

多文書間の共通性の分析

川谷隆彦

日本ヒューレット・パッカード(株) ヒューレット・パッカード研究所

takahiko.kawatani@hp.com

本報告では、(1)複数の文書がどの程度話題を共通にするかをスコア-で表す、(2)文書間で共通する話題に対する近さに応じて各文書、各文に対してスコア-を与える、(3)複数の文書から話題の共通する文書のみを抽出する、ことを可能とする多文書間の共通性分析方法を提案する。提案手法は、「全文書からひとつづつ文を取り出して文の組み合わせると、組み合わせの数は各文書の文の数の積通り存在する。文書間の話題に関する共通度のスコア-は、文の組み合わせの各々において各文共通に存在する単語数の和、もしくは2乗和で決まる」という考えをベースとするものである。簡単な実験を通じて、提案手法の妥当性を確認し、結果を観察する。

Commonality Analysis among Multi-documents

Takahiko KAWATANI

Hewlett-Packard Labs Japan, Hewlett-Packard Japan

takahiko.kawatani@hp.com

This paper proposes a method of multi-document commonality analysis which enables (1) to give a score a document set according to how much the documents share the same topic, (2) to give a score each document or each sentence according to how degree its topic is similar to the common topic, (3) to extract documents with the common topic from among documents. The proposed method is based on the idea that the topic commonality score is determined by the sum or by the squared sum of the number of common words existing in combination of sentences each of which is extracted from each document. Through simple experiments this paper confirms validness of the method and observes the results.

1. まえがき

本報告では、複数とくに3個以上の文書からなる文書集合に対する文書間の話題の共通性の分析を目的として、以下の方法を提案する。

- A) 文書集合に共通の話題が存在するか否かを判断できるよう、各文書の話題がどの程度共通しているか数値で示す。
- B) 共通の話題に近い話題の文書、または文を選択して全文書の共通の話題を把握できるよう、共通の話題への近さに応じて各文書、または各文にスコア - を与える。

- C) 話題が全文書に共通していなくとも、話題の共通する文書群があればそれを抽出する。

これらの内、A)については、2文書の場合には話題の共通性のスコア-はその2文書間の類似度そのものと考えることができ、最も代表的ものとして余弦類似度が知られている。これは文書に現れる各単語の頻度を成分とするベクトルで文書を表現しておき、2文書間の類似度をそれぞれのベクトルのなす余弦によって定義するというものである。しかし、3文書以上の場合に対しては、2文書のときの類似度に相当する尺度は知られてい

ない。文書集合の共通性を議論するうえではこのような尺度は必須と考えられる。

上記 B)、C) は文書集合からの共通話題の抽出に関わる技術である。このような処理は複数文書要約や TDT (Topic Detection and Tracking) などで重要な技術となっている。従来は、共通話題の抽出は、文書のクラスタリングを行った後、クラスター毎にクラスターを代表する文や文書タイトルを選択することにより行われていた[1][2]。また、最近では文やパッセージ単位にクラスタリングを行い、クラスター毎に重要なパッセージを選択することで共通話題を抽出する方法も現れている[3][4]。何れにせよこれまでは共通話題の抽出にクラスタリングは欠かせない技術となっているが、問題も存在する。クラスタリングは階層的な手法と非階層的な手法に大別される[5]。階層的な手法は、さらにボトムアップのアプローチとトップダウンのアプローチに分けられる。前者では、初期状態として各文書をクラスターの核とし、最も近いクラスターをマージするという処理を繰り返す。これにより文書集合は木構造で表現されるようになるが、各レベルのクラスターが意味のあるグルーピングとなっている保証はない。意味のあるグルーピングを指向するには、類似度が閾値を超えるクラスター対のみをマージするようにすればよいが、閾値を如何に決定するかが問題となる。後者では、全文書が1つのクラスターに属するという状態から出発し、例えばひとつのクラスター中のあらゆる文書対の中で最も低い類似度が閾値以下の場合、そのクラスターを分割するという処理を繰り返す。この場合も閾値をどのように決めるかは任意性がある。また、階層的な手法では処理量の問題も無視できない。非階層的な手法では、予め指定された数のクラスターが何らかの基準を満たすように作成される。従って、この手法を用いるには与えられた文書集合が何個のクラスターから構成されるか事前の知識が要求されるが、これは一般的には得られない情報であり、クラスター数を正しく入力することは困難であった。このようなことから、本報告では従来のクラスタリングに依らない共通話題の抽出法を目指す。

