{2,g}, \ldots, A_{m,1}, \ldots, A_{m,g}\} \\ &\quad \text{with } A_{i,p} = \{a_{i,p} + zk\}_{k\in\mathbb{N}}, a_{i,p} = a_i + (p-1)b \text{ for } 1 \le i \le m, 1 \le p \le g, \\ N &= \{C_{1,1}, C_{1,2}, \ldots, C_{1,h}, C_{2,1}, \ldots, C_{2,h}, \ldots, C_{n,1}, \ldots, C_{n,h}\} \\ &\quad \text{with } C_{j,l} = \{c_{j,l} + zk\}_{k\in\mathbb{N}}, c_{j,l} = c_j + (l-1)d \text{ for } 1 \le j \le n, 1 \le l \le h. \end{aligned}$$

Let $A_{i,p} \cap C_{j,t} \neq \emptyset$. Then by Lemma 1, $a_{i,p} - c_{j,t} = gcd(z,z)s = zs$ for some $s \in \mathbb{N}$. Moreover, $a_{i,p} < z$ and $c_{j,t} < z$ implies $|a_{i,p} - c_{j,t}| < z$. Thus, necessarily $a_{i,p} = c_{j,t}$. This implies immediately $A_{i,p} \subseteq C_{j,t}$. Then, by Lemma 4, $ISO_{A_{i,p},C_{j,t}} = 1$. Moreover, since all the $c_{j,t}$ are different, for any $A_{i,p} \in M$, there exists at most one $C_{j,t} \in N$ such that $A_{i,p} \cap C_{j,t} \neq \emptyset$.

If $A_{i,p} \cap C_{x,y} = \emptyset$, then $ISO_{A_{i,p},C_{x,y}} = 0$. Thus we get

$$\sum_{C_{j,t}\in N} ISO_{A_{i,p},C_{j,t}} = \begin{cases} 1 \text{ if } A_{i,p} \cap N \neq \emptyset \\ 0 \text{ if } A_{i,p} \cap N = \emptyset \end{cases}.$$

Now we obtain

$$\begin{aligned} URLO_{M,N} &= \frac{1}{mg} \sum_{\substack{A_{i,p}\in M \\ C_{j,t}\in N}} ISO_{A_{i,p},C_{j,t}} \\ &= \frac{1}{mg} \sum_{A_{i,p}\in M} \Big(\sum_{C_{j,t}\in N} ISO_{A_{i,p},C_{j,t}} \Big) \\ &\le \frac{1}{mg} \sum_{A_{i,p}\in M} 1 \qquad (8) \\ &= \frac{1}{mg} \cdot mg = 1. \end{aligned}$$

Analogously, we have $URLO_{N,M} \le 1$. Thus, by the definition of $URLS_{M,N}$, we get $URLS_{M,N} \le 1$. □

Lemma 10. *Let M and N be CURLs. $URLO_{M,N} = 1$ if and only if $M \subseteq N$.*

Proof. We consider the presentations given in the proof of Lemma 9.

Let us suppose $M \subseteq N$. Since $M \subseteq N$ if only if $A_{i,p} \subseteq N$ for any $A_{i,p} \in M$, we get $A_{i,p} \cap N \neq \emptyset$ for any $A_{i,p} \in M$. Therefore, by the proof of Lemma 9, $\sum_{C_{j,t}\in N} ISO_{A_{i,p},C_{j,t}} = 1$ for any $A_{i,p} \in M$. Thus we obtain an equality in (8), which proves that $URLO_{M,N} = 1$.

Conversely, $URLO_{M,N} = 1$ if and only if $\sum_{C\in N} ISO_{A,C} = 1$ for any $A \in M$. Therefore, $A \cap N \neq \emptyset$ for any $A \in M$. Consequently, for any $A \in M$, there is a $C \in N$ such that $A \cap C \neq \emptyset$. As in the proof of Theorem 9, we can show that

$A \cap C \neq \emptyset$ implies $A \subseteq C$. Thus, for any $A \in M$, there is a $C \in N$ with $A \subseteq C$. This implies $A \subseteq N$ for any $A \in M$ which gives $M \subseteq N$. □

We have shown that, for any CURLs M and N, $0 \leq URLS_{M,N} \leq 1$. We will now show that also the converse holds, i.e., every number x with $0 \leq x \leq 1$ can be obtained as a similarity.

