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　発表資料
2005.03.31

飯箸泰宏
明治大学･大正大学･法政大学･慶應義塾大学非常勤講師
株式会社サイエンスハウス代表取締役

目次

0．なぜ、今、類似度の汎用的尺度か

0-1.公正な情報コミュニケーション

- 汎用的な尺度がない。

  現状は、用途別のバリエーションに向かっている。
  共通の尺度を失っている。
- 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。
- 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が必要であると考える。

☆私事、編集者歴35年、システム技術者歴25年、教員歴23年です。

0-2.悪い例

- 悪い例1

  ネット上のコピペ乱用。

  ネットのコピペで作られる学生のレポート
- 悪い例2

  学生間のレポートの不正コピー
- 悪い例3

  出版による作品盗用、新聞コラムの盗用記事

0-3.応用可能性

- 応用可能性1

  アミノ酸配列･アラインメントの研究で、蛋白質やDNAの類似度を数値化できる。

- 応用可能性2

  学生のレポートの類似度を数値化できる。

- 応用可能性3

  作品盗用、盗用記事の疑惑を数値化できる。

- 応用可能性4

  各種手法による類似判定の結果を同一の尺度で比較できる。

0-4.字面の類似性

- 一般に、次のような尺度が考えられる。
  - ①任意長のテキスト文書が、他のテキスト文書と「意味」においてどの程度近いか。
  - ②任意長のテキスト文書が、他のテキスト文書と「文体」がどの程度近いか。
  - ③任意長のテキスト文書が、他のテキスト文書と「字面」においてどの程度近いか。
- 本報告は、③「字面」の類似性を図る尺度の要請に応えるものである。

1. これまでの発表と今回の発表

- 1996年3月9日　“ﾃｷｽﾄ類似度の研究”,カオス研究会
- 1999年2月27日　“テキストの類似度の計測法”,ＳＨ情報文化研究会

今回の発表では、よく使用されている類例の尺度との比較を中心に汎用性を主張する。

↑

計算の根拠の詳細については、過去の発表に譲る。

2．類例の尺度--Levenshtein距離

2-1．Levenshtein距離の例

- **例**1

test

vs

street

→Levenshtein**距離**=3

- **例**2

aaaaa! To be or not to be. That is question!

vs

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx aaaaa!

→ Levenshtein**距離**=39

2-2.問題点

- **長いストリングの全体にわたって、使用すると、類似度が正しく反映しない。**

(例1)

aaaaa! To be or not to be. That is question!

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

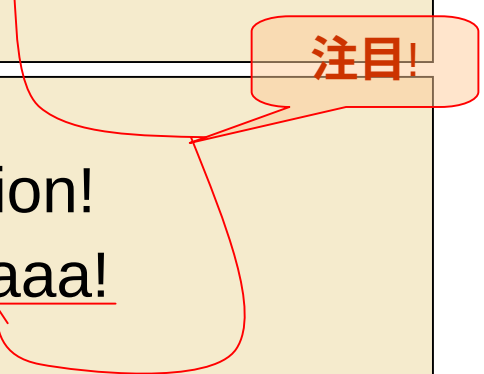
この場合、Levenshtein**距離**=39

(例2)

aaaaa! To be or not to be. That is question!

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx aaaaa!

この場合も、Levenshtein**距離**=39 !



“aaaaa!”の共通性が評価に反映しにくい。

2-3.解決策について

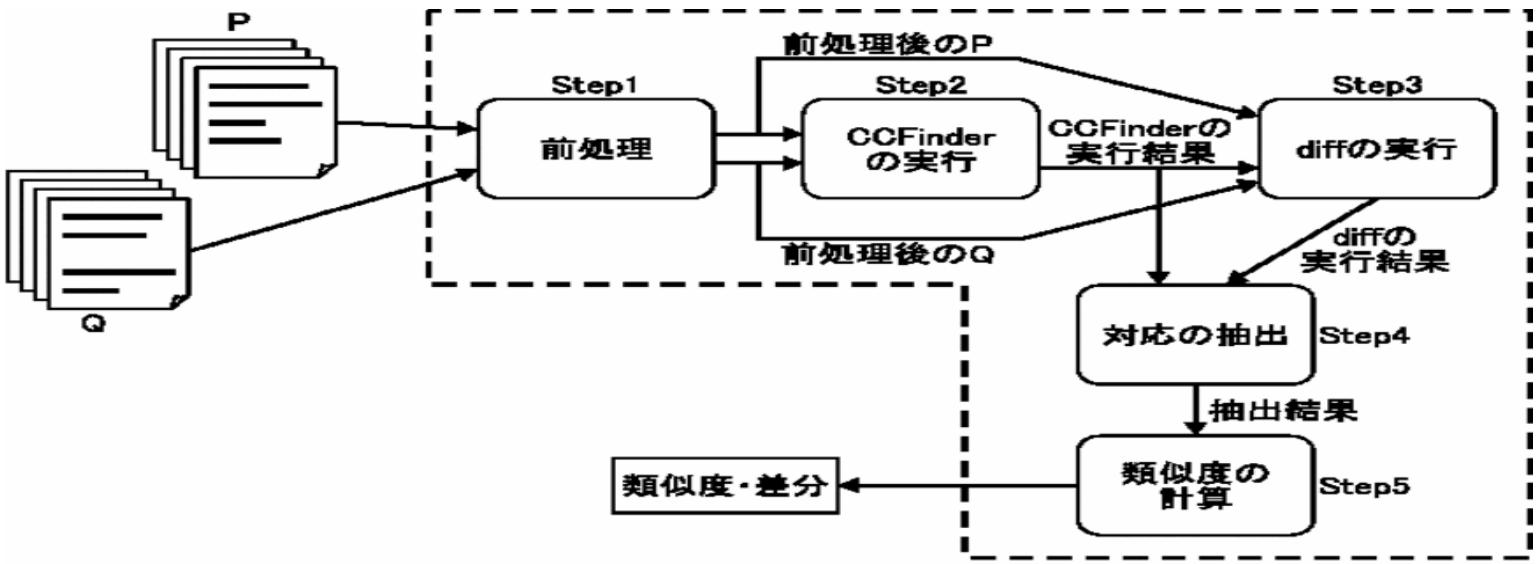
- **類似性の高い部分（クローン部分）をあらかじめ特定しておいて、その部分のみの**Levenshtein**距離を求める。すなわち、前処理、後処理を加えて、**Levenshtein**距離の利点を生かす工夫である。特定分野での利便性は向上するが、尺度としての汎用性は犠牲になっている。**

