1C5-05

発信者情報が付与されたテキストコーパスの分析について

On the Analysis of Source Identified Text Corpora

相澤彰子[*1]
Akiko AIZAWA

[*1]国立情報学研究所
National Institute of Informatics

This paper aims at establishing a methodological basis for analyzing source identified texts such as newspaper stories with their reporters' names and academic papers with their authors. We first presents a suffix-tree based clustering method for identifying reused term sequences with their corresponding document subsets, and then show some analytical results where the clustering scheme was applied to the actual text collections.

1. はじめに

現実の文書データベースや Web 上には、引用、編集、改訂等により、文章や段落にわたる長い単位でテキストが字句どおり一致する文書が多数存在する。たとえば、Reuters などの代表的なテストコレクションを分析すると、単語の分布特性に基づく類似度が上位となる文書グループでは、グループ内の文書どうしがほぼ同一であることがわかる [1]。この性質は、単語分布に基づく重複文書検出の前提条件ともなっており、一般に成り立つものであると考えられる。

ここで注意しなければならないのは、単語分布の統計的な類似性と、単語列の比較的長い単位にわたる一致は、互いに強い相関はあるが、本質的に異なる事象であるということである。すなわち前者は、文書どうしの話題の共通性に起因するものであるが、後者は、著者や発信者の同一性、あるいは同一コミュニティ内での参照・被参照関係といった、いわば社会的なつながりを示唆するものである。しかしながら、従来のテキスト処理研究のほとんどは前者に注目するもので、文書を順序関係を持たない単語の集合として捉えるか、単語 N グラムを用いる場合でも、N の値は文法的な結びつきが強い範囲（2 から高々 10 程度）に設定される場合が大半であった。

上記の観察に基づき本稿では、複数の文書にまたがり反復される比較的長い単位の単語列を抽出し、文書をグループ化するための基本的な手法について検討する。また、実際に発信者情報が付与されたテキスト文書を対象として、話題や著者関連性の分析を試みた結果を報告する。

検討に先立ち本稿では、文書間で反復される様々な長さの単語列を、分析の目的に応じて以下の 3 つのタイプに分類する。

(1) 複合語や定型句
辞書登録の対象となるような、用法が確立した語彙や言い回しで、比較的高い頻度でコーパス中に広く出現するもの。

(2) 限定的定型句
特定の話題やコミュニティの範囲で、一時的に頻出する比較的長い単位の言い回しで、著者や発信者を特定することなく使い回されるもの。

(3) 引用テキスト
著者や発信者に特有の言い回しで、基本的に反復利用にあたっては引用の断り書きが必要であるとみなされるもの。

連絡先: 相澤彰子（国立情報学研究所）
〒 101-8430 東京都千代田区一ツ橋 2-1-2
Email:akiko@nii.ac.jp

本稿で注目するのは第二、第三のタイプ、すなわち「限定的定型句」および「引用テキスト」である。これらのタイプの単語列は従来の自然言語処理ではあまり注意が払われることがなかったが、与えられた文書の話題や発信者を強く特徴づけるものである。反復して利用されることから往々にして有用な情報を含んでおり、たとえば、新聞記事の場合には、ある事件に対する一連の報道で一貫して使われる組織名や人物の肩書き、複数の地方版で共通する記事本文の重要部などが該当する。

なお本稿では、文書に記名された名前を「発信者」としており、特に著作権者については意識していない。したがって、特定の定型句や引用文に著作権が想定されるのかといった問題は本稿の考慮の対象外である。

2. 接尾辞木構造に基づくクラスタリング

2.1 ＳＴクラスタの定義

文書集合を $\mathcal{D}$、文書中に出現するすべての単語（異なり）の集合を $\mathcal{W}$ として、単語 N グラム（$N \geq 1$）は、N 個の単語の順序つき集合として以下で与えられる。

$$w_1^N = (w_1, \cdots, w_N) \qquad (w_i \in \mathcal{W}) \tag{1}$$

D の接尾辞木（suffix-tree）とは、D 中のすべての単語列を含む圧縮した木構造、すなわちトライである。接尾辞木上の各ノードは、異なる単語 N グラム w_1^N に対応しており、「w_1^N を含むすべての文書」として唯一に定まる文書集合と対応づけられている。以下、このような文書集合を $S_d(w_1^N)$ （$\subset \mathcal{D}$）と表記する。 ここで複数のノードが同じ文書集合に対応する場合があることに注意して、ＳＴクラスタを、同じ文書集合に対応付けられる単語 N グラムの集合を用いて、以下のように定義する。

