多文書間の共通性分析に基づく文書クラスタリング

川 谷 隆 彦†

本論文では多文書間の共通性分析に基づく非階層的な文書クラスタリング法を提案する．文書クラスタリングにおいては，同じ話題を有する文書がグループ化されるので同じクラスタに属する文書にはなんらかの共通性が存在するはずである．また各話題には特有の単語や単語対が存在する．提案手法ではこのような点に着目し，文書・クラスタ間の類似度を，対象文書とその時点のクラスタに含まれる文書の共通情報との間で，単語の生起情報ばかりでなく共起情報も用いて定義する．また，話題特有の単語や単語対を用いて類似度を算出し，複数の話題に共通する情報の影響を排除する．提案手法ではクラスタは 1 つずつ検出され，しかるべき方法で抽出された種文書と同じ話題の文書をマージさせつつ順次クラスタを成長させるという処理が繰り返される．TDT2 のコーパスから選択した 21 イベント 6,788 文書，31 イベント 7,306 文書，38 イベント 7,546 文書のそれぞれに対し，検出クラスタ数 21，30，36，クラスタリング精度 95.17%，95.09%，94.82%を得た．また，上記の 38 イベント 7,546 文書に対する kNN（教師ありの分類法）の分類精度は 97.02%であり，提案手法は教師なしでありながら，教師ありの分類手法に近い精度が得られることが確認された．

Document Clustering Based on Commonality Analysis among Multiple Documents

Takahiko Kawatani†

This paper proposes a flat clustering method based on multi-document commonality analysis. In document clustering, documents with the same topic are grouped into a cluster so that documents in the cluster have certain commonalities. Furthermore, any topic has its own specific terms and term-pairs. Based on these aspects, the proposed method defines the document-cluster similarity between the given document and common information among the documents in the cluster. The similarity features that it uses not only term occurrence information but also term co-occurrence information. The similarity is obtained using specific terms and term-pairs of the cluster to avoid any impact from terms and term-pairs shared by two or more topics. The cluster seed grows by merging documents with high similarity into the current cluster. Through experiments using TDT2 as a corpus, it was confirmed that a proper number of clusters is obtained and that documents are assigned to clusters with high accuracy.

1. まえがき

文書クラスタリングは，インターネットを通じて行き交う情報が増大するにつれ，情報検索[1]，文書要約[2),3)]，文書コレクションのブラウジングやナビゲーション[4),5)]，トピック検出・追跡（TDT: Topic Detection and Tracking）などの分野で重要性がますます高くなってきている．文書クラスタリングは，入力文書集合に含まれる話題に応じてクラスタを検出し，各文書をしかるべきクラスタに配置する技術であるから，その性能は，

- 各話題に対応してクラスタを検出し，クラスタの数を正しく得る能力
- 文書を正しいクラスタに配置する能力

によって測られる．クラスタリング技術としてはこれまで様々な方法が提案されてきているが，上記の能力に関して完璧とはいい難く，後述のように改善の余地が多分に残されている．本論文では上記能力の両方について同時に改善を図る方法を提案する．提案手法はクラスタを 1 つずつ検出するという戦略に基づくものであり，以下のような手順を踏む．

ステップ 1：繰返し処理の初回では全文書から，2 回目以降ではその時点のどのクラスタに対しても類似度が一定値以下の文書（残存文書）の中から，

† メディアドライブ株式会社
Media Drive Corporation

本研究は，日本ヒューレット・パッカード（株）ヒューレット・パッカード研究所にて行われたものである．

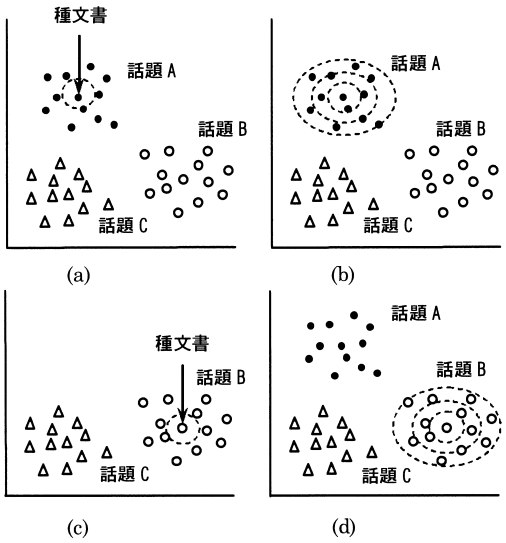

図 1 種文書の検出とクラスタの成長
Fig. 1 Detection of cluster seed and growth of a cluster.

クラスタの種の候補となる文書を複数検出する．

ステップ 2：入力文書集合全体から各候補文書と類似度が一定値以上の文書を近隣文書として求め，近隣文書の最も多い文書を種文書とする．

ステップ 3：種文書の近隣文書を初期クラスタとし，入力文書集合全体の中でカレントクラスタと類似度の高い文書をマージしてクラスタを成長させるという処理を，クラスタ内の文書数が増えなくなるまで繰り返す．