Theorem 11. *The measure URLS is dense, i.e., for any (rational) number $x \in [0,1]$ and any $\varepsilon \geq 0$, there are CURLs M and N such that*

$$|URLS_{M,N} - x| \leq \varepsilon.$$

Proof. Obviously, for the sequences $M = \{0 + 2k\}_{k\in\mathbb{N}}$ and $N = \{1 + 2k\}_{k\in\mathbb{N}}$, we get $URLS_{M,M} = 1$ and $URLS_{M,N} = 0$.

Let $0 < x < 1$. Then we choose prime numbers p and q sufficiently large such that $p < q$, $xp \leq p - 1$ and $\frac{1}{2p} + \frac{1}{2q} \leq \varepsilon$. Then we also have $xq \leq q - 1$. We now choose $m = \lceil xp \rceil$ and $n = \lceil xq \rceil$ and

$$M = \{\{i + pk\}_{k\in\mathbb{N}}\}_{i=1,2,\ldots,m} \text{ and } N = \{\{j + qk\}_{k\in\mathbb{N}}\}_{j=1,2,\ldots,n}.$$

Since the greatest common divisor of p and q is 1 and any difference $i-j$, $1 \leq i \leq m$ and $1 \leq j \leq n$, is a multiple of 1, any pair of successions $\{i + pk\}$ and $\{j + qk\}$ has an non-empty intersection. Thus we get

$$\begin{aligned} URLS_{M,N} &= \frac{URLO_{M,N} + URLO_{N,M}}{2} \\ &= \frac{n\frac{1}{q} + m\frac{1}{p}}{2} \\ &= \frac{np + mq}{2pq}. \end{aligned}$$

If we take into consideration that $xp \leq m \leq xp + 1$ and $xq \leq n \leq xq + 1$, we get

$$x = \frac{xpq + xqp}{2pq} \leq \frac{np + mq}{2pq} \leq \frac{(xp + 1)q + (xq + 1)p}{2pq} = x + \frac{1}{2p} + \frac{1}{2q} \leq x + \varepsilon.$$

Now the statement follows immediately. □

4. Jaccard Coefficient and Sørensen Coefficient for CURLs

4.1. *Definition of a Jaccard Coefficient and Sørensen Coefficient for CURLs*

The Jaccard coefficient given in the Introduction is a well-known measure for the similarity of finite sets. As we said before, we can not use this measure directly for CURLs because both sets are infinite if the intersection is non-empty. In this section, we propose an appropriate definition of the Jaccard coefficient which can be used for CURLs.

Let

$$M = \{\{a_i + bk\}_{k\in\mathbb{N}}\}_{i=1,\ldots,m} \text{ and } N = \{\{c_j + dk\}_{k\in\mathbb{N}}\}_{j=1,\ldots,n}$$

be two CURLs consisting of n and m sequences, respectively. Let

$$M_{s,t} = M \cap \{i \mid s \leq i \leq t\} \text{ and } N_{s,t} = N \cap \{i \mid s \leq i \leq t\}$$

be the subsets of M and N, consisting of all numbers greater or equal to s and smaller or equal to t. Then a natural definition of a Jaccard coefficient would be

$$\lim_{t\to\infty} \frac{|M_{0,t} \cap N_{0,t}|}{|M_{0,t} \cup N_{0,t}|}.$$

In order to use this definition we have to show that the limit exists. This will now be done. Let z be the lowest common multiple of b and d. Then it is clear that

$$M_{rz,(r+1)z-1} = M_{0,z-1} + rz = \{y + rz \mid y \in M_{0,z-1}\}$$

for all $r \geq 0$. Hence, for $t = rz + u$,

$$|M_{0,t} \cap N_{0,t}| = r|M_{0,z-1} \cap N_{0,z-1}| + |M_{0,u} \cap N_{0,u}|,$$
$$|M_{0,t} \cup N_{0,t}| = r|M_{0,z-1} \cup N_{0,z-1}| + |M_{0,u} \cup N_{0,u}|.$$