> **（定義）** S を単語 N グラム集合、D を文書集合として、$s \subset S$ なるすべての s について $S_d(s) = D$ であり、かつ、$s \not\subset S$ なるすべての s について $S_d(s) \neq D$ であるとき、$c = (S, D)$ をＳＴクラスタと呼ぶ。

2.2 ＳＴクラスタの生成手順

ＳＴクラスタ生成では、まず単語列の配列 (suffixy array) を求め、次に接尾辞木を生成して、これに基づきＳＴクラスタを数え上げる。基本的な手順を以下に示す [2, 3]。

(1) １つの文書を１つの単語列に対応させて、各単語へのポインタを格納した配列を生成する。日本語の場合には、形

態素解析を適用して分かち書き処理を行う。正規化や不要語の削除は行わない。

(2) 上記のポインタを一括してソートし、単語列の配列表現を生成する。このとき、接尾辞木上のノードは配列上の連続した区間に対応しており、区間内にあるポインタのアドレスから文書集合がただちに特定される。これを利用して接尾辞木を生成する。また、各ノードに対応する文書集合を配列に格納し、あらかじめ定めた順序にしたがって別々にソートしておく。

(3) 生成した接尾辞木ノードすべてを、(2) の文書配列をキーとしてソートする。ソート後のリストにおいて、同じ文書集合に対応づけられたノードどうしは隣接しており、ＳＴクラスタは連続した１区間として求められる。

2.3 ＳＴクラスタの評価尺度

(1) 単語列一致度

抽出した反復単語列の一致の強さを数量的に評価するために、「単語列一致度」(term sequence coincidence) を、(個別) 相互情報量 $M(w_1^N)$ の定義にしたがって次式で求める。

$$M(w_1^N) = log\frac{P(w_1^N)}{P(w_1)\cdots P(w_N)} \tag{2}$$

すなわち、$M(w_1^N)$ は、(i) k 個の単語 $(w_1, \cdots, w_k)$ が独立に生じたと仮定する場合のエントロピーと、(ii) 実際に観測された生起に基づき求めたエントロピーの差分である。

直感的には、式 (2) の $M(w_1^N)$ の値は長い単語列の場合ほど大きくなる。単純に単語列長を尺度として用いる場合との違いは、式 (2) では頻度を考慮して特殊な事象ほど高く重みづけている点である。別途行った予備実験において、単語列長と式 (2) による一致度の両者を、クラスタ内文書のカテゴリや著者のばらつきとの相関を用いて比較したところ、後者の方がよい値を示すことを確認している。

式 (2) の $P(w_1^N)$ は単語列 w_1^N の出現確率であり、w_1^N の全文書集合 $\mathcal{D}$ 中での出現頻度 $freq(w_i)$ および総語数 $F = \sum_{w_i \in \mathcal{W}} freq(w_i)$ を用いて次式で計算する。

$$P(w_1^N) = \frac{freq(w_1^N)}{F} \ . \tag{3}$$

単語 w_i の出現確率 $P(w_i)$ についても同様に、w_i を長さ１の単語列であるとみなして式 (3) から求める。ここでは未知語の確率推定を行う必要がないことから、単純さを重視して、平滑化による確率補正は適用していない。

ＳＴクラスタ (S, D) が与えられた場合、まず S の中のすべての単語列について上式 (2) による単語列一致度を計算し、クラスタ全体としての一致度は、これらの値の総和または最大値として求める。いずれを評価値とするかは、分析の目的に応じて定めるものとする。本稿の実験では統計的な特性を調べるため、一貫して最大値を尺度として用いている。