ステップ 4：1 つのクラスタの成長が終わった時点で終了条件をチェックし，満たされれば処理を終了する．満たされない場合にはステップ 1 に戻る．

図 1 はステップ 3 までのクラスタリングの過程を図示するものであり，単語空間上に分布する話題 A，B，C の文書を対象としている．図 1 (a) は，ステップ 1，2 の過程を経て，話題 A の文書の 1 つが種文書として選択され，その周りに初期クラスタが形成された状況を示している．図 1 (b) は，図 1 (a) における初期クラスタが順次成長し，話題 A の全文書をカバーして成長が終わった様子を示している．図 1 (c) では，ステップ 1，2 に立ち戻り，全体の文書集合から図 1 (b) で求められたクラスタに属する文書を除いて残存文書が求められた様子と，残存文書から種文書（図では話題 B）が決定され，その周りに初期クラスタが形成された様子とを示している．ただし，この場合の初期クラスタのメンバは全文書集合から選択される．図 1 (d) は，図 1 (c) の初期クラスタが全文書を相手に成長していく様子を示している．このように，提案手法では残存文書から種文書が抽出されるので，種文書は検出済みのクラスタに含まれる文書とは話題が異なるようになる．また，クラスタは全文書を相手に成長するので，たとえその時点までに検出済みのクラスタに含まれる文書であっても，カレントクラスタの種文書と同じ話題を含んでいれば，カレントクラスタにも帰属するようになる．その結果，入力文書が複数の話題を含んでいても，各文書をその文書が含むすべての話題に対応するクラスタに帰属させる非排他的なクラスタリングが可能である．排他的なクラスタリングを指向する場合には，ステップ 4 の処理の後で各文書を類似度が最も高いクラスタに帰属させればよい．

このような処理においては，各クラスタは対応する話題の文書を過不足なくカバーし，異なる話題の文書を含まないように成長することが要求される．このような要求が満たされていれば，各文書を正しいクラスタに配置する能力は高くなっているはずである．また，各クラスタを 1 つずつ正しく検出することが可能になるので，クラスタの検出能力も高められる．このような要求条件を満たすには文書・クラスタ間類似度をなるべく正確に求める必要がある．文書・クラスタ間類似度が不正確になる要因としては，

A) 複数の話題で共有される単語の影響
B) 着目クラスタに誤配置された少数の文書の影響

が考えられる．本論文では，

- 文書クラスタリングにおいては同じ話題を有する文書がグループ化されるので，同じクラスタに属する文書には何らかの共通性が存在する，
- 各話題にはその話題の文書には出現するが，他の話題の文書にはほとんど出現しない特有の単語や単語対が存在する，

という前提のもとに，上記 A)，B) の影響を受けにくい文書・クラスタ間類似度を提案する．この文書・クラスタ間類似度は着目文書とクラスタに含まれる文書の共通情報との間で定義され，単語の生起情報ばかりでなく共起情報も用いることを特徴としている．

以下，2 章では，従来手法のレビューを行い提案手法と対比する．3 章では，多文書間の共通性分析に基づく文書間類似度，文書・クラスタ間類似度について述べる．4 章ではクラスタリングの手順について述べる．5 章では TDT2 コーパスを用いた実験結果を紹介し，6 章で既存手法との比較，本論文で提案された類似度の正確さ，パラメータの影響に関する考察を行う．最後に，7 章で全体をまとめる．

2. 関連研究

文書クラスタリング技術には数十年に近い研究の歴史があり，様々な方法が提案されてきているが，AHC と呼ばれる Agglomerative Hierarchical Clustering[14]，k-means[14]，シングルパスがよく知られている．これらの中では AHC が多く用いられているといわれている[8]．AHC は階層的なクラスタリング法であり，初期状態では各文書は 1 つのクラスタを構成すると見なされ，後続の逐次的な処理において最も近いクラスタどうしのマージが，クラスタ数が 1 になるまで，もしくは一定の停止条件が満たされるまで繰り返される．k-means は非階層的なクラスタリング法であり，まず，ユーザにより検出すべきクラスタ数として指定された数の文書を初期クラスタとしてランダムに選択する．次いで，各クラスタの重心の算出，各クラスタの重心と各文書の距離の算出，最も近いクラスタへの各文書の配置という一連の処理が，各クラスタの構成メンバに変化がなくなるまで繰り返される．ストリームの形で文書が入力される TDT タスクでは，シングルパスが多く用いられている[6]．この方法では，新しい文書が入力されると，まずその時点で存在するクラスタの各々との類似度が求められる．そして，ある入力文書と類似度が一定値以上のクラスタがあれば入力文書はそのクラスタに配置され，なければ入力文書は新しいクラスタとされる．

まず，クラスタ検出能力に関して述べる．AHC において正しい数のクラスタを求めるには処理を途中で打ち切る必要がある．文献 10) にはいくつかの打ち切り方法が述べられているが，それらではクラスタ数が一定数になったときに処理を打ち切るという方法が採用されている．これらの方法では打ち切り条件が適正でない場合，複数のクラスタがマージされて意味のないクラスタが構成されてしまう．k-means ではユーザはクラスタ数を正しく指定する必要があるが，これは入力文書集合に対する事前知識なしには不可能な問題である．Liu らは，非階層的クラスタリングのクラスタ検出能力の改善を図るため，クラスタ数を変えながら得られた複数の結果からもっともらしい結果を選ぶことにより正しい数のクラスタを得る方法を提案し，TDT2 コーパスを用いて実験を行っている[11]．実験結果では，TDT2 からランダムに選択された 2〜6 のイベントを含む 12 通りのデータセットに対し，正しい数のクラスタが得られたのは 9 個であり，つねに正しい数のクラスタが求められているとは限らない．このような状況から，AHC や k-means のように複数のクラスタを同時並行的に求める方法では正しい数のクラスタを得るのは困難と思われる．これが，クラスタを 1 つずつ検出するという戦略を採用した理由である．既存方法の中で，文書がクラスタに成長するという点でシングルパスはクラスタを 1 つずつ検出しているといえる．しかし，シングルパスでは入力文書集合に対して 1 方向かつ 1 度のスキャンを行ってクラスタを求めているので，クラスタリングの結果は提示する文書の順番の影響を受けるといわれている．シングルパスはオンラインのクラスタリングには適するにしても，オフラインのクラスタリングには適さないように思われる．