本報告で、上記 A)、B)、C) を実現するためのアプローチは以下のとおりである。まず、2つの文を考えると、この2つの文の間の共通度は共通す

る単語の数で決まると考える。また、2つの文書間の共通度は、各文書から文を1つずつ取り出して組み合わせられた文の対における共通単語数の全ての対に対する和、もしくは2乗和で決まるとする。この場合文の対は各文書の文数の積通り存在することになる。3文書以上の場合も、文書間に存在する全ての文の組み合わせを考えればよい。ここでは、組み合わせられた文の全てに共通する単語で構成される文を共通文と呼ぶ。共通単語数の算出を容易にするため、本報告では、各文を各成分が対応する単語の有無を表す2値ベクトルで表したうえで、各文書を文ベクトルの集合で表す。さらに、各共通文に対して共通文ベクトルを定義し、共通文ベクトルの集合に対する処理により共通度を求めていく。

以下、2.では、文書の表現法、共通文ベクトルの求め方を述べた後、文書や共通文ベクトル集合の共起行列を定義し、これらの性質について述べる。3.では、文書集合がどの程度共通の話題を有しているかを示す文書集合共通度を定義する。また、文書集合内の各文書、各文が文書集合の共通の話題とどれだけ近いかを示す文書(文)-文書集合共通度を定義する。4.では、全文書での共通話題の存在を前提としない文書集合共通度を定義し、これを用いて話題の共通する文書のみを抽出する方法について述べる。5.では、簡単な実験を通じて提案手法の有効性を示す。

2. 文書の表現と共通文ベクトル

2.1 文書の表現

現れる単語集合が $\{w_1, \dots, w_M\}$ で与えられ、 R 個の文書から成る集合 D を考える。ここで、 r 番目の文書を D_r とすると、 D_r は Y_r 個の文からなるものとし、 y 番目の文及びその文ベクトルを D_{ry} 、 $d_{ry}=(d_{ry1}, \dots, d_{ryM})^T$ とする。ここで、 T は転置を表す。 d_{ry} はバイナリベクトルであり、 d_{rym} は m 番目の単語の有無を表す。

次に、次式で定義される行列 S^r を考える。

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T \quad (1)$$

式(1)から分かるように、 S^r の mn 成分は

$$S^r_{mn} = \sum_{y=1}^{Y_r} d_{rym} d_{ryn}$$
により与えられる。従って、

S^r_{mn} は文書 D_r において単語 m が生起する文の数、 S^r_{mn} は単語 m と n とが共起する文の数を表すことになる。そこで、行列 S^r を文書 D_r の生起・共起行列、または簡単に共起行列と呼ぶこととする。

共起行列には次のような性質がある。ここでは同じ単語は同じ文で2回以上現れないものとする。

- (1) S^r の対角成分の和は各文に現れる単語数の和と等しい。 $\text{trace}(S^r) = \sum_m S^r_{mn}$ から分かるように、 S^r の対角成分の和は文書 D_r に現れる単語の総数に等しく、従って、各文に現れる単語数の和とも等しくなる。
- (2) S^r の全成分の和は文書 D_r の各文に現れる単語数の2乗和に等しい。これは、文書 D_r の文 y における単語数を f_{ry} とすると、下記により示される。

$$\begin{aligned} \sum_{y=1}^Y f_{ry}^2 &= \sum_{y=1}^Y (d_{ry1} + \dots + d_{ryM})^2 \\ &= \sum_{y=1}^Y \sum_{m=1}^M \sum_{n=1}^M d_{rym} d_{ryn} \\ &= \sum_{m=1}^M \sum_{n=1}^M S^r_{mn} \end{aligned} \quad (2)$$

2.2 共通文ベクトル

3つのベクトル $a=(a_n)$ 、 $b=(b_n)$ 、 $c=(c_n)$ が与えられた時、ここでは、 $c_n = \min(a_n, b_n)$ により成分が定義されるベクトル c をベクトル a, b の共通ベクトルと定義する。例えば、ベクトル $(4,2,1,2)$ と $(1,3,1,4)$ との共通ベクトルは $(1,2,1,2)$ となる。本報告の場合、文ベクトルはバイナリなので、共通ベクトルの成分は $c_n = a_n \times b_n$ によっても求めることができる。3個以上のベクトルの共通ベクトル成分は対応する成分の中の最小値で定義するが、バイナリベクトルの場合は対応する成分同士の積となる。

ここで、3つの文書、 D_1, D_2, D_3 の間の全ての文の組み合わせに対して求められる $Y_1 \times Y_2 \times Y_3$ 通りの共通文ベクトルの共起行列 S^C を求めてみる。 D_1, D_2, D_3 のそれぞれの i, j, k 番目のベクトル d_{1i}, d_{2j}, d_{3k} の共通文ベクトルを $c^{ijk} = (c^{ijk}_m)$ で表すと、前述のように、 c^{ijk}_m は

$$c^{ijk}_m = d_{1im} d_{2jm} d_{3km} \quad (3)$$

で求められる。 S^C の各成分は

$$\begin{aligned} S^C_{mn} &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} c^{ijk}_m c^{ijk}_n \\ &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} d_{1im} d_{1in} d_{2jm} d_{2jn} d_{3km} d_{3kn} \\ &= S^1_{mn} S^2_{mn} S^3_{mn} \end{aligned} \quad (4)$$