2-4. Levenshtein距離の特徴

- 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。
- 応用分野としては、文字列あいまい検索、遺伝子配列の類似度など。
- 長いストリングの一部に着目して使用することが多い。

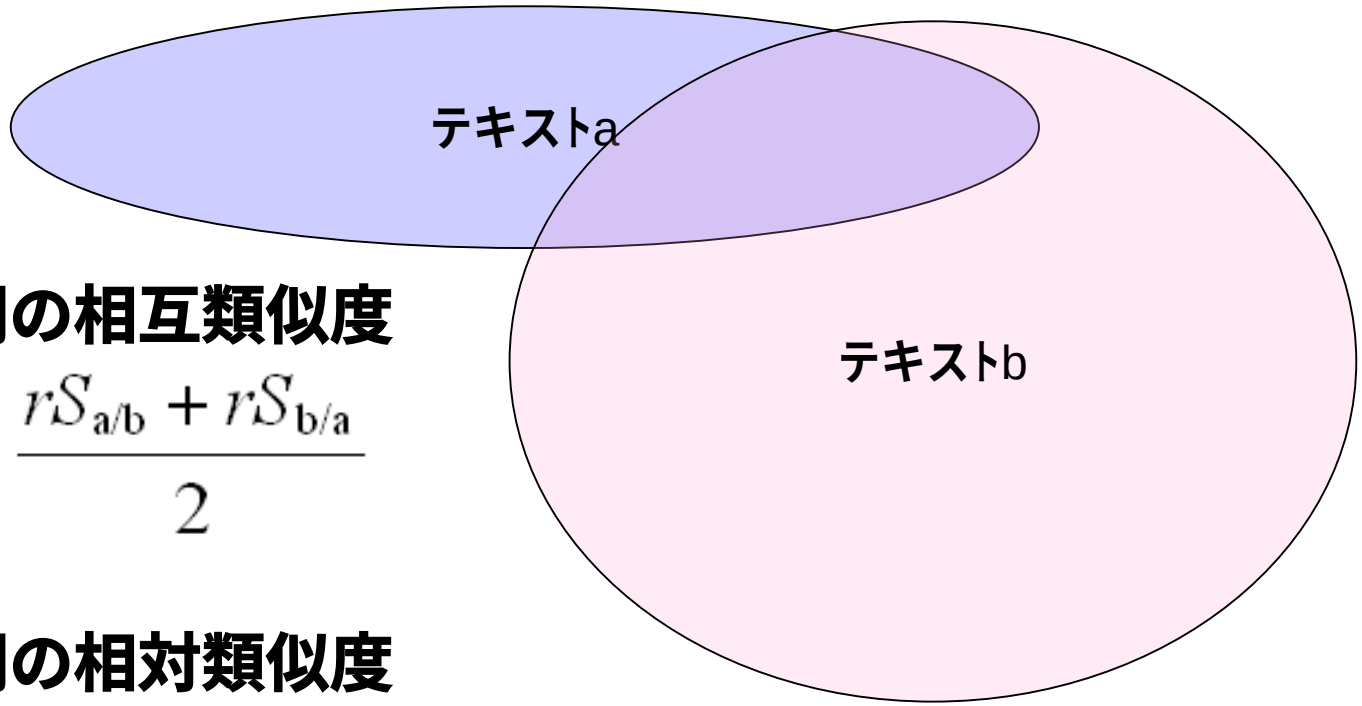
2-5.解決策の例

- **（例）「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



3．汎用的尺度―提案

3-1.テキスト間の相互類似度



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

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

3-3.テキスト間の相対類似度とオリジナル度

→ 14

テキスト間の類似度

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

$$= \sum_{i=1}^{m_{\text{a}}} \left(k_i \ln(N) - \ln\left(t_{\text{a}} - k_i + 1\right)\right)$$