文書・クラスタ間類似度尺度に関していえば，AHC やシングルパスではシングルリンク法，完全リンク法，グループ平均法が多く用いられている．これらは 2 文書間の類似度をベースとするものであり，クラスタ内の文書の共通情報を反映するものではない．クラスタリングの途中段階のクラスタにおいて，大多数の文書の話題とは異なる話題の少数の文書が誤って含まれてしまうことがある．そのような場合，上記の尺度を用いると，そのクラスタは話題の異なる文書ともゼロでない類似度を有してしまい，後続の処理で異なる話題の文書の混入を招きやすくなる．一方，k-means では文書とクラスタの重心とのユークリッド距離などが類似度尺度として多く用いられている．クラスタの重心はクラスタ内の各文書の平均として求められるが，これはクラスタの共通情報とは異なるものである．また，AHC や k-means などでは単語選択は通常行われていないので，文書・クラスタ間類似度は複数の話題に共有される単語の影響を受け，同じ単語を共有する他の話題の文書の排除を難しくする．Liu らは，非階層的クラスタリング処理を行ったのち各クラスタに特有な単語を求め，特有単語の存在個数によりクラスタリングの結果をリファインする処理を行って，精度を向上させている[11]．この方法は，各クラスタに特有でない単語の影響の排除という点で提案手法と共通しているが，提案手法ではクラスタリングの過程でこの処理を行う点が異なっている．また，Dharanipragada らは，着目単語を含むクラスタ内の文書数と，着目単語を含む全文書数とクラスタ内文書数の和との比で決まる値を，Okapi における各単語の重みに加えることにより，各単語のクラスタ特有性を反映させている[12]．しかし，この方法では，特有でありながら頻度が低い単語に大きな重みが与えられないという問題があるように思われる．

3. 共通性分析と文書・クラスタ間類似度

本章では以下を目的に多文書間の共通性分析に関する議論を行い，文書・クラスタ間類似度の新しい尺度を提案する．

i. 与えられた文書集合における各文書の話題がどの程度共通しているか数値（共通度）で示す．
ii. 文書集合の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．

3.1 共通性分析のアプローチ

R 個の文書からなる文書集合 D を考える．いま，各文書から 1 つずつ文を取り出して R 個の文からなる文の組を作ったとする．このような文の組は各文書の文の数の積通り存在する．着目する文の組において R 個の文のすべてに現れる単語を共通単語，共通単語が構成する文を共通文と呼ぶ．共通文の集合は文書集合 D の共通の内容を表すと考えられる．さらに各共通文で共通単語数を求め，その和もしくは 2 乗和を求めると，その値は各文書が共通に有する情報の量に応じていると考えられる．そこでこれらの値を，文書数や各文書のサイズで正規化して共通度を定義することとする．

3.2 文書および共通文集合の単語共起行列

次のような文書集合 D を考える．

$\{w_1, \ldots, w_M\}$：集合 D に出現する M 個の単語の集合
D^r：Y_r 個の文からなる r 番目の文書
D^r_y：文書 D^r の y 番目の文
$\boldsymbol{d}^r_y = (d^r_{y1}, \ldots, d^r_{yM})^T$：文 D^r_y のバイナリベクトル．d^r_{ym} は m 番目の単語の有無を表す．T は転置である．

文書 D^r に対し行列 $\boldsymbol{S}^r$ を次式で定義する．

$$\boldsymbol{S}^r = \sum_{y=1}^{Y_r} \boldsymbol{d}^r_y \boldsymbol{d}^{r\,T}_y, \tag{1}$$

式 (1) から分かるように，$\boldsymbol{S}^r$ の mn 成分は $S^r_{mn} = \sum_{y=1}^{Y_r} d^r_{ym} d^r_{yn}$ により与えられる．したがって，S^r_{mm} は文書 D^r において単語 m が出現する文の数，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 の単語共起行列と呼ぶこととする．同じ単語は同じ文で 2 回以上現れないと仮定すると，単語共起行列には次のような性質がある．

(1) 対角成分 S^r_{mm} は単語 m の文書 D^r における出現頻度と等しい．また，対角成分の和は文書 D^r 中のすべての単語の出現頻度の総和に等しく，したがって各文の単語数（各文の長さ）の和，すなわち文書の長さとも等しくなる．

(2) S^r の全成分の和は文書 D^r の各文の長さの 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^r_y とすると，下記により示される．

$$\begin{aligned}\sum_{y=1}^{Y_r} (f^r_y)^2 &= \sum_{y=1}^{Y_r} (d^r_{y1} + \cdots + d^r_{yM})^2 \\ &= \sum_{y=1}^{Y_r} \sum_{m=1}^{M} \sum_{n=1}^{M} d^r_{ym} d^r_{yn} \\ &= \sum_{m=1}^{M} \sum_{n=1}^{M} S^r_{mn} \end{aligned} \tag{2}$$

次に共通文集合の単語共起行列を求める．簡単な例として 3 文書 D^1，D^2，D^3 からなる文書集合を考える．D^1，D^2，D^3 のそれぞれの i，j，k 番目のベクトル $\boldsymbol{d}^1_i$，$\boldsymbol{d}^2_j$，$\boldsymbol{d}^3_k$ の共通文ベクトルを $\boldsymbol{c}^{ijk} = (c^{ijk}_m)$ とすると，c^{ijk}_m は

$$c^{ijk}_m = d^1_{im} d^2_{jm} d^3_{km} \tag{3}$$

により表すことができる．共通文集合の単語共起行列を $\boldsymbol{T} = (T_{mn})$ とすると，T_{mn} は次のように求められる．

$$\begin{aligned} T_{mn} &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} c^{ijk}_m c^{ijk}_n \\ &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} d^1_{im} d^1_{in} d^2_{jm} d^2_{jn} \\ &\quad \cdot d^3_{km} d^3_{kn} \\ &= S^1_{mn} S^2_{mn} S^3_{mn}. \end{aligned} \tag{4}$$

