社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－DBS－136（2）
2005－FI－79（2）
2005／5／19

異なるデータ長に対応したデータ圧縮による類似データ同定

―近代日本文学データの著者推定による検証―

安形輝　　agata@asia-u.ac.jp
亜細亜大学国際関係学部

Benedetto らによる圧縮プログラムを用いた類似データの同定手法は高い精度を出しているにも関わらず、その後の研究が十分に行われてきたとはいえない。本研究では彼らの手法の欠点を改良した圧縮改善率からの推定手法を提案し、日本語データへの応用可能性を検証した。第一に、先行研究と同様にテキストを加工し固定長データを用いて、第二に、何も加工しない可変長データを用いて実験を行った。前者については 50 のテスト集合に対する平均成功率は Benedetto らの手法が 90.5%、本研究が提案する圧縮改善率からの手法が 97.7%となり、先行研究での最高値 96.0%を上回る結果が得られた。また、圧縮改善率による手法はデータが短い場合にも他の手法に比べて性能劣化がほとんど起こらないことが明らかとなった。後者の実験についても圧縮改善率による手法は 95.7%と高い成功率が得られた。

Measuring Similarity of Variable Length Data with Compression Program

- Authorship Attribution in Japanese Modern Literature -

Teru AGATA　　agata@asia-u.ac.jp
Asia University

Benedetto et al. recently proved the validity of a method of measuring similarity using with compression program. In spite of its potential, however, this method hasn't been applied to information science. The present study proposed a modified method using with the improvement ratio of compression and conducts two experiments in authorship attribution in Japanese modern literature. (1) Comparing the results of the modified method and Benedetto's method applied to the same test collection. (2) Experiment in variable length data. Experiment (1) shows the average precision of the present method is 97.7%, while that of Benedetto's was 90.5%, and that the performance in the case of short data is far better than other methods. Experiment (2) proved the effectiveness of the present method with the good average precision rate: 95.7%.

1. はじめに

1.1 圧縮プログラムによる類似データ同定

Benedetto らは”Language Trees and Zipping”[1]において ZIP 系列の圧縮プログラムによる分類や類似データの同定手法を提案している。一見、奇抜な手法ではあるが、

(1)一般的なプログラムを利用するため導入コストが低く汎用性が高いこと、
(2)テキストデータだけでなく、機械可読データであれば画像データや DNA データなど種類に関わらずどのようなデータにも応用可能であること

などの利点を持っている。一般誌に掲載されるなど、この研究に対する注目は大きいと言えるが、掲載誌が物理学 Physical Review Letter 誌であることから、情報学分野での注目は少なく、応用もほとんど行われていない。

1.2 圧縮による類似データ同定の既往研究

Benedetto らは 90 文献[1]から構成されるコーパスに対して著者推計実験を行い、93.3%という高い精度の結果を残している。しかし、この実験環境に対する記述をほとんど公表しておらず、既往研究と同様の実験集合を用いてもないため、データの信頼性、汎用性、再現性については疑問が残る。

また、Kukushkina ら[2]は Benedetto らとほぼ同様の形で圧縮プログラムを応用したテキストの自動分類に関して研究を行っている。結果では使う圧縮プログラムによって大きく性能にばらつきがあるが、一部のプログラムを使った場合には、自然言語処理において定評のあるマルコフ連鎖を応用した手法よりも高い性能を示している。

内山[3]は Benedetto らの手法を 7 人の書き手による日本語学術論文 34 編の原著者推定を行い、高い精度での著者推定を行うことができるとしている。しかし、独自の小規模データを使用しており、既往研究との比較ができない。また、従来の著者推定手法との比較も行われていない。

1.3 著者推定に関する既往研究

著者推定に関しては、計量文体学を中心として、コンピュータの登場以前から様々な手法が提案され、継続的に研究がなされてきた比較的活発な研究領域といえる[4]。これは、著者推定や真贋判定[5]に対する社会的需要があったためと考えられる。研究対象データは、古文書[6]からウェブ[7]から調書までとさまざまであり、頑強性が高い著者推定手法は応用可能性が高くなることが示唆される。