となり、文書 D_1, D_2, D_3 の共起行列の対応する

成分同士の積として求められる。これは文書の数とは無関係に成り立つ関係である。結局、共通文ベクトルの共起行列は共通文ベクトルを実際に求めることなく得ることができる。

S^C_{mm} は、 $Y_1 \times Y_2 \times Y_3$ 個の共通文の中で単語 w_m の生起する共通文の数、 S^C_{mn} は単語 w_m と w_n の共起する共通文の数を表す。 S^C_{mm} は3文書全てで単語 w_m が生起しない限り、 S^C_{mn} は3文書全てで w_m と w_n が共起しない限り0以外の値を持たない。

3. 共通性の評価法

3.1 文書集合共通度

まず、文書集合として話題がどの程度共通するかを示す尺度として、文書集合共通度を提案する。前述のように、本報告では文書集合から得られる共通文の各々に存在する単語数（共通単語数）をもとに文書集合共通度を求めていく。これには、共通単語数の和をベースとする方法と、2乗和をベースとする方法とが考えられる。理解を容易にするため、先ず D が3文書 D_1, D_2, D_3 から成る場合について説明する。

共通単語数の和をベースとする方法（線形モデル）

D_1, D_2, D_3 のそれぞれの i, j, k 番目の文 D_{1i}, D_{2j}, D_{3k} に共通する単語数を $g(D_{1i}, D_{2j}, D_{3k})$ とすると、これは共通文ベクトル c^{ijk} において値が1の成分数となる。従って、各共通文に現れる単語数の総和を $G(D_1, D_2, D_3)$ とすると、これは

$$\begin{aligned} G_1(D_1, D_2, D_3) &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} g(D_{1i}, D_{2j}, D_{3k}) \\ &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} (c^{ijk}_1 + \Lambda + c^{ijk}_M) \\ &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} (c^{ijk}_1^2 + \Lambda + c^{ijk}_M^2) \\ &= \sum_{m=1}^M S^C_{mm} \end{aligned} \quad (5)$$

により求められ、共通文ベクトルの共起行列の対角成分の総和で与えられることが分かる。 $G_1(D_1, D_2, D_3)$ は文書 D_1, D_2, D_3 の情報共通量とも呼ぶべきものであるが、このままでは文書の長さの影響を受ける。また、文書数によって値域が容易に変動する。これらの影響を排除するための正規化を行うことにより、線形モデルでの文書集合共通度 $com_1(D)$ は以下のように定義できる。

$$\begin{aligned}
com_l(D) &= \left[\frac{G_l(D_1, D_2, D_3)}{\sqrt[3]{G_l(D_1, D_1, D_1)G_l(D_2, D_2, D_2)G_l(D_3, D_3, D_3)}} \right]^{1/2} \\
&= \left[\frac{\sum_{m=1}^M S^C_{mm}}{\sqrt[3]{\sum_{m=1}^M S^1_{mm}{}^3 \sum_{m=1}^M S^2_{mm}{}^3 \sum_{m=1}^M S^3_{mm}{}^3}} \right]^{1/2} \\
&= \left[\frac{\sum_{m=1}^M S^1_{mm} S^2_{mm} S^3_{mm}}{\sqrt[3]{\sum_{m=1}^M S^1_{mm}{}^3 \sum_{m=1}^M S^2_{mm}{}^3 \sum_{m=1}^M S^3_{mm}{}^3}} \right]^{1/2}
\end{aligned} \tag{6}$$

式(6)において、正規化のポイントは、3文書が同一のとき共通度は1になるようにしたこと、及び3文書の時は文書の突合わせは(3-1)回行われたので、(3-1)乗根を求めるようにしたことにある。後述のように、 R 文書の場合は $R-1$ 回の文書の突合わせが行われるので $R-1$ 乗根を求めるようにする。

共通単語数の2乗和をベースとする方法(2次モデル)

各共通文に現れる単語数の2乗和を $G_s(D_1, D_2, D_3)$ とすると、

$$\begin{aligned}
&G_s(D_1, D_2, D_3) \\
&= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} g(D_{1i}, D_{2j}, D_{3k})^2 \\
&= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} (c^{ijk}_1 + \Lambda + c^{ijk}_M)^2 \tag{7} \\
&= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} \sum_{m=1}^M \sum_{n=1}^M c^{ijk}_m c^{ijk}_n \\
&= \sum_{m=1}^M \sum_{n=1}^M S^C_{mn}
\end{aligned}$$