Therefore we get

$$\begin{aligned}
&\frac{|M_{0,t} \cap N_{0,t}|}{|M_{0,t} \cup N_{0,t}|} = \frac{r|M_{0,z-1} \cap N_{0,z-1}| + |M_{0,u} \cap N_{0,u}|}{r|M_{0,z-1} \cup N_{0,z-1}| + |M_{0,u} \cup N_{0,u}|} \\
&= \frac{r|M_{0,z-1} \cap N_{0,z-1}|}{r|M_{0,z-1} \cup N_{0,z-1}| + |M_{0,u} \cup N_{0,u}|} + \frac{|M_{0,u} \cap N_{0,u}|}{r|M_{0,z-1} \cup N_{0,z-1}| + |M_{0,u} \cup N_{0,u}|} \\
&= \frac{|M_{0,z-1} \cap N_{0,z-1}|}{|M_{0,z-1} \cup N_{0,z-1}| + \frac{|M_{0,u} \cup N_{0,u}|}{r}} + \frac{|M_{0,u} \cap N_{0,u}|}{r|M_{0,z-1} \cup N_{0,z-1}| + |M_{0,u} \cup N_{0,u}|}
\end{aligned}$$

which implies

$$\begin{aligned}
&\lim_{t\to\infty} \frac{|M_{0,t} \cap N_{0,t}|}{|M_{0,t} \cup N_{0,t}|} \\
&= \lim_{r\to\infty} \frac{|M_{0,z-1} \cap N_{0,z-1}|}{|M_{0,z-1} \cup N_{0,z-1}| + \frac{|M_{0,u} \cup N_{0,u}|}{r}} + \frac{|M_{0,u} \cap N_{0,u}|}{r|M_{0,z-1} \cup N_{0,z-1}| + |M_{0,u} \cup N_{0,u}|} \\
&= \frac{|M_{0,z-1} \cap N_{0,z-1}|}{|M_{0,z-1} \cup N_{0,z-1}|}.
\end{aligned}$$

Therefore we give the following definition.

Definition 12. *For two cyclic unary regular languages $M = \{\{a_i+bk\}_{k\in\mathbb{N}}\}_{i=1,\ldots,m}$ and $N = \{\{c_j+dk\}_{k\in\mathbb{N}}\}_{j=1,\ldots,n}$, we define the Jaccard coefficient $JC_{M,N}$ of M and N by*

$$JC_{M,N} = \frac{|M_{0,z-1} \cap N_{0,z-1}|}{|M_{0,z-1} \cup N_{0,z-1}|},$$

where z is the smallest common multiple of b and d.

Let us see that the measure JC does not depend on the representation of the CURLs.

Theorem 13. *Let $M = \{\{a_i+bk\}_{k\in\mathbb{N}}\}_{i=1,2,\ldots,m}$ and $\overline{M} = \{\{a'_i+b'k\}_{k\in\mathbb{N}}\}_{i=1,2,\ldots,m'}$ be two descriptions of the regular set R, and let $N = \{\{c_i + dk\}_{k\in\mathbb{N}}\}_{i=1,2,\ldots,n}$ and $\overline{N} = \{\{c'_i + d'k\}_{k\in\mathbb{N}}\}_{i=1,2,\ldots,n'}$ be two descriptions of the regular set S. Then*

$$JC_{M,N} = JC_{\overline{M},\overline{N}}.$$

Proof. Let z be the lowest common multiple of b and d, and u the lowest common multiple of b, b', d, d'. Then $u = tz$ for some $t \in \mathbb{N}^+$. Then $|M_{0,u-1} \cap N_{0,u-1}| = t|M_{0,z-1} \cap N_{0,z-1}|$ and $|M_{0,u-1} \cup N_{0,u-1}| = t|M_{0,z-1} \cup N_{0,z-1}|$ and therefore

$$JC_{M,N} = \frac{|M_{0,u-1} \cap N_{0,u-1}|}{|M_{0,u-1} \cup N_{0,u-1}|}.$$

Analogously,

$$JC_{\overline{M},\overline{N}} = \frac{|\overline{M}_{0,u-1} \cap \overline{N}_{0,u-1}|}{|\overline{M}_{0,u-1} \cup \overline{N}_{0,u-1}|}.$$

Now the equality $JC_{M,N} = JC_{\overline{M},\overline{N}}$ follows because $M_{0,u-1} = \overline{M}_{0,u-1}$ and $N_{0,u-1} = \overline{N}_{0,u-1}$ since the same languages R and S are described. □

We now determine $JC_{M,N}$ for two CURLs $M = \{\{a_i + bk\}_{k\in\mathbb{N}}\}_{i=1,\ldots,m}$ and $N = \{\{c_j + dk\}_{k\in\mathbb{N}}\}_{j=1,\ldots,n}$. Let us assume that there are q pairs (i, j) such that $\{a_i + bk\} \cap \{c_j + dk\} \neq \emptyset$.