しかし、どのような著者推定手法もテキストについて何らかの言語的、構造的、内容的な解析を必要としている。例えば、最も簡便とされる著者推定手法に、古典的かつ精度の高い平均文長を使った手法があるが、句読点や改行から文の長さを判定する必要がある。

一方で、圧縮プログラムを応用した手法ではテキストの言語、構造、内容を解析する必要がなく、単にファイルとしてプログラムに投入するのみであり、応用範囲は広いと考えられる。

1.3 研究目的

本研究では、圧縮プログラムを使った類似データの同定手法の検証を目的として、既往研究との比較実験を行った。日本語文献の著者推定に関する研究は、計量国語学などの領域で数多くなされているが、

① 既存の複数の手法の結果を残している
② 青空文庫から実験データが入手可能である

という理由から、松浦ら(2000)による「n-gram の分布を利用した近代日本語文の著者推定」[8]を比較の対象とした。この研究中で比較されている著者推定手法は、松浦らによる dissim、Tankard[9]による手法、基本データとしてのダイバージェンス手法である。

また、Benedetto らの手法を用いた準備実験から明らかとなった欠点を補う新たな圧縮プログラムを用いた類似データ同定手法を提案し、その検証も行った。

2. 類似データの同定の基本原理

2.1 Benedetto らの手法

Benedetto らの類似データ同定手法で用いられる LZ77 法[10]に代表されるデータ圧縮法は、原則的に頻出データをより短い他のデータに置き換えることによって、データ圧縮を行う。

[1] http://www.liberliber.it/

基準となるデータ X があり、LZ_X をデータ X を圧縮したときのファイル長としたとき、X に類似するデータは、候補となるすべてのデータについて、$LZ_{A_i+x} - LZ_{A_i}$ が最小になる A_i となるという考えに基づくものである。X と A_i に共通するデータ部分が多ければ多いほど、また、圧縮プログラムの共通データの検出率が高いほど、$LZ_{A_i+x} - LZ_{A_i}$ は小さくなると考えられ、このような類似データの同定手法が実際に成立することが理解できる。

2.2 圧縮改善率からの推定手法

Benedetto らの手法では、元テキストと候補テキストを連結した連結ペアの圧縮ファイル長と候補テキストの圧縮ファイル長の最小差を類似度として用いていた。しかし、予備的な実験からは、

(1)各テキストの単体での圧縮されやすさが連結ペアの圧縮ファイル長に大きく影響すること

(2)テキスト連結の順序が圧縮率に影響すること

の二点が明らかとなった。そこで、ペアとなるものの圧縮率から単体での圧縮率の影響とテキストの連結順序の影響を排除する目的で、以下の数式で表される圧縮改善率(CIR)を考案した[11]。

$$CIR = \left(\frac{\mathrm{LZ_X}}{\mathrm{L_X}} + \frac{\mathrm{LZ_{A_i}}}{\mathrm{L_{A_i}}} \right) - \left(\frac{\mathrm{LZ_{X+A_i}} + \mathrm{LZ_{A_i+X}}}{\mathrm{L_{X+A_i}}} \right) \quad (1)$$

ここで、L_X は X のファイル長を、LZ_X は X を圧縮した場合のファイル長を、$\mathrm{LZ_{X+A_i}}$ は X を先に、A_i を後として連結した場合の圧縮ファイル長を、$\mathrm{LZ_{A_i+X}}$ は逆に連結した場合の圧縮ファイル長を示している。式(1)は、前半が各データ単体での圧縮されやすさを、後半が連結ペアを作った場合の圧縮されやすさを表現しており、全体としてペアを作ったことでどの程度圧縮率が上がったかを表している。後半部で $\mathrm{LZ_{X+A_i}}$ と $\mathrm{LZ_{A_i+X}}$ の二つを算出する理由は、圧縮プログラムの辞書作成アルゴリズムを考慮した場合に、二つのテキストをどの順序で投入するかが与える影響を排除するためである。