と求められる。 $G_s(D_1, D_2, D_3)$ は共通文ベクトルの共起行列の各成分の総和に他ならない。結局、2次モデルでの文書集合共通度 $com_s(D)$ は以下のように定義できる。

$$\begin{aligned}
com_s(D) &= \left[\frac{G_s(D_1, D_2, D_3)}{\sqrt[3]{G_s(D_1, D_1, D_1)G_s(D_2, D_2, D_2)G_s(D_3, D_3, D_3)}} \right]^{1/2} \\
&= \left[\frac{\sum_{m=1}^M \sum_{n=1}^M S^C_{mn}}{\sqrt[3]{\sum_{m=1}^M \sum_{n=1}^M S^1_{mn}{}^3 \sum_{m=1}^M \sum_{n=1}^M S^2_{mn}{}^3 \sum_{m=1}^M \sum_{n=1}^M S^3_{mn}{}^3}} \right]^{1/2} \\
&= \left[\frac{\sum_{m=1}^M \sum_{n=1}^M S^1_{mn} S^2_{mn} S^3_{mn}}{\sqrt[3]{\sum_{m=1}^M \sum_{n=1}^M S^1_{mn}{}^3 \sum_{m=1}^M \sum_{n=1}^M S^2_{mn}{}^3 \sum_{m=1}^M \sum_{n=1}^M S^3_{mn}{}^3}} \right]^{1/2}
\end{aligned} \tag{8}$$

式(6)で与えられる $com_l(D)$ と比較すると、式(8)の $com_s(D)$ では共通文ベクトルの共起行列の非対角成分によって共通文における単語間の共起の関係が反映されているのが特長である。

上記では、3文書を対象に2つの文書集合共通度 $com_l(D)$ 、 $com_s(D)$ を導いたが、 R 文書を対象と

する場合にも以下のように定義できる。

$$com_l(D) = \left[\frac{\sum_{m=1}^M S^C_{mm}}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M (S^r_{mm})^R}} \right]^{1/(R-1)} \tag{9}$$

$$\begin{aligned}
com_s(D) &= \left[\frac{\sum_{m=1}^M \prod_{r=1}^R S^r_{mm}}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M (S^r_{mm})^R}} \right]^{1/(R-1)} \\
&= \left[\frac{\sum_{m=1}^M \sum_{n=1}^M S^C_{mn}}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M \sum_{n=1}^M (S^r_{mn})^R}} \right]^{1/(R-1)} \\
&= \left[\frac{\sum_{m=1}^M \sum_{n=1}^M \prod_{r=1}^R S^r_{mn}}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M \sum_{n=1}^M (S^r_{mn})^R}} \right]^{1/(R-1)}
\end{aligned} \tag{10}$$

3.2 2文書の文書集合共通度と類似度との関係

$R=2$ として文書 D_1 、 D_2 を対象文書とすると、式(9)は以下のように変形できる。

$$com_l(D) = \frac{\sum_{m=1}^M S^1_{mm} S^2_{mm}}{\sqrt{\sum_{m=1}^M S^1_{mm}} \sqrt{\sum_{m=1}^M S^2_{mm}}} \tag{11}$$

式(11)は、2文書の場合の文書集合共通度は、それぞれの文書の共起行列の対角成分を成分とする2つの M 次元ベクトルの余弦類似度に他ならないことを示している。2.1で述べたように、各文書の共起行列の対角成分は対応する単語を含む文の数を表す。従って、同じ単語が同じ文に2回以上現れないと仮定すると、2文書の文書集合共通度は文書内の単語頻度を成分とする文書ベクトルの類似度と全く同じとなる。

一方、 $R=2$ の場合は式(10)は以下のように変形できる。

$$com_s(D) = \frac{\sum_{m=1}^M \sum_{n=1}^M S^1_{mn} S^2_{mn}}{\sqrt{\sum_{m=1}^M \sum_{n=1}^M (S^1_{mn})^2} \sqrt{\sum_{m=1}^M \sum_{n=1}^M (S^2_{mn})^2}} \tag{12}$$

式(12)は、2文書の場合の文書集合共通度は、それぞれの文書の共起行列の各成分を成分とする2つの M^2 次元ベクトルの余弦類似度として求められることを示している。前節で述べたように、式(12)の分子は各共通文に現れる単語数の2乗和 $G_s(D_1, D_2)$ を表している。また、2つの文 D_{1i} 、 D_{2j} の共通単語数 $g(D_{1i}, D_{2j})$ は $d_{1i}^T d_{2j}$ と表すことが出来るので、

$$G_s(D_1, D_2) = \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} (d_{1i}^T d_{2j})^2 \tag{13}$$