T_{mn} は文書 D^1，D^2，D^3 の単語共起行列の対応する成分どうしの積として求められている．これは文書の数とは無関係に成り立ち，文書数が R の場合は

$$T_{mn} = \textstyle\prod_{r=1}^{R} S^r_{mn} \tag{5}$$

で与えられる．結局，共通文ベクトル集合の単語共起行列は，共通文ベクトルを求めることなく得ることができる．行列 $\boldsymbol{T}$ も単語共起行列である以上，文書の単語共起行列と同じ性格を持つ．したがって，T_{mm} と T_{mn} はそれぞれ共通文集合における単語 m の生起回数，単語 m と n の共起回数を表すこととなる．また，行列 $\boldsymbol{T}$ の対角成分の総和は各共通文の共通単語数の総和，全成分の総和は各共通文の共通単語数の 2 乗和となる．

表 1，**図 2** は上記を例示している．表 1 (a) は単語総数を 5 とし，文書 1，2 が 2 個の文，文書 3 が 3 個の文からなるとして，各文書の文ベクトルを表している．表 1 (b) は各文の組に対して共通文ベクトルと共

表 1 文書の例とその共通文集合
Table 1 Examples of documents and their common sentences.

(a) 文書の例

文書	文	単語 1	2	3	4	5
1	1	1	1	0	1	1
	2	1	0	1	0	1
2	1	1	1	0	1	0
	2	0	0	1	0	1
3	1	0	1	1	1	0
	2	1	1	0	1	1
	3	0	1	1	0	1

(b) 共通文の集合

組み合わせ	単語 1	2	3	4	5	Σc_i	$(\Sigma c_i)^2$
1-1-1	0	1	0	1	0	2	4
1-1-2	1	1	0	1	0	3	9
1-1-3	0	1	0	0	0	1	1
1-2-1	0	0	0	0	0	0	0
1-2-2	0	0	0	0	1	1	1
1-2-3	0	0	0	0	1	1	1
2-1-1	0	0	0	0	0	0	0
2-1-2	1	0	0	0	0	1	1
2-1-3	0	0	0	0	0	0	0
2-2-1	0	0	1	0	0	1	1
2-2-2	0	0	0	0	1	1	1
2-2-3	0	0	1	0	1	2	4
合計	2	3	2	2	4	13	23

$$\begin{bmatrix} 2&1&1&1&2\\ 1&1&0&1&1\\ 1&0&1&0&1\\ 1&1&0&1&1\\ 2&1&1&1&2 \end{bmatrix} \begin{bmatrix} 1&1&0&1&0\\ 1&1&0&1&0\\ 0&0&1&0&1\\ 1&1&0&1&0\\ 0&0&1&0&1 \end{bmatrix} \begin{bmatrix} 1&1&0&1&1\\ 1&3&2&2&2\\ 0&2&2&1&1\\ 1&2&1&2&1\\ 1&2&1&1&2 \end{bmatrix} \begin{bmatrix} 2&1&0&1&0\\ 1&3&0&2&0\\ 0&0&2&0&1\\ 1&2&0&2&0\\ 0&0&1&0&4 \end{bmatrix}$$

文書 1　　文書 2　　文書 3　　共通文集合

図 2 文書 1，2，3 および共通文集合の共起行列
Fig. 2 Co-occurrence matrices of document 1, 2, 3 and common sentence sets.

通単語数 $\sum c_i$，およびその 2 乗値 $(\sum c_i)^2$ を示す．たとえば，組合せ 2-2-3 は文書 1，2 の 2 番目，文書 3 の 3 番目の文の組合せを示す．この場合，この 3 個の文に共通する単語は 3 と 5 であり，共通文ベクトルでは 3，5 番目の成分のみが 1 となる．また，共通単語数は 2，その 2 乗値は 4 となる．図 2 は表 1 の文書 1，2，3，および共通文集合に対する単語共起行列を示す．図 2 より，共通文集合の単語共起行列の各成分は各文書の単語共起行列の対応する成分どうしの積であることが分かる．また，共通文の単語共起行列の対角成分の和は 13，全成分の和は 23 となるが，これらは表 1 の各共通文の共通単語数の総和，2 乗和と等しい．

3.3 共 通 度

前述のように，各共通文の共通単語数の総和，あるいは 2 乗和は各文書が共通に有する情報の量と見なされるので，文書数や各文書のサイズで正規化して共通度を定義することができる．共通単語数の総和をベースとする場合を線形モデル，2 乗和をベースとする場合を 2 次モデルと呼ぶこととする．$com_l(D)$，$com_q(D)$ をそれぞれ線形モデル，2 次モデルにおける共通度と定義すると，これらは以下のように定義できる．

$$com_l(D) = \left[\frac{\sum_{m=1}^{M} T_{mm}}{\sqrt[R]{\prod_{r=1}^{R} \sum_{m=1}^{M} (S_{mm}^r)^R}} \right]^{1/(R-1)} \tag{6}$$

$$com_q(D) = \left[\frac{\sum_{m=1}^{M} \sum_{n=1}^{M} T_{mn}}{\sqrt[R]{\prod_{r=1}^{R} \sum_{m=1}^{M} \sum_{n=1}^{M} (S_{mn}^r)^R}} \right]^{1/(R-1)} \tag{7}$$

正規化は R 個の文書が同一の場合に共通度が 1.0 になるように行われる．また，式 (6)，(7) では $(R-1)$ 乗根が求められているが，これは文書数が R 個のときには $(R-1)$ 回の文書の突き合わせが行われているからである．R 個の文書が互いに異なる場合，$(R-1)$ 乗根を求めないと R が大きいほど共通度の値は小さくなってしまう．これらの正規化により，文書数の異なる複数の文書集合の共通度の比較が可能となる．

