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）

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る。

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目指すものではなく、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

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

テキスト間の相対類似度

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

$$\text{類似度} S_{\mathrm{a/b}} = \ln(I_{\mathrm{a/b}}) = \sum_{i=1}^{m_{\mathrm{a}}} k_i \ln(N) - \ln\left(\prod_{i=1}^{m_{\mathrm{a}}}(t_{\mathrm{a}} - k_i + 1)\right)$$
$$= \sum_{i=1}^{m_{\mathrm{a}}}\left(k_i \ln(N) - \ln(t_{\mathrm{a}} - k_i + 1)\right)$$

$$\text{類似度} S_{\mathrm{b/a}} = \ln(I_{\mathrm{b/a}}) = \sum_{i=1}^{m_{\mathrm{b}}} k_i \ln(N) - \ln\left(\prod_{i=1}^{m_{\mathrm{b}}}(t_{\mathrm{b}} - k_i + 1)\right)$$
$$= \sum_{i=1}^{m_{\mathrm{b}}}\left(k_i \ln(N) - \ln(t_{\mathrm{b}} - k_i + 1)\right)$$

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

$\mathrm{A}: k_1, k_2, k_3, \cdots, k_{\mathrm{i}}, \cdots, k_{m_{\mathrm{a}}}$
$\mathrm{B}: k_1, k_2, k_3, \cdots, k_{\mathrm{i}}, \cdots, k_{m_{\mathrm{b}}}$

とする。

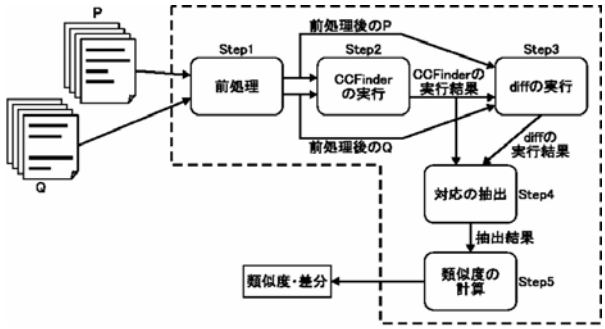
3 Levenshtein 距離の特徴

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。

2 つの文字列から別の文字列を作るために、要素の文字を最小何回「挿入」「削除」「入替え」する必要があるかを示す数を距離とみなしている。

考え方は単純でわかりやすいが、文字列が長くなればそれだけで Levenshtein 距離が大きくなってしまう傾向があり類似性の高い部分が相当部分を占めていても距離が大きい(類似性が低い)と判定されてしまう。そのため、類似性の高い部分（クローン部分）をあらかじめ特定しておいて、その部分のみの Levenshtein 距離を求めることが良く行われる。すなわち、前処理、後処理を加えて、Levenshtein 距離の利点を生かす工夫である。特定分野での利便性は向上するが、尺度としての汎用性は犠牲になっている。

例として、井上らの「類似度計測システム」(特許コード P03P000860、出願日 平成 14 年 1 月 24 日（2002.01.24)、独立行政法人 科学技術振興機構、井上 克郎、松下 誠、山本 哲男）を取り上げる。ここにも、類似性の高い部分を抽出してから比較する方法が提案されている。



4 本手法と Levenshtein 距離の比較

本手法と **Levenshtein 距離には、おのずと性格に違いがある。**

表 1 尺度の性格比較

	飯箸の汎用的尺度	**Levenshtein 距離**
類似性	○	×
距離	△	○
独立性	○	△

これらの違いに踏まえて次の 2 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 xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx aaaaa!

表 2 計算例（1）

	飯箸の汎用的尺度	**Levenshtein 距離**
相互類似性	25 %	
距離		3
相互独立性	75 %	

表 3 計算例（2）

	飯箸の汎用的尺度	**Levenshtein 距離**
相互類似性	16 %	
距離		74
相互独立性	84 %	

5 結論

計算例にも見られるように、例 1 に比べて例 2 は Levenshtein 距離では大きく離れた存在であるかのように示されるが、飯箸の汎用的な汎用的尺度によれば、類似性も独立性も適度に認めていることが示されている。

本報告で示した文字列の類似度に関する汎用的な尺度は、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、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。

ただし、計算適用例はまだわずかなため、さらに事例を増やして妥当性を検討すべきであると思われる。

なお、キーテキストに重み付けする場合などの研究は未発表であり、準備中である。

6 参考文献（既発表含む）

1)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.

2)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.

3)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

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）