と書くことができ、これを用いると $com_s(D)$ は

$$com_s(D) = \frac{\sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} (d_{1i}^T d_{2j})^2}{\sqrt{\sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} (d_{1i}^T d_{1j})^2} \sqrt{\sum_{i=1}^{Y_2} \sum_{j=1}^{Y_2} (d_{2i}^T d_{2j})^2}} \quad (14)$$

とも表すことができる。式(14)は筆者が先に提案した SVSM 類似度そのものである[6][7]。SVSM 類似度では式(14)の右辺で示されるように、対象となる文書間の全ての文ベクトルの組み合わせから求められる内積の2乗和をベースにしている。また、上記の議論から SVSM 類似度は式(11)のように両文書の共起行列の対応する成分同士の積和から求められることが分かるが、これについては前報では述べていなかった。

以上により文書集合共通度は従来から知られている2文書間の類似度の自然な延長となっていることが分かる。

3.3 文書、文の共通度

対象とする文書または文を P として、本節では P が文書集合 D の共通の話題にどれだけ近いかを示す尺度として、文書 文書集合共通度を定義する。これには、次の2つの方法が考えられる。

第1の方法は、 P を文書集合 D に加えた新しい文書集合の文書集合共通度を文書 文書集合共通度とする考え方である。文書 P の共起行列を S^P として、線形モデル、2次モデルの場合の文書 文書集合共通度を $com_l(D+P)$ 、 $com_s(D+P)$ とすると、これらは以下のように定義できる。

$$com_l(D+P) = \left[\frac{\sum_{m=1}^M S^C_{mm} S^P_{mm}}{R+1 \sqrt{\sum_{m=1}^M (S^P_{mm})^{R+1}} \prod_{r=1}^R \sum_{m=1}^M (S^r_{mm})^{R+1}} \right]^{1/(R)} \quad (15)$$

$$com_s(D+P) = \left[\frac{\sum_{m=1}^M \sum_{n=1}^M S^C_{mn} S^P_{mn}}{R+1 \sqrt{\sum_{m=1}^M \sum_{n=1}^M (S^P_{mn})^{R+1}} \prod_{r=1}^R \sum_{m=1}^M \sum_{n=1}^M (S^r_{mn})^{R+1}} \right]^{1/(R)} \quad (16)$$

第2の方法は、 P を構成する文ベクトルと文書集合 D における共通文ベクトルとの類似度により文書 文書集合共通度を定義する方法である。これには余弦類似度を用いる場合と SVSM 類似度を用いる場合の2通りが考えられる。前者の場合、 P を構成する文ベクトルの和ベクトルと共通文ベクトルの和ベクトルとの間での余弦類似度を

求めることになるが、共通文ベクトルの和ベクトルの各成分は、 S^C の対応する対角成分となるので余弦類似度は容易に求めることができる。後者の場合には、式(12)をベースとすることができる。得られる文書 文書集合共通度を前者について $com_l(D, P)$ 、後者について $com_s(D, P)$ と表記すると、

$$com_l(D, P) = \frac{\sum_{m=1}^M S^C_{mm} S^P_{mm}}{\sqrt{\sum_{m=1}^M (S^C_{mm})^2} \sqrt{\sum_{m=1}^M (S^P_{mm})^2}} \quad (17)$$

$$com_s(D, P) = \frac{\sum_{m=1}^M \sum_{n=1}^M S^C_{mn} S^P_{mn}}{\sqrt{\sum_{m=1}^M \sum_{n=1}^M (S^C_{mn})^2} \sqrt{\sum_{m=1}^M \sum_{n=1}^M (S^P_{mn})^2}} \quad (18)$$

と定義することができる。

文書集合 D の共通の話題を把握するには、 P を文書集合 D に含まれる各文書の各文として文 文書集合共通度を求め、その値の大きい幾つかの文を選択してユーザに見せるのが効果的である。しかしながら、文 文書集合共通度は内容の違いまで示すものではないので、同じような文が選択されてしまう可能性がある。このような重複の排除には MMR (Maximal Marginal Relevance) [8] の利用が考えられる。

4. 部分的に存在する共通性の検出

3章で述べた文書集合共通度は全文書の話題が揃っているかどうかの尺度を与えるものであり、文書 文書集合共通度は全文書の話題が揃っていることを前提に各文書、各文の共通話題に対する近さの程度を与えるものであった。本章では、話題が必ずしも共通でない文書群から同じ話題の文書の抽出を可能にするため、部分的に存在する共通性の検出を試みる。