3.4 共通度と文書間類似度

文書数が 2 の場合（$R=2$）を考える．2 文書間の共通度は文書間類似度そのものと見なすことができる．この場合，式 (6)，(7) は以下のように書き直すことができる．

$$com_l(D) = \frac{\sum_{m=1}^{M} S_{mm}^1 S_{mm}^2}{\sqrt{\sum_{m=1}^{M} (S_{mm}^1)^2} \sqrt{\sum_{m=1}^{M} (S_{mm}^2)^2}} \tag{8}$$

$$com_q(D) = \frac{\sum_{m=1}^{M} \sum_{n=1}^{M} S_{mn}^1 S_{mn}^2}{\sqrt{\sum_{m=1}^{M} \sum_{n=1}^{M} (S_{mn}^1)^2} \sqrt{\sum_{m=1}^{M} \sum_{n=1}^{M} (S_{mn}^2)^2}} \tag{9}$$

3.1 節で述べたように，同じ単語が同じ文に 2 回以上現れないと仮定すると，各文書の単語共起行列の対角成分は対応する単語の出現頻度を表す．したがって，式 (8) で与えられる文書間類似度は文書内の単語頻度を成分とする文書ベクトルの余弦類似度，すなわちベクトル空間モデルにおける類似度とまったく同じとなる．

また，式 (9) は 2 文書間の類似度はそれぞれの文書

の単語共起行列の対応する成分どうしの積和をもとに与えられることを示している．この場合，2 つの文書の類似度が高いためには，2 つの文書の間で単語の出現傾向だけではなく，単語共起の傾向まで似ている必要がある．さらに，2 つの文 D_i^1，D_j^2 の共通単語数は ${\boldsymbol d_i^1}^T \boldsymbol d_j^2$ と表すことができるので，共通単語数の 2 乗和は $\sum_{i=1}^{Y_1}\sum_{j=1}^{Y_2}\left({\boldsymbol d_i^1}^T \boldsymbol d_j^2\right)^2$ によっても表される．したがって，式 (9) は

$$com_q(D) = \frac{\sum_{i=1}^{Y_1}\sum_{j=1}^{Y_2}\left({\boldsymbol d_i^1}^T \boldsymbol d_j^2\right)^2}{\sqrt{\sum_{i=1}^{Y_1}\sum_{j=1}^{Y_1}\left({\boldsymbol d_i^1}^T \boldsymbol d_j^1\right)^2}\sqrt{\sum_{i=1}^{Y_2}\sum_{j=1}^{Y_2}\left({\boldsymbol d_i^2}^T \boldsymbol d_j^2\right)^2}} \quad (10)$$

と表すこともできる．式 (10) は対象となる文書間のすべての文ベクトルの組合せから求められる内積の 2 乗和をベースにしている．

3.5 文書・クラスタ間類似度

文書・クラスタ間類似度は着目文書とクラスタ内文書集合の共通文集合との類似度として定義することができる．着目文書を P，クラスタ内文書集合を D とすると，線形モデルの文書・クラスタ間類似度 $sim_l(D,P)$，および 2 次モデルのそれ $sim_q(D,P)$ は式 (8)，(9) を適用することにより次のように表される．

$$sim_l(D,P) = \frac{\sum_{m=1}^{M} z_{mm}T_{mm}S_{mm}^P}{\sqrt{\sum_{m=1}^{M}(z_{mm}T_{mm})^2}\sqrt{\sum_{m=1}^{M}(S_{mm}^P)^2}}, \quad (11)$$

$$sim_q(D,P) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{M} z_{mn}T_{mn}S_{mn}^P}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M}(z_{mn}T_{mn})^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M}(S_{mn}^P)^2}} \quad (12)$$

ここで，

$\boldsymbol S^P$：着目文書 P の単語共起行列

z_{mm}，z_{mn}：それぞれ単語 m，および単語 m，n の対に対する重み

である．以下，対角成分，非対角成分に対する重みをそれぞれ単に z_{mm}，z_{mn} と表記する．

3.6 変　　形

上記により原理的には文書・クラスタ間類似度を定義することができるが，実際にはサイズの大きい文書集合の場合，次のような問題がある．

i. 話題が同じであってもすべての文書に出現する単語が存在するとは限らない．
ii. 共通文集合の単語共起行列の各成分は各文書の単語共起行列の対応する成分どうしの積で求められるので，極端に大きな値をとる場合がある．

これらの問題に対し次のように対処する．まず，i の問題に対しては，u_m を単語 m の出現する文書数（文書頻度）として，以下のように便宜的に求められる行列 $\boldsymbol T^A$ を $\boldsymbol T$ の代わりに用いる．

$$T_{mn}^A = \prod_{\substack{r=1 \\ S_{mn}^r>0}}^{R} S_{mn}^r \quad \text{if } u_m > A \text{ and } u_n > A \quad (13\text{-a})$$

$$T_{mn}^A = 0 \quad \text{otherwise} \quad (13\text{-b})$$

T_{mn}^A は文書頻度が A 以上の 2 つの単語 m，n が共起する文書（$m=n$ のときは単語 m が生起する文書）のみを用いて求められる．この結果，全文書に出現する単語ではなく，文書頻度が A 以上の単語を用いて類似度が算出されるので，問題 i に対処することが可能となる．

問題 ii は単語共起行列 $\boldsymbol T^A$ の各成分の値域を小さくすることで対処する．そのため，まず行列 $\boldsymbol Q^A = (Q_{mn}^A)$ を次式で定義する．

$$Q_{mn}^A = \log(T_{mn}^A) \quad \text{if } T_{mn}^A > 1$$
$$= 0 \quad \text{otherwise} \quad (14)$$