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

$$= \sum_{i=1}^{m_{\text{b}}} \left(k_i \ln(N) - \ln\left(t_{\text{b}} - k_i + 1\right)\right)$$

テキスト間のオリジナル度

$$O_{\text{a}} = \ln\left(L_{\text{a}}\right) = t_{\text{a}} \ln(N)$$

$$O_{\text{b}} = \ln\left(L_{\text{b}}\right) = t_{\text{b}} \ln(N)$$

3-4.緒元

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

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

とする。

3-5.定義

定義1：「オリジナリィの強度」と「類似の強度」

テキストの「オリジナリィの強度」とは、他のテキストと区別される程度の強さである。したがって、あるテキストが存在するとき、偶然に成立する確率の逆数を、ここではテキストの「オリジナリィの強度」と定義する。

「類似の強度」とは、2つのテキストの中に共通する文字列（単語とは限らないことに注意）が生ずる確率の逆数とする。共通する文字列を同定する方法は、最長一致を原則としてここでは採用した。共通する文字列をキーテキストと称する。

定義2：「オリジナル度」と「類似度」

「オリジナル度」　＝ln(オリジナリィの強度)
「類似度」　＝ln(類似の強度)

定義3：「相対オリジナル度」と「類似度」

「相対オリジナル度」＝（「オリジナル度」－「類似度」）／「オリジナル度」
「相対類似度」　＝「類似度」／「オリジナル度」

定義4：「相互類似度」

「相互類似度」　＝（「類似度a/b」＋「類似度b/a」）／ 2

3-6.計算例 (1)

（例） テキストa test　　cf. Levenshtein距離=3
テキストb street

ここでは、便宜的に使用されている文字は、s,t,r,eの5文字とする。アルファベットのすべての文字を採用しても良いが、例1としては煩雑なので、5文字にとどめる。
共通するキーテキストは、最長一致を原則とすると、st,e,r,tの4つである。
これらは、テキストa test　には、 st,e,t　の3つ、
テキストb street　には、 st,e,r,tの4つが使用されている。
1)「テキストa のオリジナル度」= 4 ln(5) = 6.437752・・・
「テキストb のオリジナル度」= 6 ln(5) = 9.656627475・・・
「テキストa の類似度」= 2.566550639・・・
「テキストb の類似度」= 1.062473242・・・
2)「テキストa の相対類似度」= 0.398671893・・・
「テキストb の相対類似度」= 0.11002529・・・
「テキストa の相対オリジナル度」= 0.601328107・・・
「テキストb の相対オリジナル度」= 0.88997471・・・
3)「テキストaとテキストb の相互類似度」= 0.254348592・・・
「テキストaとテキストb の相互オリジナル度」= 0.745651408・・・

「両者は相互に25%の類似度である」
「両者は相互に75%のオリジナル度である」

3-6.計算例 (2)

(例)　テキストa: aaaaa! To be or not to be. That is question!
　　テキストb: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx aaaaa!
cf. Levenshtein距離=39

ここでは、便宜的に使用されている文字数を255文字とする。
テキストの長さはいずれも44文字である。共通するキーテキストは、最長一致を原則とすると、aaaaa!," "(space), の2個である。
これらは、テキストa には、 aaaaa! 1個と" " (space) 9個、テキストbには、 aaaaa! 1個と" " (space) 1個が使用されている。

1)「テキストa のオリジナル度」= 243.815596･･･
　「テキストb のオリジナル度」= 243.815596･･･
　「テキストa の類似度」= 45.39768483･･･
　「テキストb の類似度」= 31.34109354･･･
2)「テキストa の相対類似度」= 0.186196804･･･
　「テキストb の相対類似度」= 0.128544253･･･
　「テキストa の相対オリジナル度」= 0.813803196･･･
　「テキストb の相対オリジナル度」= 0.871455747･･･
3)「テキストaとテキストb の相互類似度」= 0.157370528･･･
　「テキストaとテキストb の相互オリジナル度」= 0.842629472･･･

「両者は相互に16%の類似度である」
「両者は相互に84%のオリジナル度である」

4．今後の展開

- 残された課題

1)計算適用例を増やして、人間的な類似度感覚との差異を調査する。

2)各種応用場面での普及手法と本手法を比較して、汎用性を確かめる。

- 今後の研究の展望

1)分野別キーテキスト群、日本語汎用キーテキスト群、などを選定する。

2)キーテキストに重み付けをするケースについての研究を発展させたい。

- 楽しみ

1)学生の課題レポートや課題プログラムのソースがどの程度「相互類似性」を持っているか、計算してみたい。

5．参考文献（発表を含む）

- 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．
- 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．
- 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）

-終わり-