この式(1)を、データの長さが異なった場合に、各データの圧縮率が最終的な圧縮率にどの程度寄与するかを考慮に入れて改良したものが、式(2)である。

$$CIR = \left(\frac{\mathrm{LZ_X}}{\mathrm{L_X}} \cdot \frac{\mathrm{L_X}}{\mathrm{L_{X+A_i}}} + \frac{\mathrm{LZ_{A_i}}}{\mathrm{L_{A_i}}} \cdot \frac{\mathrm{L_{A_i}}}{\mathrm{L_{X+A_i}}} \right) - \left(\frac{\mathrm{LZ_{X+A_i}} + \mathrm{LZ_{A_i+X}}}{2 \cdot \mathrm{L_{X+A_i}}} \right)$$
$$= \frac{2\left(\mathrm{LZ_X} + \mathrm{LZ_{A_i}}\right) - \left(\mathrm{LZ_{X+A_i}} + \mathrm{LZ_{A_i+X}}\right)}{2 \cdot \mathrm{L_{X+A_i}}} \quad (2)$$

以下の実験ではこの式(2)を採用している。

圧縮改善率を使い、データペアの類似度を測定する場合、各データの単体での場合、と比較した圧縮率が改善されたペアから順に候補テキストを並べることとなる。

3. 実験環境

3.1 実験テキスト

実験対象テキスト集合の作成は、松浦らの研究とまったく同じ手順によって行った。ただし、ここで作成されたテキスト集合セットは、手順の一部に無作為な選択を含む部分があるため、まったく同じテキスト集合ではなく、ほぼ同じ性質を有すると考えられるテキスト集合となる。

実験テキストの対象とするのは、著作権の切れた作品のデジタル化を行っている青空文庫[2]から入手した、岡本綺堂、芥川龍之介、梶井基次郎、菊池寛、国木田独歩、水野仙子、樋口一葉、有島武郎の 8 人の近代日本文学者による

[2] http://www.aozora.gr.jp/

92 作品である。これらの作品は明治から昭和初期にかけて執筆された作品であり、歴史的仮名遣いで書かれたものと現代仮名遣いに改めた作品が混在しているが、先行研究と同様に手法の頑強性をも検証するため、あえて統一はしていない。使用した全 92 作品は、72 本の小説、9 本のエッセイ、5 本の書簡形式文章、3 本の戯曲、2 本の日記、1 本の談話から構成される（表１）。