Let $g = \frac{d}{gcd(b,d)}$ and $h = \frac{b}{gcd(b,d)}$. We mention the following fact.

If the two successions M and N have a non-empty intersection, then

$$\{a_i, a_i + b, a + 2b, \ldots a_i + (g-1)b\} \cap \{c_j, c_j + d, c_j + 2d, \ldots, c_j + (h-1)d\}$$

consists only of one element.

(Assume that the intersection contains at least two elements x and y. Without loss of generality let $x < y$. Then

$$x = a + x'b = c + x''d \text{ and } y = a + x'b + y'b = c + x''d + y''d$$

which gives $y'b = y''d = p$. Since b and d are divisors of p, we have $p \geq z$. Thus $y > z$ in contrast to the choice of y.)

We construct the sets M' and N' as in the proof of Theorem 8 and show that $U = M'_{0,z-1} \cap N'_{0,z-1}$ contains exactly q elements. By the fact given above, U has at most q elements, since we have only q pairs of intersecting successions of M and N. However, if the intersection of two pairs are equal, then $a_{i_1} + v_1 b = c_{j_1} + v'_1 c = a_{i_2} + v_2 b = c_{j_2} + v'_2 c$, which gives by Lemma 2 that $a_{i_1} = a_{i_2}$ and $c_{j_1} = c_{j_2}$, i.e., the two pairs coincide.

Furthermore, $M'_{0,z-1} \cup N'_{0,z-1}$ contains $mg + nh - q$ elements because we have $mg + nh$ successions and q elements are counted twice. Hence

$$JC_{M,N} = JC_{M',N'} = \frac{q}{mg + nh - q}. \tag{9}$$

Obviously, $0 \leq JC_{M,N} \leq 1$ for all CURLs M and N. We now show the denseness of the measure JC.

Theorem 14. *For any rational number $r \in [0, 1]$, there are CURLs M and N such that $JC_{M,N} = r$, i.e., the measure JC is dense.*

Proof.

If $r = 0$, we can choose $M = \{0 + 2k\}_{k\in\mathbb{N}}$ and $N = \{1 + 2k\}_{k\in\mathbb{N}}$ and then $JC_{M,N} = 0$.

Let $r \in (0, 1]$ be a rational number, then $r = \frac{x}{y}$ for any $x, y \in \mathbb{N}$ with $x \leq y$.

Let $b \in \mathbb{N}$ such that $b > y$, let us define

$$M = \{\{i + bk\}_{k\in\mathbb{N}}\}_{i=1,2,\ldots,x} \text{ and } N = \{\{j + bk\}_{k\in\mathbb{N}}\}_{j=1,2,\ldots,y}$$

Since the greatest common divisor of b and b is b, given $i \in \{1, 2, \ldots, x\}$ and $j \in \{1, 2, \ldots, y\}$, $\{i + bk\}_{k\in\mathbb{N}}$ and $\{j + bk\}_{k\in\mathbb{N}}$ have a non-empty intersection if and only if $i - j = 0$. Therefore, there are x pairs (i, j) such that $i - j = 0$.

Therefore, by the equation 9 taking into consideration that $g = h = 1$, we have

$$JC_{M,N} = \frac{q}{xg + yh - q} = \frac{x}{x + y - x} = \frac{x}{y}$$

Now the statement follows immediately. □

Analogously, we can consider the Sørensen coefficient as the limit (for $t \to \infty$) of the Sørensen coefficients of the initial parts $M_{0,t}$ and $N_{0,t}$. This leads to the following definition.

Definition 15. *For two cyclic unary regular languages $M = \{\{a_i + bk\}_{k\in\mathbb{N}}\}_{i=1,\ldots,m}$ and $N = \{\{c_j + dk\}_{k\in\mathbb{N}}\}_{j=1,\ldots,n}$, we define the Sørensen coefficient $SC_{M,N}$ of M and N by*

$$JC_{M,N} = \frac{2 \cdot |M_{0,z-1} \cap N_{0,z-1}|}{|M_{0,z-1}| + |N_{0,z-1}|},$$

where z is the smallest common multiple of b and d.