4.1 共通文ベクトルの共起行列の求め方

2.2 においては式(4)のように各文書の共起行列同じ成分の積により S^C の対応する成分を求めていたため、0以外の値をとるのは全文書で生起・共起する単語、単語対に対応する成分に限られていた。ここでは、共通文ベクトルの共起行列の各成分の算出には、各文書の共起行列の内その成分の値がゼロでない共起行列のみを用いるようにする。従って、各単語、単語共起が文書集合 D に必ず現れる限り、共通文ベクトルの共起行列の成分は0以外の値をとる。

このような共通文ベクトルの共起行列を T とし、

さらに、各単語、または各単語対の生起、または共起した回数を保持する行列を U とした場合、これらは以下のように求められる。

1. $r=1$ とおく。 T の全成分は 1、 U のそれは 0 とする。

2. $S^r_{mn} > 0$ のとき、

$$T_{mn} = S^r_{mn} T_{mn} \quad (19)$$

$$U_{mn} = U_{mn} + 1 \quad (20)$$

3. $r=R$ で終了。そうでなければ $r=r+1$ として 2へ。

さらに、 A を閾値として行列 T^A を以下のように決める。

$$\begin{aligned} T^A_{mn} &= T_{mn}, & \text{if } U_{mn} \geq A, \\ T^A_{mn} &= 0 & \text{otherwise.} \end{aligned} \quad (21)$$

上記のように、行列 T^A は A 文書以上で生起した単語、共起した単語対に対応する成分のみが値を有し得る共起行列である。また、2.2 では共通文ベクトルにおいて値が 1 となる成分は文の組み合わせにおいて全ての文ベクトルで値が 1 となる成分のみであったが、ここでは、 A 文書以上で現れる単語に対応する成分が値 1 となるように共通文ベクトルを決定したことになる。行列 T^A は、そのように決定された全共通文ベクトルから求められる生起・共起行列である。

4.2 用い方

例えば、式(10)において行列 S^C の替りに行列 T^A を用いた文書集合共通度を

$$com_l(D; T^A) = \left[\frac{\sum_{m=1}^M T^A_{mn}}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M (S^r_{mn})^R}} \right]^{1/(R-1)} \quad (22)$$

$$com_s(D; T^A) = \left[\frac{\sum_{m=1}^M \sum_{n=1}^M T^A_{mn}}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M \sum_{n=1}^M (S^r_{mn})^R}} \right]^{1/(R-1)} \quad (23)$$

のように定義する。 $com_l(D; T^A)$ 、 $com_s(D; T^A)$ は行列 T^A を共通文ベクトルの共起行列として用いて求められる文書集合共通度という意味である。これを不一致許容形の文書集合共通度と呼ぶことにする。 $com_l(D; T^R)$ 、 $com_s(D; T^R)$ は $com_l(D)$ 、 $com_s(D)$ とそれぞれ等価である。

ここで、文書集合 D では $B (< R)$ 個の文書が話題を共通にしており、他の文書は互いに関連のないものと仮定する。このとき、 A が B と等しいか B よりも小さければ、話題を共通にする B 個の文書に現れる単語の寄与によって行列 T^A の各成分の

和は大きくなり、 $com_s(D; T^A)$ や $com_l(D; T^A)$ も大きくなる筈である。一方、 A が $R \sim B+1$ の間にあるときは、偶発的に B 個以上の文書で現れる単語があったにしてもその単語の各文書内の頻度は高くないものと想定され、 T^A の全成分の和は小さくなり、 $com_s(D; T^A)$ や $com_l(D; T^A)$ も小さいものと考えられる。従って、 B の値を求めるには A の値を R から 1 ずつ減らしていき、 $com_s(D; T^A)$ の値が十分大きくなる A を検出すればよい。

しかし、この段階では R 個のうち、どの B 個の文書が話題を共通にするのかは分からない。話題が共通な文書は $com_s(D; T^A)$ の値を大きくする A により決定される共通文ベクトル集合と類似性が高い筈である。そこで式(17)(18)を以下のように変形し話題を共通にする文書の検出に用いることができる。

$$\begin{aligned} com_l(D, D_r; T^A) &= \left[\frac{\sum_{m=1}^M T^A_{mn} S^r_{mn}}{\sqrt{\sum_{m=1}^M (T^A_{mn})^2} \sqrt{\sum_{m=1}^M (S^r_{mn})^2}} \right]^{1/(R-1)} \end{aligned} \quad (24)$$

$$\begin{aligned} com_s(D, D_r; T^A) &= \left[\frac{\sum_{m=1}^M \sum_{n=1}^M T^A_{mn} S^r_{mn}}{\sqrt{\sum_{m=1}^M \sum_{n=1}^M (T^A_{mn})^2} \sqrt{\sum_{m=1}^M \sum_{n=1}^M (S^r_{mn})^2}} \right]^{1/(R-1)} \end{aligned} \quad (25)$$

