

# 文字列類似度の汎用的尺度 General Mesurment for Similarity of strings

飯箸泰宏\*  
Yasuhiro IIHASHI

\*明治大学  
Meiji University

あらまし：現在、文字列の類似度を測る汎用的な尺度がない。一般的に文字列の類似度の計測には Levenshtein 距離とその用途別のバリエーションが使われているが、Levenshtein 距離の使いにくさから用途別の分化が進んで、共通の尺度を失っている。比較空間に共有されるキーテキストを用いて、オリジナリティと類似性に関する正規化された汎用的尺度を提案する。

キーワード：テキストの類似度、オリジナリティ、汎用的尺度、類似度の関数、計算例、遺伝子配列

## 1 はじめに

一般的に文字列の類似度の計測には Levenshtein 距離とその用途別のバリエーションが使われているが、Levenshtein 距離の使いにくさから用途別の分化が進んで、共通の尺度を失っている。比較空間に共有されるキーテキストを用いて、オリジナリティと類似性に関する正規化された汎用的尺度を提案する。

知的所有権の争いには、テキストの類似度、ソースの類似度が争われることが多い。将来にわたる公正な情報コミュニケーションにとって、根拠のある汎用的尺度は必要であると考えます。

本報告では、汎用的な尺度を与える関数を提案する。この提案は、処理速度の向上などの実用的な目的を指すものではなく、各種の計測法や尺度を較正する目的で使用されることを目指している。

今回は、ランダムテキスト列に関する考察をベースに、比較空間に共有されるキーテキストを用いて、オリジナリティと類似性に関する正規化された汎用的尺度を提案する。実際に2組のテキストを用いて、この尺度を用いた計算結果と Levenshtein 距離の値を比較する。

## 2 文字列類似度の汎用的尺度

テキスト間の相互類似度

$$rS_{a-b} = \frac{rS_{a/b} + rS_{b/a}}{2}$$

テキスト間の相対類似度

$$rS_{a/b} = \frac{S_{a/b}}{O_a}, rS_{b/a} = \frac{S_{b/a}}{O_b}$$

$$\begin{aligned} \text{類似度 } S_{a/b} &= \ln(I_{a/b}) = \sum_{i=1}^{m_a} k_i \ln(N) - \ln\left(\prod_{i=1}^{m_a} (t_a - k_i + 1)\right) \\ &= \sum_{i=1}^{m_a} (k_i \ln(N) - \ln(t_a - k_i + 1)) \end{aligned}$$

$$\begin{aligned} \text{類似度 } S_{b/a} &= \ln(I_{b/a}) = \sum_{i=1}^{m_b} k_i \ln(N) - \ln\left(\prod_{i=1}^{m_b} (t_b - k_i + 1)\right) \\ &= \sum_{i=1}^{m_b} (k_i \ln(N) - \ln(t_b - k_i + 1)) \end{aligned}$$

諸元は次のとおりである。

使用される文字種の数	$N$
テキストAのテキスト長	$t_a$
テキストBのテキスト長	$t_b$
キーテキストの数	$m_a, m_b$

それぞれのキーテキストのテキスト長

A:  $k_1, k_2, k_3, \dots, k_i, \dots, k_{m_a}$

B:  $k_1, k_2, k_3, \dots, k_i, \dots, k_{m_b}$

とする。

## 3 Levenshtein 距離の特徴

Levenshtein 距離は、現在、文字列の類似度を測る方法として最も多用されている手法である。



情報コミュニケーション学会第2回全国大会  
CIS2005 (2005.3.30~3.31)

(2002.01.24)