(2) 単語分布一致度

ＳＴクラスタ内の文書間類似度の数量尺度として、「単語分布一致度」を以下で定義する。まず、各文書について、標準的な正規化や不要語削除手順を適用して索引語を抽出し、次に、*tf-idf* 重みづけによる文書ベクトルを求める。さらに、これらの文書ベクトルの単純加算によりクラスタ全体の代表ベクトルを生成し、最後に、クラスタ内の各文書とクラスタの代表ベクトルのコサイン尺度による類似度の平均値を、クラスタ全体の類似度として計算する。すなわち、文書集合 D の単語分布類似度は、$\vec{d}$ を各文書の文書ベクトル、$\vec{c}$ をクラスタの代表ベクトルとして、ＳＴクラスタ (S, D) の単語分布一致度を次式で定める。

$$Sim(D) = \sum_{d \in D} \frac{\vec{d} \cdot \vec{c}}{|\vec{d}||\vec{c}|} \tag{4}$$

ここで、$Sim(D)$ の値は $0 \leq Sim(D) \leq 1$　であり、ＳＴクラスタ内の文書の統計的な単語分布が類似すればするほど、その値は１に近づく。

3. 実験結果

3.1 対象とするテキスト文書セット

実験では表 1 に示す６つの文書セットを用いた。Reuters [14] および SJM (San Jose Mercury) [15]　は英文のニュース記事、Mainichi [16] および NIKKEI [17] は、毎日新聞および日本経済新聞 CD-ROM 版から抽出した和文のニュース記事である。また、ntc-IPSJ および ntc-JSCE はそれぞれ、NTCIR-1[18] に収録された学会発表文献抄録データのうち、情報処理学会および土木学会における発表文献に対応している。新聞記事については、発信者が記事中に示されている記事だけを選んで実験に用いた。具体的には、Reuters および San Jose Mercury については、明示的に <byline> でタグつけされた文字列を、Mainichi および NIKKEI については、記事末尾に特定のフォーマットで付与された発信者情報（“【ワシントン１５日＜記者名＞】” など）を分析して用いた。日本語の分かち書き処理には形態素解析ツール茶筅 [19] を用いた。

なお表 1 では、各文書セットの収録期間、言語、文書数に加えて、前節で述べた手法を適用して抽出したＳＴクラスタの総数、および 2.8GMHz Xeon/Linux（1CPU）による実行時間を示している（ただし、比較のため形態素解析および辞書作成の時間は含んでいない）。さらに参考のため、各文書セットごとにテキストを文単位に分割し、式 (2) による一致度の文あたり平均値もあわせて示している。括弧内の数値は文あたり平均単語数である。

3.2 実験 1: 限定的定型句に注目した分析

最初の実験では、話題の重なりが少ない２つの文書セットを選び、これらに共通する単語列を抽出して、一致度の分布を調べた。ここでの主な目的は、発信者の直接参照や見聞なく再現される単語列について、対象とするテキストの違いによる一致度分布の揺らぎを調べることである。

このために、まず、(a)Reuters と San Jose Mercury (SJM)、(b) Mainichi と Nikkei、および (c) ntc-IPSJ と ntc-JSCE、それぞれの文書セットのペアを混合した上でＳＴクラスタを生成した。次に、その中から、Reuters と SJM のように異なる由来の文書が混在する「混合クラスタ」を抽出し、その一致度分布を調べた。ここで (a)(b)(c) それぞれのペアは、互いに年代や分野をかえることで、話題の重複がなるべく少なくなるように配慮してある。

図１に、(a)(b)(c) の３ペアに関して抽出した混合クラスタの一致度分布について、面積が１になるよう正規化したヒストグラムの比較結果を示す。また、抽出した混合クラスタの 95%がその値以下になるような「95%境界値」は以下のようになった。

Reuters-SJM	···	38.7	(6)
Mainichi-Nikkei	···	47.4	(7)
IPSJ-JSCE	···	45.5	(8)

表 1: 実験で用いたテキスト文書集合

文書セット	収録期間	言語	文書数	ＳＴクラスタ数	実行時間	文あたり一致度
Reuters	1996.8.20-1997.8.19	Eng	109,433	1,338,735	2644 sec.	330 (24.5)
SJM	1991.1.1 -1991.12.31	Eng	72,947	320,457	595 sec.	361 (30.1)
Mainichi	1998.1.1 -1998.12.31	Jpn	10,855	111,406	78 sec.	394 (33.6)
Nikkei	1996.1.1 -1996.12.31	Jpn	911	19,745	10 sec.	274 (27.5)
ntc-IPSJ	1988.5.19-1997.7.25	Jpn	26,796	226,640	99 sec.	420 (32.4)
ntc-JSCE	1991.9.17-1996.9.17	Jpn	21,259	180,538	70 sec.	434 (35.3)