Moreover, using the same arguments as above, we show that this definition is independent of the representation and that, for two cyclic unary regular languages $M = \{\{a_i + bk\}_{k\in\mathbb{N}}\}_{i=1,\ldots,m}$ and $N = \{\{c_j + dk\}_{k\in\mathbb{N}}\}_{j=1,\ldots,n}$,

$$SC_{M,N} = SC_{M',N'} = \frac{2q}{mg + nh}, \tag{10}$$

where q is the number of pairs (i, j) such that $\{a_i + bk\} \cap \{c_j + dk\} \neq \emptyset$, $g = \frac{d}{gcd(b,d)}$ and $h = \frac{b}{gcd(b,d)}$.

Theorem 16. *For any rational number $r \in [0,1]$, there are CURLs M and N such that $SC_{M,N} = r$, i.e., the measure SC is dense.*

Proof. Any rational number $r \in [0,1]$ can be given in the form $r = \frac{2x}{x+y}$ with $x \leq y$ (since $r = \frac{x}{b} = \frac{2x}{2b}$ for some $x \leq b$ and then $2b = x + y$ for some $y \geq x$). Now the sets given in the proof of Theorem 14 and the considerations in that proof show the statement. □

4.2. *Comparing the measure URLS with the Jaccard and Sørensen Coefficients*

Now, given two CURLs M and N, let us compare the similarity measure $URLS_{M,N}$ with the Jaccard coefficient $JC_{M,N}$ and the Sørensen coefficient $SC_{M,N}$ that has been defined in the previous subsection.

Theorem 17. *$URLS_{M,N} \geq SC_{M,N} \geq JC_{M,N}$ for any CURLs M and N.*

Proof. Let $M = \{\{a_i + bk\}_{k\in\mathbb{N}}\}_{i=1,\ldots,m}$ and $N = \{\{c_j + dk\}_{k\in\mathbb{N}}\}_{j=1,\ldots,n}$. Moreover, let us assume that there are q pairs (i,j) such that $\{a_i + bk\} \cap \{c_j + dk\} \neq \emptyset$, and let $g = \frac{d}{gcd(b,d)}$ and $h = \frac{b}{gcd(b,d)}$.

Obviously, $(mg-nh)^2 = (mg)^2 - 2mgnh + (nh)^2 \geq 0$ which implies $(mg+nh)^2 = (mg)^2 + 2mgnh + (nh)^2 \geq 4mgnh$ or equivalently

$$\frac{q(mg+nh)}{2mgnh} \geq \frac{2q}{mg+nh},$$

i.e., $URLS_{M,N} \geq SC_{M,N}$ by (4), (6) and (10).

Furthermore, $mg \geq q$ and $nh \geq q$. Therefore $mg+nh-2q \geq 0$. By multiplication with q and adding $mgq+nhq$, we get $2q(mg+nh)-2q^2 \geq q(mg+nh)$ or equivalently

$$\frac{2q}{mg+nh} \geq \frac{q}{mg+nh-q},$$

i.e., $SC_{M,N} \geq JC_{M,N}$ by (9) and (9). □

Corollary 18. *Let M and N be two CURLs. Then $URLS_{M,N} = SC_{M,N} = JC_{M,N}$ if and only if $M = N$.*

Proof. Let us suppose $M = N$. By Lemma 10, $M = N$ if and only if $URLO_{M,N} = URLO_{N,M} = 1$. Since $URLO_{M,N} = \frac{q}{mg}$ and $URLO_{M,N} = \frac{q}{nh}$, we have $M = N$ if and only if $q = mg = nh$.

Since $q = mg = nh$ if and only if $qmg = (mg)^2$ and $qnh = (nh)^2$, we have that

$$(mg)^2 + (nh)^2 = qmg + qnh.$$

Moreover,

$$(mg)^2 + (nh)^2 = qmg + qnh$$

if and only if

$$q((mg)^2 + (nh)^2) - q^2(mg + nh) = 0$$

if and only if

$$q((mg)^2 + (nh)^2) - q^2(mg + nh) + 2qnhmg = 2qnhmg$$

if and only if

$$q(mg + nh)(mg + nh - q) = 2qnhmg$$

if and only if

$$URLS_{M,N} = \frac{q(mg + nh)}{2nhmg} = \frac{q}{mg + nh - q} = JC_{M,N}.$$

The statement follows by Theorem 17 □

Theorem 19. *Let $M_1, M_2, \dots M_i \dots$ be an infinite sequence of CURLS. Then the following three statements are equivalent*
i) $\lim_{i\to\infty} URLS_{M_i,N} = 1$,
ii) $\lim_{i\to\infty} SC_{M_i,N} = 1$,
iii) $\lim_{i\to\infty} JC_{M_i,N} = 1$.