式 (11)，(12) において $\boldsymbol T$ の代わりに $\boldsymbol Q^A$ を用いることにより，類似度は次のように定義される．

$$sim_l(D,P) = \frac{\sum_{m=1}^{M} z_{mm}Q_{mm}^A S_{mm}^P}{\sqrt{\sum_{m=1}^{M}(z_{mm}Q_{mm}^A)^2}\sqrt{\sum_{m=1}^{M}(S_{mm}^P)^2}} \quad (15)$$

$$sim_q(D,P) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{M} z_{mn}Q_{mn}^A S_{mn}^P}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M}(z_{mn}Q_{mn}^A)^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M}(S_{mn}^P)^2}} \quad (16)$$

上記 A の値は得られるクラスタの話題の揃い方を規定するパラメータとなる．A の値を大きくとると得られたクラスタの話題は揃うことになり，クラスタリングの途中段階で他の話題の文書が雑音として紛れ込む可能性を抑えることができる．しかし，A の値を必要以上に大きくとると，クラスタが十分に成長せず，狭い話題のサブクラスタに分割されてしまうことも考えられる．したがって，A の値は各話題の広がりの程

度に応じて適切な値を設定する必要がある．具体的には，訓練データを用いて A の値を変えながら実験を行い，クラスタの話題の揃い方を観察しながら決定することになる．

4. クラスタリングの方法

4.1 手順の詳細

1 章で述べたクラスタリングの各処理ステップ，および他に必要となる処理について詳細を述べる．

ステップ 1：クラスタの種の複数の候補文書を，初めての処理のときには全文書から，2 回目以降の繰返しのときには残存文書から検出する．

具体的には，まず異なる話題の文書は同じ単語対を共有しないと仮定する．このような仮定の下では，ある文書集合と，単語共起を用いた式 (16) による類似度の大きい文書はその文書集合の中の優勢な話題を有し，かつ同じ話題の文書集合の中でも中心的な文書となると考えられる．そこで式 (16) を用いて全文書集合もしくは残存文書集合とその構成メンバとの類似度を求め，値の大きな一定数の文書を種文書候補とする．ただし，複数の話題で共有される単語の影響を排除するため，z_{mm} $(m = 1, \ldots, M)$ は 0 とする．

ステップ 2：クラスタの種の各候補文書につきすべての文書との類似度（式 (8) もしくは (9)）を求め，一定値以上の類似度を有する文書を近隣文書として抽出する．近隣文書集合は同じ話題を述べた文書から構成されていると見なすことができる．近隣文書の多い文書ほど話題の中心を記述していると考えられるので，近隣文書数が最も多くなる文書をクラスタの種とする．

ステップ 3：種文書の近隣文書集合を初期クラスタとし，その時点のクラスタ文書集合から特有単語・単語対を検出する（後述）．次いで，特有単語・単語対を選択的に用いて各文書との間で類似度（式 (15) もしくは式 (16)）を求め，一定値以上の類似度を有する文書をそのクラスタに仮に帰属させることによりクラスタを成長させる．クラスタに仮に帰属する文書数が一定になれば，全文書に対する類似度を保持してステップ 4 へ．そうでなければ本ステップを繰り返す．

ステップ 4：終了条件（後述）を満たせばステップ 5 へ．そうでなければステップ 1 に戻って続行．

ステップ 5：クラスタ間の重なりをチェックし，重なりの大きいクラスタがあれば冗長なクラスタを検出して削除する（後述）．

ステップ 6：各文書と残された全クラスタとの類似度をもとに，各クラスタに帰属する文書を最終的に決定する（後述）．

4.2 終了条件

ステップ 3 の終了後，残存文書集合を求めたとする．クラスタリングが正常に終了していれば残存文書は理想的には存在しないはずである．そこで，各話題から取り残される文書の存在も考慮して，残存文書数が一定値以下となれば終了という条件が考えられる．しかし，どの話題にも含まれない不詳数の例外的な文書が含まれるような場合は不詳数の残存文書が存在することになり，上記終了条件は不適当である．そこで，ステップ 4 では前回のクラスタ検出からの残存文書の減少数を求め，一定値以下の残存文書の減少が 2 回連続した場合に終了するようにした．ただし，最終回に得られたクラスタは採用しない．残存文書数の減少数が一定値以下になればただちに終了としなかったのは，クラスタがすべて検出される前に冗長クラスタ（後述）が検出された場合，冗長クラスタのときには残存文書は減少しないのでそのまま処理が終了してしまうからである．上記一定値が小さいと冗長クラスタが発生することが多くなり，また大きいと小さなクラスタが未検出に終わる場合がある．上記一定値は実験的に 10 と決めた．なお，冗長クラスタの 2 回連続検出の場合も全クラスタ検出の前に処理が終了しうるが，実際にはそのような事態は起こらなかった．

4.3 特有単語・単語対検出

$\boldsymbol{U}^0$，$\boldsymbol{U}$ を入力文書集合全体，その時点での着目クラスタ文書集合から求められた文書頻度行列とする．すなわち，

U^0_{mm}，U_{mm}：それぞれ，入力文書集合全体，および着目クラスタ文書集合で単語 m の生起する文書数

U^0_{mn}，U_{mn}：それぞれ，入力文書集合全体，および着目クラスタ文書集合で単語 m，n の共起する文書数

である．したがって，U_{mm}/U^0_{mm} もしくは U_{mn}/U^0_{mn} の値は単語 m，もしくは単語 m，n の対の着目クラスタ文書集合への集中の度合い（集中度）を表すことになる．特有単語・単語対の集中度は，着目クラスタが対応する話題の文書を過不足なくカバーするように成長した場合には，1.0 に近い値をとるはずである．また，クラスタの成長過程でも非特有単語・単語対よりも大きい値をとる．したがって，集中度が閾値以上の単語，単語対を選択することにより，着目クラスタに特有な単語，単語対を検出することができる．具体的には，着目クラスタ文書集合における頻度の高い一定数（実験では 30 個）の単語のうち，U_{mm}/U^0_{mm} の値の大きい一定数（同 5 個）は着目クラスタに特有と仮定したうえで，それらの平均を特有単語の平均集中度