言語や分野の違いを考慮すると、3つの文書セットすべてについて、一致度の分布は一貫した傾向を示しているといえる。すなわち、互いに話題の重なりが少ない文書の場合、値が 50∼100 よりも大きくなるような反復単語列の存在は、比較的まれである。言い換えれば、それより一致度が高い単語列の存在は、何らかの話題あるいは分野的なつながりを示すことが推察される。ただし、特定の文書間で話題が一致したからといって、必ずしも単語列一致度が高い値を示すわけではないので、上記の 95% 境界値は、話題の一致不一致を判別するためのものではないことに注意が必要である。

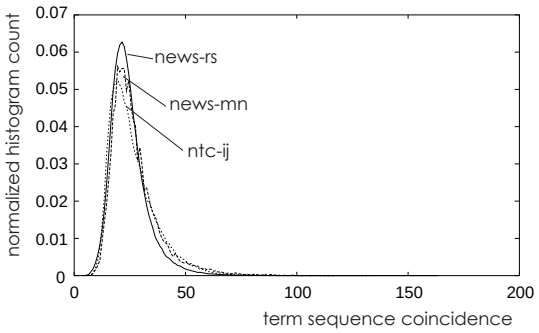

図 1: 話題の重なりが少ない場合の一致度の分布

3.3 実験 2: 引用テキストに注目した分析

次の実験では、Reuters, Mainichi, ntc-IPSJ の 3 つの文書セット各々について、クラスタ内文書の著者が互いに異なるＳＴクラスタを抽出し、単語列一致度の分布を調べた。ここで著者が「異なる」とは、クラスタ内の文書すべてに共通であるような著者（共著者を含む）が存在しない場合を指している。

図 2 に、正規化したヒストグラムによる一致度分布の比較を示す。また、著者が異なるクラスタの「95%境界値」は、それぞれ以下のようになった。

Reuters ··· 907.4 (84)
Mainichi ··· 164.2 (20)
ntc-IPSJ ··· 110.2 (15)

図 1 の場合と比較すると、対象文書セットによる違いが見られることがわかる。

この文書セットによる違いをさらに明らかにするために、クラスタ内文書の日付偏差による 95%境界値の変化を調べた。日付偏差 t を、0 ∼ 30 日の範囲で変化させ、偏差 t 以下となるクラスタについて 95%境界値を求めたところ、Mainichi および ntc-IPSJ では日付による大きな変化は見られないのに対して、Reuters では偏差が数日以内で 95%値は大きく減少して Mainichi の場合よりも小さくなった。すなわち、Reuters における一致度の値は、記事の日付に依存しており、数日以内に発信される記事の間では記者名によらず反復が行われている。

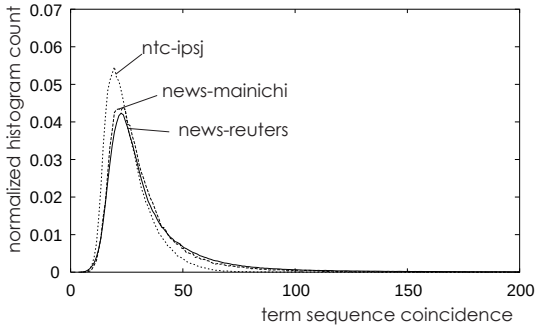

図 2: 著者が異なるＳＴクラスタの単語列一致度の分布

このことからテキスト反復利用の形態は、メディアや分野に依存して様々であることが推測される。

3.4 実験３: 限定的定型句の例

現実問題として、「限定的定型句」と「引用テキスト」の間には明確な定義上の区別は存在しないと考えられ、数量的にも両者を区別することは困難である。しかしながら、発信者情報を手がかりとすることにより、話題語の候補をある程度絞り込むことはできる。例として、Mainichi から抽出した一致単語列を先頭と末尾の助詞を取り除くなどして簡単に整形し、異なり著者数が 5 以上、日付偏差が 3 日以上の条件のもとで文書間一致度に基づく上位 10 個を選んだ結果を表 2 に示す。一般には、カテゴリが付与されていない文書から統計的にこのような単語列を識別することは容易ではなく、著者や日付等を手がかりとした分析の有効性がうかがえる。