表1 実験集合リスト

著者名	タイトル	著者名	タイトル
岡本綺堂	化け猫杏	菊池寛	恩讐の彼方に
岡本綺堂	弁天娘	菊池寛	勝負事
岡本綺堂	菊人形の昔	菊池寛	出世
岡本綺堂	狐と僧	菊池寛	忠直卿行状記
岡本綺堂	帯取りの池	菊池寛	父帰る
岡本綺堂	お照の父	菊池寛	藤十郎の恋
岡本綺堂	津の国屋	菊池寛	若殿草紙
岡本綺堂	柳原堤の女	菊池寛	ゼラール中尉
岡本綺堂	幽霊の観世物	国木田独歩	源おじ
芥川龍之介	あばばばば	国木田独歩	牛肉と馬鈴薯
芥川龍之介	アグニの神	国木田独歩	非凡なる凡人
芥川龍之介	秋	国木田独歩	恋を恋する人
芥川龍之介	あの頃の自分の事	国木田独歩	武蔵野
芥川龍之介	或阿呆の一生	国木田独歩	忘れ得ぬ人々
芥川龍之介	或敵討の話	国木田独歩	酒中日記
芥川龍之介	或旧友へ送る手記	国木田独歩	たき火
芥川龍之介	或日の大石内蔵助	国木田独歩	運命論者
芥川龍之介	浅草公園―或シナリオ―	水野仙子	響
芥川龍之介	一塊の土	水野仙子	輝ける朝
梶井基次郎	愛撫	水野仙子	神樂坂の半襟
梶井基次郎	ある崖上の感情	水野仙子	道―ある妻の手紙―
梶井基次郎	ある心の風景	水野仙子	女
梶井基次郎	泥濘	水野仙子	四十餘日
梶井基次郎	冬の蠅	水野仙子	嘘をつく日
梶井基次郎	冬の日	樋口一葉	十三夜
梶井基次郎	筧の話	樋口一葉	にごりえ
梶井基次郎	過古	樋口一葉	大つごもり
梶井基次郎	器楽的幻覚	樋口一葉	たけくらべ
梶井基次郎	Kの昇天―或はKの溺死	樋口一葉	うつせみ
梶井基次郎	交尾	樋口一葉	わかれ道
梶井基次郎	檸檬	樋口一葉	ゆく雲
梶井基次郎	のんきな患者	有島武郎	小さき者へ
梶井基次郎	路上	有島武郎	二つの道
梶井基次郎	桜の樹の下には	有島武郎	片信
梶井基次郎	雪後	有島武郎	卑怯者
梶井基次郎	城のある町にて	有島武郎	広津氏に答う
梶井基次郎	蒼穹	有島武郎	一房の葡萄
梶井基次郎	闇の絵巻	有島武郎	小作人への告別
梶井基次郎	橡の花―或る私信―	有島武郎	水野仙子氏の作品について
菊池寛	青木の出京	有島武郎	溺
菊池寛	入れ札	有島武郎	宣言一つ
菊池寛	勲章を貰う話	有島武郎	想片
菊池寛	身投げ救助業	有島武郎	私の父と母
菊池寛	M侯爵と写真師	有島武郎	火事とポチ
菊池寛	無名作家の日記	國木田獨歩	少年の悲哀
菊池寛	大島が出来る話	國木田獨歩	石清虚

3.2 推計実験の評価尺度

著者推定実験の評価は、先行研究と同様の手順で行った。あるデータと集合内の他のデータを比較し、同じ著者の他のデータを見つけられたときに著者推定に成功したものとし、全推定試行数に対して、著者推定の成功数の割合を算出している。これを平均成功率と呼び、式で表すと以下のとおりである。

$$\text{平均成功率} = \frac{\text{成功例数}}{\text{全推定数}}(\%) \qquad (3)$$

3.3 圧縮プログラム

Benedetto らや圧縮改善率による著者推定手法では、圧縮率が高い圧縮プログラムほど、より共通する部分を発見できることになる。そのため、複数の圧縮形式について考慮すべきであるが、今回は Microsoft 社の Windows に標準添付され、世界的に最も普及していると考えられる ZIP 形式を採用した。実際の圧縮には、ZLIB 圧縮ライブラリを最も圧縮率の高いパラメータを設定し使用した。

4. 既往研究との比較実験

4.1 実験集合の作成

松浦らの研究と同様の実験を行うために、青空文庫からの 92 作品を基にして 50 のテスト集合を作成した。各テスト集合の作成手順は以下のとおりである。

(1) 92 作品プールから擬似乱数(Mersenne Twister 法[3])によって作品を一つ選択する。
(2) 作品中のテキストが 30,000 文字よりも多い場合、先頭の 30,000 文字を取り出し、テスト集合に追加する。該当作品を全作品プールから削除する。
(3) ３万文字よりも少ない場合、30,000 文字未満の同じ著者の作品群から一つずつ作品

[3] きわめて長い周期 2^{19937}-1 を持ち、623 次元

を選択し選んだ順に連結する。30,000 文字を超えた時点で、テキストの先頭 30,000 文字を一つの実験テキストとし、テスト集合に追加する。連結したすべての作品を全作品プールから削除する。

(4) 92 作品プールに作品が残っている場合、(1)に戻る。

(5) テスト集合に複数の作品が登録されなかった著者の場合は、著者推定が不可能となるためその作品を除去する。また、著者による偏りをなくすために、一著者あたりの最大実験テキスト数を5とする。