として求め，平均集中度の $1/\alpha$ 倍を特有単語検出の閾値とした．同様に，頻度の高い一定数（同 100 個）の単語対のうち U_{mn}/U^0_{mn} の値の大きい一定数（同 50 個）は特有と仮定し，それらの集中度の平均値の $1/\alpha$ 倍を特有単語対検出の閾値とした．式 (15)，(16) において特有とされなかった単語 m，単語 m，n の対に対しては，$z_{mm} = 0$，$z_{mn} = 0$ とされる．上記一定数の値や α は，抽出された特有単語や単語対の妥当性を判断して実験的に決定した．

4.4 冗長クラスタの検出と除去

まず，冗長クラスタがどのように生ずるかを見る．最初の例は**図 3** (a) に示される場合である．図のように，2 つのクラスタがほぼ重なってしまう場合は一方が冗長となり，取除く必要がある．すなわち，冗長なクラスタへの文書の配置は行わないようにする．このようなケースは，あるクラスタの成長が終わったときに，そのクラスタに関連する話題の文書が誤って残存文書に取り残されてしまい，後続の処理で取り残された文書が種文書として検出され，クラスタとして順調に成長したというような場合に起こる．

また，図 3 (b) のように，他のクラスタに完全に包含されたクラスタも冗長となる．これは，ある話題の種文書が順調に成長しなかった結果，その話題の文書の多くが残存文書集合に取り残され，後続の処理で残された文書から検出された種文書がクラスタとして順調に成長したというような場合に起こる．

正しい数のクラスタを求めるには，上述のような冗長クラスタの除去は必須である．冗長クラスタの検出のためには，まず，各クラスタに対し，そのクラスタに対してのみ類似度が一定値よりも大きくなる文書数をクラスタ重要度として求める．クラスタ重要度が一定値よりも小さいクラスタが 1 つ存在する場合は無条件に除去する．複数存在すれば，クラスタ重要度が最も小さいクラスタをまず除去する．このような処理を冗長なクラスタが存在しなくなるまで繰り返す．冗長なクラスタの重要度は通常 0 もしくは 0 に近い値となるので，上記一定値は実験では 5 とした．

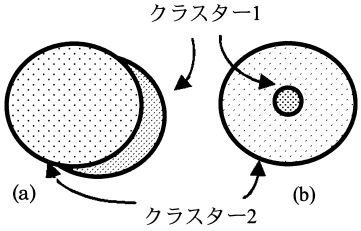

図 3 冗長クラスタの例
Fig. 3 Examples of redundant clusters.

4.5 各クラスタ帰属する文書の決定

ステップ 6 において，たとえば，排他的なクラスタリングを指向する場合には，各文書はその文書と最も類似度の高いクラスタに帰属させる．また，非排他的なクラスタリングを指向する場合には，各文書はその文書との類似度が一定値以上のクラスタに帰属させる．前者の場合，どのクラスタの特有単語，単語対も含まれない結果，どのクラスタとも類似度がゼロとなる文書を除いて，各文書はいずれかのクラスタに属することになる．後者の場合，どのクラスタとも類似度が一定値以下の文書はどのクラスタにも属さないことになる．

5. 実　　験

5.1 実験データ

用いたコーパスは TDT2 である[7]．TDT2 は 1998 年の 1 月から 6 月の間の 100 個のイベント（たとえば，“Asian Economic Crisis”，“Monica Lewinsky Case”，“Current Conflict with Iraq” など）に関するニュースストーリーの集合であり，放送系のニュースソース “ABC”，“CNN”，“VOA”，および電子ニュース系の “NYT”，“APW”，“PRI” から採取されている．各文書にはどのイベントを述べているかを示すラベルが付与されている．実験では，イベント数が約 20，30，40 となるように，文書数が 70，40，30 以上のイベントを選択し，**表 2** に示すような 3 種類のデータを作成した．実際には，21，31，38 イベントが選択された．イベントあたりの最大文書数は 1,484 で，データ 3 では最小文書数の約 50 倍である．したがって，文書数の小さなイベントがどれだけ正確に検出できるかどうかが鍵となる．また，データ 1，2，3 を通して，文書あたりの文の数は，平均 15.8，最大 157，最小 1，抽出された単語種類数は平均 123，最大 861，最小 6 であった．なお，Liu らの実験[11] では，全部で 15 のイベントから選択した 3〜9 個を組み合わせてデータセット（27 通り）を作成している．また，各データセットにおけるイベントの文書数の最大と最小の比は，

表 2 実験に用いたデータセット
Table 2 Data used in the experiments.