Proof. iii) implies i). Assume that iii) holds. Thus, for any real number $\varepsilon \geq 0$, there is a natural number n such that $1 - JC_{M_i,N} \leq \varepsilon$ for $i \geq n$. Then by Theorem 17, $1 - URLS_{M_i,N} \leq \varepsilon$ for $i \geq n$. Therefore i) holds.

ii) implies i) and iii) implies ii) can be shown analogously.

i) implies iii). Assume that i) holds. Then for any $\varepsilon \geq 0$, there is a natural number n such that $1 - URLS_{M_i,N} < \varepsilon$ for $i \geq n$. If M_i and N contain m and n successions, respectively, we get

$$1 - \frac{q(mg + nh)}{2mgnh} = 1 - \big(\frac{qmg}{2mgnh} + \frac{qnh}{2mgnh}\big) = 1 - \frac{q}{2nh} - \frac{q}{2mg} < \varepsilon.$$

Thus

$$2 - \frac{q}{nh} - \frac{q}{mg} < 2\varepsilon \text{ and } (1 - \frac{q}{nh}) + (1 - \frac{q}{mg}) < 2\varepsilon.$$

Consequently,

$$1 - \frac{q}{nh} < 2\varepsilon \text{ and } 1 - \frac{q}{mg} < 2\varepsilon,$$

or, equivalently,

$$mg - q < 2\varepsilon mg \text{ and } nh - q < 2\varepsilon nh. \tag{11}$$

Now we obtain

$$\begin{aligned} 1 - JC_{M_i,N} &= 1 - \frac{q}{mg+nh-q} = \frac{mg+nh-2q}{mg+nh-q} = \frac{(mg-q)+(nh-q)}{mg+nh-q} \\ &< \frac{2\varepsilon mg + 2\varepsilon nh}{mg+nh-q} = 2\varepsilon \frac{mg+nh}{mg+nh-q} \text{ (by (11))} \\ &= 2\varepsilon \frac{1}{1-\frac{q}{mg+nh}} \\ &< 4\varepsilon \qquad \text{(because } q \leq mg, q \leq nh, \text{ i. e., } \tfrac{q}{mg+nh} \leq \tfrac{1}{2}). \end{aligned}$$

Thus iii) is valid, too.

By a combination of the shown implications the assertion follows. □

5. A Dissimilarity Measure for CURLs

In this section, we will define a dissimilarity measure for CURLs by using the similarity measure that was defined in the previous section. That will be done in the same way as the Jaccard distance is defined by using the Jaccard coefficient.

Definition 20. *Let $n \in \mathbb{N}^+$. Let M and N be two CURLs. We define the dissimilarity measure between M and N, that we will call $URLD_{M,N}$ (for Unary Regular Languages Dissimilarity), as*

$$URLD_{M,N} = 1 - URLS_{M,N}$$

where $URLS_{M,N}$ is the similarity measure between M and N.

Then, we can say that the dissimilarity measure between CURLs is the proportion of strings that are not shared by such languages.

Given two CURLs M and N, $0 \leq URLO_{M,N} \leq 1$, we have $0 \leq URLS_{M,N} \leq 1$. Then, $0 \leq URLD_{M,N} \leq 1$ as in the Jaccard distance case.