ここで、 $com_l(D, D_r; T^A)$ 、 $com_s(D, D_r; T^A)$ は、行列 T^A を共通文ベクトルの共起行列として用いて求められる D_r 、 D 間の文書 文書集合共通度という意味である。これを不一致許容形の文書 文書集合共通度と呼ぶことにする。この値の大きな文書を選択すれば、話題を共通にしている文書が得られる筈である。選択された文書が話題を共通にしているか否かのチェックは式(9)(10)を用いることにより可能である。

また、式(24)(25)において、 D_r の代わりに文 P を用いると、不一致許容形の文 文書集合共通度を求めることができる。式(17)(18)では、全文書に現れる単語、単語対をもとに文 文書集合共通度が求められたが、この場合には A を適当な値に設定することにより、例えば大部分の文書に現れる単語、単語対をもとに文 文書集合共通度を求めるようにすることができる。

5. 実験

5.1 実験データ

実験データは Reuters-21578 から取り出した 21

記事であり、3グループに分けられる。内容は、グループ 1(G1)：カテゴリ”acquisition”から取り出した GenCorp 社の企業買収に関する 12 記事、グループ 2(G2)：カテゴリ”crude”から取り出したエクアドルの地震に関する 6 記事、グループ 3(G3)：カテゴリ”money-fx”から取り出した James Baker 氏の発言に関する 3 記事、である。ここでは、21 記事の中から文書数の最も多いグループ 1 を抽出し、さらにグループ 1 の共通話題を最もよく表す 3 つの文を選択することを目的とした実験を通して、結果の観察を行い提案手法の動作を確認する。実験は線形モデルで行われた。2 次モデルにおいても同様の結果が得られている。

行った主な前処理は、文切り出し、lemmatizing、ストップワード除去であり、出現した総単語数は 1147 であった。

5.2 実験結果

まず、表 1 に出現頻度が 13、12、11 回の単語について、 T の各単語に対応する対角成分の値(T_{mm})、出現文書における平均出現回数、各グループにおける各単語の出現する文書数を示す。14 回以上現れる単語は存在しない。表から分かるように、出現回数 12 回の ”GenCorp” は T_{mm} の値は非常に大きく、グループ 1 の話題を規定する単語となっている。図 1 は、 A の値と不一致許容形の文書集合

表1 高頻度単語の頻度、 T_{mm} 、出現した文書における平均出現回数、各グループ(G1、G2、G3)における出現文書数

頻度	単語	T_{mm}	平均出現回数	G1	G2	G3
13	dtrs	5.38×10^4	3.2	10	3	0
	industry	2.00	1.1	12	1	0
	make	5.00	1.7	9	4	0
12	100	1.00	1.0	11	1	0
	AFG	64.00	1.7	12	0	0
	Brown	4.00	1.2	12	0	0
	GenCorp	6.40×10^9	7.4	12	0	0
	offer	1.38×10^7	4.4	11	1	0
11	company	1.44×10^2	2.0	9	2	0
	general	3.50×10^6	5.2	11	0	0
	share	2.07×10^4	3.1	11	0	0
	tender	1.92×10^2	1.7	11	0	0
	Wagner	8.00	1.3	11	0	0

Z 共通度 $com_r(D;T^A)$ との関係を示す図である。図から分かるように、 A が 12 の時は $com_r(D;T^A)$ は 0.38 であり、 A が 13 のときの 0.22 よりも高くなっている。また、 A が 12 以下のとき $com_r(D;T^A)$ の値は殆ど一定となっている。これらは A が 12 以下では T_{mm} の値が大きい ”GenCorp” が $com_r(D;T^A)$ の値を支配するためである。結局、 A が 12 以下のときはグループ 1 の寄与が大きく表れており、 A を 12 以下に設定した不一致許容形の文書集合共通度 $com_r(D,D_r;T^A)$ の値が大きくなる文書を選択すればグループ 1 の文書が得られることになる。

図 2 は、 A を 12、13 に設定したときの各文書の $com_r(D,D_r;T^A)$ の値を示している。 A が 12 のときは、グループ 1 に属する文書のみが全体的に大きな値を有しており、グループ 1 に属する文書を容易に選択できることを示している。 A が 13 のときも表 1 での頻度 13 の単語を含む文書がゼロでない値をとっているが、値は A が 12 のときに比べて小さい。さらに、図 2 において $com_r(D,D_r;T^A)$ の値が 0.02 を越える文書を選択し、それらの文書の文書集合共有度を求めてみた。 A が 12 のときには 12 文書選択され文書集合共有度は 0.89 となり、 A が 13 の時に選択される 13 文書の文書集合共有度は 0.29 であった。この事実は A を 12 に設定して選択された文書群の方が A を 13 に設定した場合よ