このような手順で作成された 50 のテスト集合の総計は表2の通りである。松浦らのデータと同様の手順で作成したにも関わらず、特に水野仙子の値が異なっている。無作為抽出がデータ作成手順に含まれるため、10 回データ集合を作成し、先行研究と同様の分布になるかを試行した。しかし、近い分布の集合群は作成されなかった。データ集合の特性に差異が見られた要因としては、

(1) 青空文庫のデータに対して 1999 年時点から修正が加えられこと

(2) 無作為抽出のための擬似乱数として Mersenne Twister 法を用いていること

が考えられるが、既往研究のデータが公開されておらず、これ以上の分析は行うことができない。

結果として、データ集合の特性に違いは出てしまったが、松浦らのデータに比べテスト集合に含まれる平均著者数が増加しており、著者推定の精度からはより厳しい条件となったといえる。例えば、著者「水野仙子」のテキストデータは松浦らのデータでは半数以下の集合にのみ含まれるが、今回の集合には 2/3 以上のデータに含まれている。

なお、各テキストについて、本文以外の著者や執筆年月日などの書誌事項は除去している。また、先行研究と同様に、(1)改行、空白は原則として一文字としているが、改行後の空白は冗長であるため除去し、(2)半角英数記号は全角に変換している。結果として、すべてのデータは 30,000 文字であり、日本語 1 文字が 2 バイトであるため 60,000 バイトとなった。

表2 テスト集合の総計

著者名	50 セット中の合計	松浦ら (2000)
岡本綺堂	218	203
芥川龍之介	100	100
梶井基次郎	170	160
菊池寛	241	222
国木田独歩	147	129
水野仙子	88	48
樋口一葉	100	84
有島武郎	100	98
総計	1,164	1,044

4.2 固定長データに対する実験

固定長 60,000 バイトデータを使って、著者推定を行った実験の結果は表 3 のようになった。表における松浦らの平均成功率は n-gram の値を変化させた場合の最高値となっている。

圧縮改善率からの手法は 97.68%（1164 試行中の 1137 回成功）という非常に高い精度を得ることができた。また、50 セット中の 29 セットではすべての試行において正解著者を同定しており、ほぼ完璧に成功しているといえる。Benedetto らの手法でも 90.46%という高精度を得られているが、圧縮改善率からの手法には及ばない。また、すべての試行が成功したセットは 50 セット中の 2 セットのみであった。