In contrast to the Jaccard distance, the dissimilarity measure is not a metric distance since the triangle inequality is not satisfied. That can be proved by using the following counterexample: if M is the set of the odd numbers, N is the set of the even numbers and L is the set of the natural numbers, then

$$1 = URLD_{M,N} > URLD_{M,L} + URLD_{L,N} = 0.$$

However, the dissimilarity measure for CURLs is a semimetric distance, i.e a function d satisfying $d(x,y) \geq 0$, $d(x,y) = 0$ if and only if $x = y$, and $d(x,y) = d(y,x)$. Let us see that given two CURLs M and N, $URLD_{M,N}$ satisfies all the conditions to be a semimetric:

(1) $URLD_{M,N} \geq 0$ has been proved in the previous section.
(2) Let us see that $URLD_{M,N} = 0$ if and only if $M = N$.
First we have to show that $URLD_{M,M} = 0$. $URLD_{M,M} = 0$ holds if and only if $URLS_{M,M} = 1$. Since $URLS_{M,M} = URLO_{M,M} = \dfrac{1}{m} \displaystyle\sum_{\substack{A \in M \\ B \in M}} ISO_{A,B}$, $A \cap B \neq \emptyset$

if and only if $A = B$ (by Lemma 2) and $ISO_{A,A} = 1$ for any $A \in M$, we have $URLS_{M,M} = 1$.

Let us suppose that $URLD_{M,N} = 0$ for some $M \neq N$. Without loss of generality, let us assume that $M \nsubseteq N$, then $URLO_{M,N} < 1$ (by Lemma 10). Therefore, $URLS_{M,N} < 1$ and it implies $URLD_{M,N} \neq 0$. This is a contradiction, because we supposed $URLD_{M,N} = 0$. So, if $URLD_{M,N} = 0$, then $M = N$.

(3) Let us see $URLD_{M,N} = URLD_{N,M}$. We have

$$URLS_{M,N} = \frac{URLO_{M,N} + URLO_{N,M}}{2} = \frac{URLO_{N,M} + URLO_{M,N}}{2} = URLS_{N,M}$$

6. Discussion

In this work, we have proposed a similarity measure for CURLs by modifying the Jaccard similarity coefficient and the Sørensen coefficient. Moreover, we have defined a dissimilarity measure for CURLs by using that similarity measure.

Moreover we can also use the similarity and dissimilarity measures defined in this work for non-cyclic URLs. In that case, we consider the infinite set of strings that is generated by the loop of its respective MUDFA, since its initial word contributes to the language with only a finite number of strings, and we follow the same strategy of the cyclic case.

For two URLs (cyclic or non-cyclic) M and N, $URLD_{M,N} = 0$ if and only if $M = N$ (the identity of indiscernibles) is not always satisfied, as can be seen from the following counterexample: If $M = \{1, 4 + 2n\}_{n\in\mathbb{N}}$ and $N = \{2n\}_{n\in\mathbb{N}}$, then $URLD_{M,N} = 0$ and $M \neq N$. Thus, the dissimilarity measure for URLs is a symmetric distance and not a semimetric distance.

As a possible application of the proposed measure we can consider grammatical inference and retrieval theory. Evolutionary computation is an example of optimization technique where the search needs to be informed by a measure that compares individuals with a target. Inferring a CURL would mean just that, and this could be done with $URLS$, JC or SC. Considering the best individual in each of the generations computed by an evolutionary algorithm we would obtained a sequence of CURLs, in the form required by Theorem 19. If the algorithm performs well, this sequence would eventually converge to N with respect to some similarity (our measure, Sørensen and Jaccard coefficient), then it tends to N with respect to the two other similarities, too. We believe that a tendency can be seen easier by the use of our measure since it is greater than the two other ones, and therefore it approaches to 1 earlier. Thus we think that the new defined measure $URLS$ is more appropriate in these circumstances, i.e., URLS could be used as an indicator of convergence, outperfoming JC and SC.

Finally we mention that there are some proposals of distances $d(R, S)$ of two (unary) regular sets R and S, however, the corresponding similarities $1 - d(R, S)$ are not of interest for us, since the principle mentioned in the Introduction (sets are

more similar if the have more elements in common) is not satisfied by them, and we wanted to have a similarity measures for CURLs which follows this intuitive idea.

As examples we mention the Bodnarchuk distance for arbitrary languages, the Baire distance for unary regular languages, the Hamming distance and the information distance for cyclic unary regular languages.

