情報コミュニケーション学会第 3 回全国大会
CIS2006（2006.2.25～2.26）

文字の出現確率に基づく第２類の文字列類似度
Similarity of strings -2nd kind- based on existence probability of character

飯箸泰宏*
Yasuhiro IIHASHI

*明治大学
Meiji University

あらまし：昨年、本学会で、ランダムテキストに基づく文字列類似度を発表し、「飯箸法」として一定の普及を見ることができた。著作権や工業所有権をめぐる争いはますます増加している。それらの争いの解決には、よりいっそう精緻な文字列類似度の計測法が必要である。今回は、ランダムに出現する文字ではなく、対象文字列空間に出現する文字の出現頻度に基づく正規化された新たな尺度を提案する。文字列類似度計測法の飯箸法第２類である。
キーワード：文字列、テキスト、ソースコード、類似度、文字の出現頻度、計算例、塩基配列

1　はじめに

昨年、本学会で、ランダムテキストに基づく文字列類似度の計測法を発表した[5]。今回は現実の文字列比較空間に存在する文字の出現確率を基にしたキーテキストを採用することによって、精緻化を図る。

昨年発表したランダムテキストに基づく文字列類似度の計測法[5]は、教育的効果が高いので、「(飯箸法)第１類」として残しておくことにする。

今回は、現実の文字列比較空間に存在する文字の出現確率を基にした、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（飯箸法第２類）を提案する。また、実際に 2 組のテキストを用いて、この新しい尺度（飯箸法第２類）を用いた計算結果と、昨年発表したランダムテキストに基づく文字列類似度の計測法[5]（飯箸法第１類）と Levenshtein 距離の 3 種類の値を比較する。

2　文字列類似度の汎用的尺度

テキスト間の相互類似度

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

テキスト間の相対類似度

$$rS_{\mathrm{a/b}} = \frac{S_{\mathrm{a/b}}}{O_{\mathrm{a}}}, rS_{\mathrm{b/a}} = \frac{S_{\mathrm{b/a}}}{O_{\mathrm{b}}}$$

対象文字列空間に出現する文字の出現頻度に基づく正規化された新たな尺度（飯箸法第２類）における類似度は次のとおりである。

$$\text{類似度}S_{\mathrm{a/b}} = \ln\left(I_{\mathrm{a/b}}\right) = \sum_{i=1}^{m_{\mathrm{a}}}\left(n_{\mathrm{k}_{\mathrm{a}i}} \times p_{\mathrm{k}_{\mathrm{a}i}}\right) = \sum_{i=1}^{m_{\mathrm{a}}}\left(n_{\mathrm{k}_{\mathrm{a}i}} \times \sum_{j=1}^{k_{\mathrm{a}i}} m_{\mathrm{c}_{\mathrm{a}ij}} \times p_{\mathrm{c}_{\mathrm{a}ij}}\right)$$

$$\text{類似度}S_{\mathrm{b/a}} = \ln\left(I_{\mathrm{b/a}}\right) = \sum_{i=1}^{m_{\mathrm{b}}}\left(n_{\mathrm{k}_{\mathrm{b}i}} \times p_{\mathrm{k}_{\mathrm{b}i}}\right) = \sum_{i=1}^{m_{\mathrm{b}}}\left(n_{\mathrm{k}_{\mathrm{b}i}} \times \sum_{j=1}^{k_{\mathrm{b}i}} m_{\mathrm{c}_{\mathrm{b}ij}} \times p_{\mathrm{c}_{\mathrm{b}ij}}\right)$$

$$\text{オリジナル度}O_{\mathrm{a}} = \ln\left(\prod_{i=1}^{t_{\mathrm{a}}} p_i\right) = \sum_{i=1}^{t_{\mathrm{a}}} \ln\left(p_i\right)$$

$$\text{オリジナル度}O_{\mathrm{b}} = \ln\left(\prod_{i=1}^{t_{\mathrm{b}}} p_i\right) = \sum_{i=1}^{t_{\mathrm{b}}} \ln\left(p_i\right)$$

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　$m_{\mathrm{a}}, m_{\mathrm{b}}$

それぞれのキーテキストのテキスト長

$$A: k_1, k_2, k_3, \cdots, k_{\mathrm{i}}, \cdots, k_{\mathrm{m_a}}$$

$$B: k_1, k_2, k_3, \cdots, k_{\mathrm{i}}, \cdots, k_{\mathrm{m_b}}$$

それぞれのキーテキスト$(\mathrm{k}_{ai}, \mathrm{k}_{bi})$に含まれる文字$\mathrm{c}_{aij}, \mathrm{c}_{bij}$ $(i = 1 \text{ to } \mathrm{k}_i)$の出現確率を$p_{\mathrm{c}_{aij}}, p_{\mathrm{c}_{bij}}$とする。

3 飯箸法第２類、飯箸法第１類[5]、飯箸法第１類と Levenshtein 距離の比較

表１ 尺度の性格比較

	飯箸法 第２類	飯箸法 第１類	Levenshtein 距離
類似性	◎	○	×
距離	×	△	○
独立性	◎	○	△

これらの違いに踏まえて次の３つ例で計算結果を示す。

（例１）test と street

（例２）aaaaa! To be or not to be. That is question と xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx aaaaa!

（例３）「田子の浦ゆ うち出でて見れば 真白にそ富士の高嶺に 雪は降りける」（万葉集）[6]と「田子の浦に うち出でて見れば 白妙の 富士の高嶺に 雪は降りつつ」（新古今集）[7]

表２ 計算（例 1）

	飯箸法 第２類	飯箸法 第１類	Levenshtein 距離
相互類似性	86 %	25 %	
距離			3
相互独立性	14 %	75 %	

表３ 計算（例 2）

	飯箸法 第２類	飯箸法 第１類	Levenshtein 距離
相互類似性	12 %	16 %	
距離			74
相互独立性	88 %	84 %	

表４ 計算（例 3）

	飯箸法 第２類	飯箸法 第１類	Levenshtein 距離
相互類似性	78 %	71 %	
距離			7
相互独立性	22 %	29 %	

5 結論

本報告で示した文字列の類似度に関する汎用的な尺度（第２類）は、先に発表した第１類を超えてテキストの類似度、遺伝子配列における類似度などにおいて、より概括的普遍的な尺度を与えるものと期待できるものである。

参考文献（口頭発表含む）

[1] A Method for the Correction of Garbled Words, Based on the Levenshtein Metric; K. Okuda, E. Tanaka, T. Kasai, IEEE Transactions on Computers, Vol. c-25, No. 2, February 1976

[2] 飯箸泰宏,"テキスト類似度の研究",カオス研究会,研究発表,1996.3.9

[3]飯箸泰宏,"テキストの類似度の計測法",ＳＨ情報文化研究会,研究発表,1999.2.27

[4] 井上克郎、松下誠、山本哲男,”類似度計測システム”,独立行政法人 科学技術振興機構,特許コード P03P000860,出願日 平成 14 年 1 月 24 日（2002.01.24）

[5]飯箸泰宏、”文字列類似度の汎用的尺度”、情報コミュニケーション学会第 2 回全国大会発表論文集、pp.53-54、2005 年

[6] 山部赤人、「山部宿祢赤人望不盡山歌一首 并短歌」の反歌、『萬葉集』巻三の三一八、推定 710-733 年

[7]紀貫之、本歌取り、新古今和歌集、推定 1200-1250 年