以下で一様に分布する擬似乱数

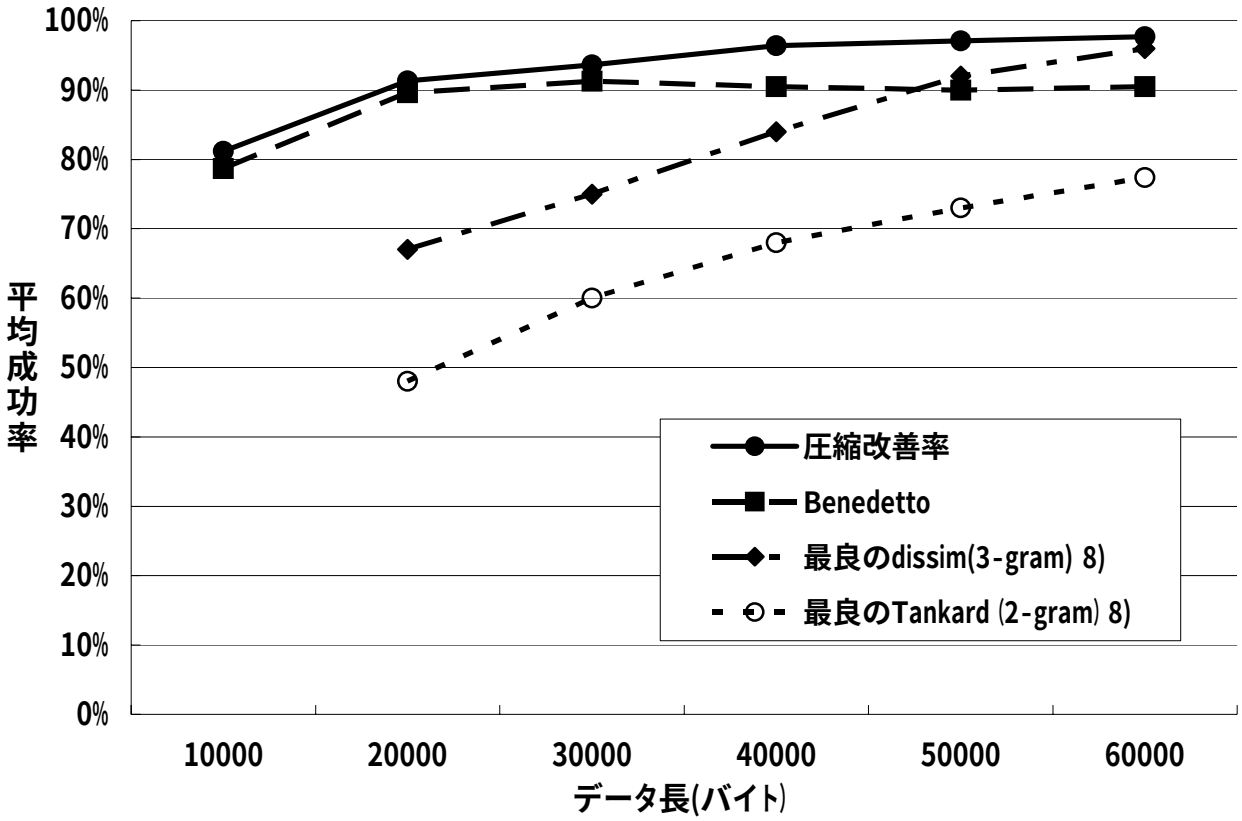

図 1 データ長と平均成功率

さらに松浦らの研究結果との比較において、今回のテスト集合は松浦らの実験時のデータよりもデータ数・著者数が多いため、判定精度からは不利と考えられるにも関わらず、判定精度が大幅に上昇していることを強調しておきたい。

表3 平均成功率

推計手法			平均成功率
松浦ら (2000)*	dissim	3-gram	96.00%
	Takardの手法	2-gram	77.40%
	ダイバージェンス	1-gram	52.50%
Benedettoら			90.46%
圧縮改善率			97.68%

4.3 データ長を変化させた場合の性能劣化

4.2 では 30,000 文字／60,000 バイトという固定長データを用いて実験を行ってきたが、書簡などの短いテキストに関する著者推定の場合、いつも十分な長さのデータが得られるとも限らない。ここでは、各データを短くしていった場合、つまり、手がかりがより少なくなった場合に、どの程度の性能劣化が起こるかを分析している。

実験手順としては 3.3 までと同様の 50 セットから構成される実験集合を使ったが、その際、各データを 10,000 バイトから 60,000 バイトまで 10,000 ずつ変化させた場合の平均成功率を算出した。実験結果は、成功率を縦軸に、データ長を横軸にとったグラフでは図１のようになった。このグラフからは、dissim や Tankard の手法はテキストデータが少なくなるにつれ、性能が徐々に落ちていることがわかるが、Benedetto らや圧縮改善率の手法による推定の場合、10,000 バイト程度のデータであっても 9 割を超える成功率を得られていることがわかる。また、二手法については性能の落ち込みが極めて緩やかであるため、既往研究で

表4 オリジナルデータに対する著者推定

		Benedetto		圧縮改善率		全試行数
		成功数	成功率	成功数	成功率	
データ長（バイト）	60000	79	85.9%	88	95.7%	92
	50000	79	85.9%	88	95.7%	92
	40000	81	88.0%	88	95.7%	92
	30000	80	87.0%	88	95.7%	92
	20000	87	94.6%	88	95.7%	92
	10000	86	93.5%	88	95.7%	92

は行われなかった 5,000 バイトというきわめて短いデータを使っての著者推定も行ったが、ここでも圧縮改善率は 80%を超える成功率となっている。また、Benedetto の手法は 10,000 から 30,000 バイトの範囲においてほんのわずかに低い性能を見せている。