表 2: Mainichi から抽出した限定的定型句の例

類似度	一致度	単語列
0.81	22.91	「 二つ の 中国 」
0.74	101.06	汪 道 涵 （ お う どう かん ） ・
0.72	76.25	台湾 政策 に関する 「 三つ の 不 支持 」
0.71	46.49	台湾 の 対 中 交流 窓口
0.71	62.16	物質 生産 禁止 （ カットオフ ） 条約
0.71	46.14	最恵国 待遇 （ ＭＦＮ ）
0.70	93.83	クルド 労働 者 党 （ ＰＫＫ ） の オジャラン 党首
0.68	36.22	胡錦涛 ・ 中国 副 主席
0.68	31.16	元 チリ 大統領 の 身柄
0.68	47.93	リー ・ クアンユー 上級 相

4. 関連研究と考察

本稿では、発信者、日付、カテゴリ等があらかじめ付与されたテキスト文書セットを用いて、長い単位での単語列の一致に

ついて統計的な分析を試みた。本稿で検討の対象とした反復単語列の抽出と利用の問題は、テキスト流通の限定的な一面を捉えたものではあるが、以下に示すように、情報検索の多様な応用分野と関係が深く、今後の発展が期待される。

(1) 著者同定への適用

従来より計量文献学の分野では、出典が不明、あるいは盗作や贋作の判定が必要な文献の著者同定問題に関して統計的なアプローチが適用されてきた [4, 5]。具体的には、著者の執筆のスタイルやクセを文の長さ、語彙、特殊な文末表現等の様々な尺度で数量化して、著者同定を行うものである。計算機の高速化や分かち書き処理技術の向上を背景に、最近の研究では、日本語文献でも単語Nグラム（$N = 3 \sim 10$）が尺度として取り入れられている [8, 9]。ここで、過去の研究の多くは、古典的な著作や文学作品の真贋の識別を目的としていたが、近年では、上記のような著者数量化のフレームワークを一般の発信者にまで適用する試みもある [6, 7]。無記名のテキストに対して、著者や話題の単一性をどのように数量化したらよいかという問題は、今後のテキスト流通において重要であり、本稿における検討結果を含め、新たな方向の模索が必要とされる。

なお本稿における検討は著作権の問題とも無関係ではないが、ここでの目的は無断引用やレポート複製の摘発を行うことではない。というのも、意図的な借用ではテキストに若干の変更を加えることが容易であり、摘発のポイントは、このような表層的な書き換えに対処するための意味処理となるためである（たとえば [20]）。著作権的な側面からはむしろ、著者自身の意図せざる借用の事前通知といった予防的な方向への展開が期待される。

(2) 重複文書検出への適用

テキストの電子化によって多数の文書の中から重複する文書を検出する問題が、近年とりわけ重要となっている。重複文書の検出では、多くの場合、文書間での若干の差異を許容している。すなわち、単純に完全一致する文書ペアを抽出するだけではなく、あらかじめ定めた類似度の閾値を超えるものを重複と判定する。Chowdhury らは、従来の重複文書検出手法を以下の２つに分類している [10]。第一は「shingles」と呼ばれる文字列の一致に基づく方法であり、連続した単語列など文書固有の要素集合を比較することで効率的に重複文書の検出を行おうというものである [11]。第二は 類似度計算に基づく方法であり、単語分布の類似度を用いて重複文書の候補を検出しようとするものである [1, 12]。本稿の手法は前者のタイプに属しているが、文書全体を１つの単位として比較するのではなく、文書中に含まれるテキストの部分的な重複に注目した検出が行えることから、より柔軟な重複チェックが期待できる。

(3) 高速文書クラスタリングと反復テキスト抽出

文書間に共通する単語列を利用した高速な文書クラスタリング法として、Zamir らは「Suffix Tree Clustering」（以下 STC） を提案している [13]。STC は接尾辞木構造を利用する点で、本稿におけるクラスタリングともっとも関連が深いが、Web 検索結果のリアルタイム再構成を目的としている点が異なる。具体的には、STC は前処理で単語の正規化や文単位でのテキストの分割を適用しており、一定長以上の単語列についてはペナルティを課している。また STC では、いったん抽出した文書クラスタを基本クラスタ（base cluster）として、これらを統合することで最終的なクラスタリング結果を生成している。これに対して本稿の手法は、正確な単語列の一致の抽出を目的として処理を単純化し、大規模なテキストへの適用を試みたものである。