The Bodnarchuk distance $BD(R,S)$ of two sets R and S of non-empty words is defined as

$$d(R,S) = \begin{cases} 0 & \text{if } R = S \\ \frac{1}{\min\{|w||w\in(R\setminus S)\cup(S\setminus R)\}} & \text{if } R \neq S \end{cases}$$

(see [6]). Thus the distance is the inverse of the length of the shortest word which gives a difference of the two languages. It is easy to see that $BD(A,B) = 1$ holds for

$$A = \{a^{101n+i} \mid n \geq 0, i \in \{1,3,4,5,\ldots 101\}\} \text{ and } B = \{a^{101n+i} \mid n \geq 0, 2 \leq i \leq 101\},$$

i.e., their distance is maximal, and thus the similarity should be small, but these sets have 99% of their elements in common, which intuitively gives similarity.

The Baire metric of two infinite sequences $r = a_1a_2\ldots$ and $s = b_1b_2\ldots$ over some set is defined as

$$d(r,s) = \begin{cases} 0 & \text{if } r = s \\ \frac{1}{2^{\min\{n|a_n\neq b_n\}}} & \text{if } r \neq s \end{cases}$$

(see [9]). A unary regular set R of words can be represented as infinite sequences $r = a_1a_2\ldots$ over $\{0,1\}$ where $a_n = 1$ if and only if $a^n \in R$. As in the the case of the Bodnarchuk metric the sets A and B given above have a relatively large distance and a large similarity, which contradicts the intuition.

In the case of CURLs, the sequences r and s can be given in the form $r = u^\omega$ and $s = v^\omega$ where u and v have the same length, i.e., they are infinite powers of some finite sequences of the same length. Then we can define the scaled Hamming distance of r and s as the number of positions where u and v differ and divided it by the length of v (by the division we ensure that the value belongs to the unit interval). However, now the sets

$$A' = \{a^{100n+1} \mid n \geq 0\} \text{ and } B' = \{a^{100n+2} \mid n \geq 0\}$$

or equivalently, $u = 10^{99}$ and $v = 010^{98}$ have a small distance $\frac{1}{50}$, but no similarity because they have no common element.

Essentially, the same holds for the information distance, which is given by the length of the minimal program (in the sense of Kolmogorov complexity) which transforms u to v (see [1])

References

[1] C. M. Bennett, P. Gács, M. Li, P. M. B. Vitanyi and W. H. Zurek, Information distance. *IEEE Transactions on Information Theory* (1998) 1407–1423.

[2] R. Carrasco, Accurate Computation of the Relative Entropy between Stochastic Regular Grammars, RAIRO, *Theoretical Informatics and Applications* **31** (5) (1997) 437–444.
[3] R. Carrasco and J. Rico, A Similarity between Probabilistic Tree Languages: Application to XML Document Families, *Pattern Recognition* **36** (9) (2003) 2197–2199.
[4] J. Dassow, G. M. Martín and F. J. Vico, Dynamical systems based on regular sets. Submitted.
[5] M. Domaratzki, k. Ellul, J. Shallit and M. Wang, Non-Uniqueness and Radius of Cyclic Unary NFAs. *Intl. J. Found. Comp. Sci.* **16** (2005) 883–896.
[6] F. Gecseg and I. Peak, *Algebraic Theory of Automata*. Academiai Kiado, Budapest, 1972.
[7] G. Gregor, Probabilistic and Nondeterministic Unary Automata, *In: Proceedings of Mathematical Foundations of Computer Science.* Lecture Notes in Computer Science **2747**, Springer-Verlag, (2003) 460–469.
[8] T. Jiang, E. McDowell and B. Ravikumar, The Structure and Complexity of Minimal NFA's over a Unary Alphabet. *Intl. J. Found. Comp. Sci.* **2** (2) (1991) 163û-182.
[9] A. Levi, Basic Set Theory. Dover, 1979.
[10] F. Mera and G. Pighizzini, Complementing Unary Nondeterministic Automata, *Theor. Comput. Sci.* **330** (2) (2005) 349–360.
[11] G. Pighizzini, Unary Language Concatenation and Its State Complexity. *5th International Conference on Implementation and Application of Automata* **2088** (2001) 252–262.
[12] G. Pighizzini and J. Shallit, Unary Language Operations, State Complexity and Jacobsthal's Function, *Intl. J. Found. Comp. Sci.* **13** (2002) 145–159.
[13] C. J. van Rijsbergen, Information Retrieval. London, 1979.
[14] G. Rozenberg and A. Salomaa, Handbook of formal languages. Springer, 1997.
[15] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining. Addison-Wesley, 2006.
[16] Wikipedia articles on Jaccard coefficient and Sørensen coefficient.