5. オリジナルデータを対象とした著者推定

5.1 実験集合

既往研究との比較ではどのデータのサイズも同じとし、一部のデータは複数の文献を連結するという操作をした実験集合を対象とした実験であった。ここでは、オリジナルの 92 作品に対して何ら操作を加えることなく用いて、著者推定実験を行った。

92 作品のうち、最もサイズが小さいものは梶井基次郎の「過古」で、書誌事項を除いたテキスト部分は 3,184 バイトであり、最もサイズが大きいのは岡本綺堂の「半七捕物帳 津の国屋」で、66,928 バイトである。

5.2 実験結果

固定長データの実験と同様に 92 作品のうち１作品選択し、それと同じ著者のデータを見つけられた場合には成功とする形式で実験を行った。さらに、各データの先頭から何バイトまでを圧縮処理の対象とするかについて 10,000 バイト単位で変化させることも行った。圧縮改善率による手法はデータ長に関係なくすべての場合において 95.7%という高い精度を示した。また、Benedetto らによる手法もおおむね 80%以上の成功率であったが、データ長と関係では手がかりとなるデータ長を短くするほど性能が高くなるという不可解な現象がみられた。

6. 結論

ZIP プログラムを用いた圧縮改善率からの著者推定法は、非常に簡便な手法であるにも関わらず、既往研究とほぼ同様の実験環境下で、従来の著者推定手法よりも高い性能が得られた。また、手がかりとなるデータが小さな場合でも他の手法に比べ、著しく性能劣化が少ないことも明らかとなった。また、青空文庫のデータを加工せずに行った実験でも 95.7%と高い成功率を得ており、実データについても十分に応用可能性は高いと考えられる。

今後は、ZIP 以外の他の圧縮形式についてもどの程度の性能が得られるか実験していくとともに、テキスト以外のデータへの応用、非可逆圧縮手法のアルゴリズムについても検討していく。

【注・引用文献】
1) Benedetto, D. et al. “Language Trees and Zipping”. Physical Review Letters, Vol.88, No.4, p.048702-1-048702-4(2002)
2) Kukushkina, O.V; Polikarpov, A.A.; Khemelev, D.V. "Using letters and grammatical statistics for authorship attribution". Problems of Transmitting of Information, Vol.37, No2, 2001(英訳が http://www.philol.msu.ru/~lex/articles/grco_e.htm より入手可能)
3）内山和也. “スタイルの計量に関する覚え書き-文体論の視点から” 計量国語学, Vol. 23, No.7, p.347-352(2002)
4）石田栄美ほか 4 名. “文体からみた学術的文

献の特徴分析”. 2004 年度三田図書館・情報学会研究大会発表論文集, p.33-36 (2004)
5）村上征勝. 真贋の科学 : 計量文献学入門. 東京, 朝倉書店, 1994. 154p.
6) 村上征勝. “著者を探る古文書の計量分析”. 電子情報通信学会誌. Vol.85, No.3 p.158-161(2002)
7) 佐藤進也ほか 2 名. 文字列出現頻度比較による情報源間の類似度判定. 情報学基礎研究会, Vol.66, No.16, p.119-126(2002)
8) 松浦司; 金田康正. "n-gram の分布を利用した近代日本語文の著者推定". 計量国語学, Vol.22, No.6, p.225-238(2000) 一部、[松浦司; 金田康正. "近代日本文学者 8 人による文章における文字 N-gram 分布を手がかりとする著者推定".『情報処理学会自然言語処理研究会報告』, Vol.99, No.95(NL134), p.31-38(1999)]も参照
9) Tankard, J. "The Literary Detective". BYTE, Vol.11, No.2, p.231-238(1986)
10) Ziv ,J; Lemple,A. “A Universal Algorithm for Sequential Data Compression”. IEEE Transactions on Information Theory, Vol.IT-23, No.3, p.337-343(1977)
11) 安形輝. “圧縮プログラムによる類似データの同定”. 2004 年度日本図書館情報学会研究大会発表要綱