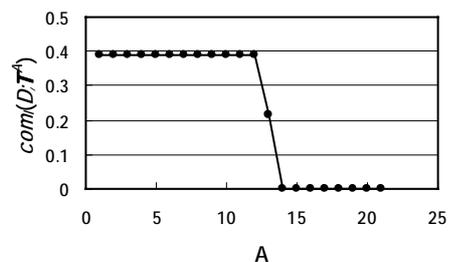


図1 A と $com_r(D;T^A)$ との関係

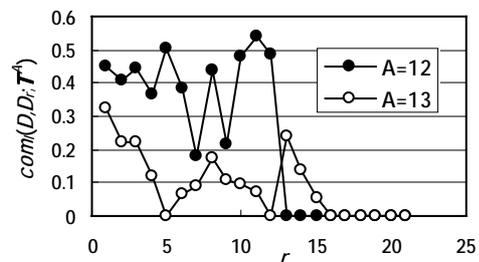


図2 各文書の不一致許容形の文書-文書集合共通度 $com_r(D,D_r;T^A)$

りも話題が揃っていることを物語っている。以上から、最も優勢な共通話題を反映する文書を選択するには、不一致許容形の文書集合共通度の値を大きくするような閾値 A を求める、その A を用いて求められる不一致許容形の文書 文書集合共通度の値の大きな文書を選択する、選択された文書の文書集合共通度を求め、話題が揃うことを確認する、というプロセスを踏めばよいことが分かる。

表 2 は、式(24)においてグループ 1 の各文書の各文を D_r の代わりに用いて求めた文 文書集合共通度の値の大きい 3 つの文を示している。この場合 T^A はグループ 1 に属する 12 文書から求めており、 A の値は 8 に設定している。また、 T^A_{mm} の値をそのまま用いた場合には“GenCorp”の影響が大きすぎるため、便宜的に T^A_{mm} の 12 乗根を用いた。MMR の処理は行っていない。表 2 からグループ 1 の文書の共通する話題が何かを窺い知ることができる。

以上の例では期待通りに共通話題が抽出できている。提案手法では、粗っぽく云えば、 S^C_{mm} 、 T^A_{mm} は単語 m の重みを表し、その値は各文書での頻度の積で決まる。そのため共通の話題を規定するような単語が各文書で現れれば、その単語には自然と大きな重みが付与される格好になる。これが提案手法の特長であるが、本実験で見られたように特定の単語が各文書で非常に高い頻度で現れる場合にはその単語のみで共通度がほぼ決まってしまうことになる。この事実の是非については議論の余地があり、単語の重み付けなど何らかの対策が必要となる可能性がある。

6. まとめ

以上、本報告では、多文書間の話題の共通性分析方法として、全文書一致形、及び不一致許容形の文書集合共通度、文書集合の共通話題と各文書・文との近さを表す文書（文） 文書集合共通度を提案し、簡単な実験を通じて妥当性を検証した。残された課題は、より大量のデータを用いて本手法の能力と限界の把握、及び本手法が適した応用を見出すことである。

参考文献

- [1] 奥村学, 難波英嗣. テキスト自動要約に関する

表 2 最も共通度の高い 3 つの文

1 位 : General Partners said it is asking GenCorp for its shareholder lists for help in disseminating the offer.
 2 位 : Earlier today, General Partners, owned by Wagner and Brown and AFG Industries Inc, launched a 100 dlr per share tender offer for GenCorp.
 3 位 : General Acquisition Co said it was disappointed by Gencorp's response to its tender offer and asked how the company might give better value to shareholders.

- 研究動向 . 自然言語処理 , 6 , 6 , pp.1-26 , 1999 .
 [2] 奥村学 , 難波英嗣 . テキスト自動要約に関する最近の話題 . 自然言語処理 , 9 , 4 , pp.97-116 , 2002 .
 [3] H. Zha. Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In *Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.113-120. Tampere, Finland, August, 2002.
 [4] H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. B. Wise and X. Zhang. Cross-Document Summarization by Concept Classification. In *Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.121-128. Tampere, Finland, August, 2002.
 [5] C. D Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.
 [6] 川谷隆彦. 文ベクトル集合モデルによるテキスト処理 . 情報処理学会自然言語処理研究報告 , 2000-NL-140 , pp.31-38(2000) .
 [7] 川谷隆彦. 文ベクトル集合モデルによるテキスト処理(II) . 情報処理学会自然言語処理研究報告 , 2001-NL-143 , pp.1-8(2001) .
 [8] J. Carbonell, Y. Geng and J. Goldstein. Automated Query-Relevant Summarization and Diversity-Based Reranking. In *Proceedings of the IJCAI-97 Workshop on AI in Digital Libraries*, pp.9-14.