抽出した反復テキストの利用例としては、文書中に埋め込まれた引用テキストの同定による情報ナビゲーションや、発信者、日付、カテゴリ等のタグを利用したイベント特有表現の抽出等が考えられる。ただし、抽出したテキストから、さらに慣用句や重要文を取り出すためには、構文あるいは形態素を手がかりとしたテキストの整形、正規化や類語辞書を用いた意味処理、メディア固有表現の削除、文中で明示される引用箇所の切り分け等、さまざまな自然言語処理の要素技術が必要であり、詳細の検討が今後の課題となっている。

参考文献

[1] M. Sanderson: “Duplicate Detection in the Reuters Collection,” Technical Report of the Department of Computing Science at the University of Glasgow, TR-1997-5 (1997).

[2] 相澤: “テキストからの再利用文字列の抽出と分析”, 情報処理学会研究会報告、FI-71-24 (2003) （発表予定）.

[3] A. Aizawa: “Analysis of Source Identified Text Corpora: Exploring the Statistics of the Reused Text and the Authorship”, ACL2003 (2003) （発表予定）.

[4] 村上征勝. 「行動計量学シリーズ：真贋の科学」朝倉書店. 1994.

[5] Tony McEnery and Michael Oakes: “Authorship Identification and Computational Stylometry,” in Handbook of Natural Language Processing, Marcel Dekker Inc., 545–562 (2000).

[6] Y. Tsuboi and Y. Matsumoto: “Authorship Identification for Heterogeneous Documents,” SIG Notes of IPSJ, NL-148, 17–24 (2002).

[7] 佐藤、原田、風間: “文字列出現頻度比較による情報源間の類似性判定”, 情報処理学会研究会報告、FI-66-16, 119–126 (2002).

[8] 松浦, 金田: “近代日本小説家８人による文章の n-gram 分布を用いた著者判別”, 情報処理学会研究会報告, NL-137-1, 1–8 (2000).

[9] 吉田, 延澤, 平石, 斎藤: “著者判別に有効な特徴量の推定”, 情報処理学会研究会報告, NL-145-13, 83–90 (2001).

[10] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe: “Collection Statistics for Fast Duplicate Document Detection,” ACM Trans. on Information Systems, 20(2), 171–191 (2002).

[11] A. Z.Broder, S. C. Glassman, M. S. Manasse, and G. Zweig: “Syntactic Clustering of the Web,” Proc. of the Sixth International World Wide Web Conference, 391–404 (1997).

[12] H. García-Molina, L. Gravano, and N. Shivakumar: “dSCAM: Finding Document Copies Across Multiple Databases,” Proc. of Fourth International Conference on Parallel and Distributed Information System, 68–79 (1996).

[13] O. Zamir and O. Etzioni: “Web Document Clustering: A Feasibility Demonstration,” Proc. of SIGIR’98, 46–54 (1998).

[14] Reuters: “Reuters Corpus, Volume 1, English language, 1996-08-20 to 1997-08-19” (2000).

[15] D. Harman and M. Liberman: “TIPSTAR Complete” Linguistic Data Consortium (1993).

[16] 毎日新聞社1999. CD-毎日新聞 98 年版.

[17] 日本経済新聞社: “1996∼2000 年版 日経全文記事データベース” (2001).

[18] National Center for Science Information Systems: “NTCIR Test Collection 1” (1999).

[19] 松本, 北内, 山下, 平野, 松田, 浅原: “日本語形態素解析システム「茶筌」Version 2.0 使用説明書”, NAIST Technical Report, NAIST-IS-TR99012, 奈良先端科学技術大学院大学 (1999).

[20] 深谷, 山村, 工藤, 松本, 竹内, 大西: “頻度統計と概念辞書を用いた文章の類似性の定量化”, 情報処理学会研究会報告、NL-153-10, 73–79 (2003).