データ	イベント数	文書数	イベント当りの文書数		
			平均	最大	最小
1	21	6788	323.2	1484	70
2	31	7306	235.7	1484	41
3	38	7546	198.6	1484	30

21.5 を唯一の例外としてそのほかはすべて 10 以下である．本論文で実験に用いたデータセットは，Liu らが用いたものに比べクラスタリングが難しくなるように作成したものである．

5.2 実験条件

前処理としては，文切り出しの後，品詞付け，各単語の活用の基本形への変換（lemmatization），ストップワード除去を行った．クラスタリング処理には固有名詞を含む名詞，動詞，形容詞に品詞付けされた単語を用いた．さらに，対象となる入力文書集合の各単語の文書頻度を求め，文書頻度がデータ 1 では 20 文書未満，データ 2，3 では 15 文書未満の単語は棄却した．その結果，単語種類数はデータ 1，2，3 でそれぞれ，4,552，5,769，5,914 となった．データ 3 における最小イベントサイズ 30 に比べ，15 文書という閾値は小さな値ではないため，サイズの小さなイベントでは特有単語が抽出されない恐れがある．しかし，文書頻度の高い単語の組合せが特有単語対になることは珍しくなく，サイズの小さなイベントにおいて有効な特有単語が検出されなかったにしても，特有単語対によって十分にカバーしうると考えられる．逆にいえば，提案手法では，文書頻度の高い単語の組合せが特有単語対となることを期待して閾値を大きく設定し，単語種類数を少なくして処理量の増大を防いでいる（提案手法では単語共起行列を用いるため，処理量は単語種類数の 2 乗に比例する）．

文書間類似度，文書・クラスタ間類似度については，予備実験の結果，線形モデルでは複数のイベントを含むクラスタが生成しやすいことが確認されたので，2 次モデルのみを用いた．また，異なる話題の文書間で同じ単語が共有されることは珍しくないが，単語対に関しては，単語単独の場合に比べ，より共有されにくいと考えられる．したがって，単語共起の情報を強調することにより，異なる話題の文書間もしくは文書・文書集合間の類似度を低く抑えられると考えられる．そこで，式 (16) で類似度を算出するときは非対角成分の重み z_{mn} に 1 以上の大きな値を設定した．

また，各パラメータは，値を振りながら実験を行って最も良い結果を与える値を決定した．TDT2 では各文書に与えられるイベントラベルは 1 つなので，クラスタリング精度の算出を可能とするため排他的なクラスタリングを行った．また，処理時間の増大を抑えるため，ステップ 3 におけるクラスタ成長処理の繰返しは最大 7 回で打ち切るようにした．これによる性能の低下はなかった．

評価法について述べる．結果の評価はクラスタリング精度比較と得られたクラスタの分析により行った．クラスタリング精度（各文書を正しいクラスタに配置する精度）は以下のように定義した．まず，各文書を類似度の最も大きいクラスタに帰属させたうえで，各クラスタに帰属する文書のイベントラベルをチェックし，最も優勢なイベントラベルをそのクラスタのラベルとする．次いで，各クラスタにおいてクラスタラベルと一致するイベントラベルを持つ文書をカウントし，正解数とする．複数のクラスタが同じラベルを持つときは，正解文書数の最も多いクラスタ以外のクラスタに帰属する文書はすべて誤りとする．さらに，各イベントについて，そのイベントの文書を最も多く含むクラスタをそのイベントの対応クラスタと呼ぶこととする．あるクラスタが単一/複数のイベントの対応クラスタとなっているとき，そのクラスタを単一/複数イベントクラスタと呼ぶ．複数イベントクラスタでは，少ない文書数のイベントはより多い文書数のイベントにマージされていることになる．また，複数のクラスタが同じラベルを持つときは，それらをイベント分割クラスタと呼ぶ．

5.3 クラスタリングの実験結果

表 3 に最良のクラスタリング結果を示す．表 3 において，検出クラスタの欄の 4 つの数字は，左から順に，検出されたクラスタ数，単一イベントクラスタ数，複数イベントクラスタ数，イベント分割クラスタ数である．したがって，データ 1 ではすべてのイベントが単一イベントクラスタとして抽出されている．データ 2 では 31 イベントに対し 30 個のクラスタが検出され，単一/複数イベントクラスタは 29/1 個である．データ 3 では 38 イベントに対し 36 個のクラスタが検出され，単一/複数イベントクラスタは 34/2 個である．なお，どのデータでも，2〜3 個の冗長クラスタが発生しているほか，どのクラスタにも属さない文書が 6〜9 個存在した．また，ステップ 3 におけるクラスタ成長処理の平均繰返し数は，各データでクラスタあたり 4.8〜5.1 回であった．

データ 2 で検出に失敗し，他のイベントに吸収されたイベントは，“State of the Union Address（大統

表 3 クラスタリングの実験結果
Table 3 Clustering results.

データ	イベント数	検出クラスター	クラスタリング精度 (%)
1	21	21 21 0 0	95.17
2	31	30 29 1 0	95.09
3	38	36 34 2 0	94.82

表 4 文書頻度の高い 10 個の単語，単語対における特有単語対，および単語対
Table 4 Specific terms and term-pairs among the top 10 terms and term-pairs with highest document frequency.

イベント	Asian Ecnomic Crisis	Monica Lewinsky Case	Current Conflict with Iraq	1998 Winter Olympics
単語	economic	* lewinsky	* iraq	* olympic
	crisis	president	weapon	* medal
	* economy	* monica	* u.n.	* gold
	government	clinton	united	* olympics
	year	house	state	win
	country	white	* iraqi	game
	financial	* starr	military	* nagano
	* international	* counsel	* inspector	team
	* asia	jury	u.s.	time
	* asian	* grand	president	world
単語対	* international monetary	* lewinsky monica	* iraq weapon	* gold medal
	* international fund	clinton president	* u.n. iraq	* win medal
	* fund monetary	house white	state united	* gold win
	* crisis economic	* president lewinsky	* u.n. weapon	* olympic medal
	* crisis financial	* grand jury	* hussein saddam	* gold olympic
	state united	* independent counsel	* weapon inspector	* game winter
	* crisis asian	* clinton lewinsky	* united iraq	olympics winter
	president suharto	* independent starr	* u.n. inspector	state united
	* asia economic	* counsel starr	* iraq inspector	game olympic
	* asia crisis	president house	united nation	* win olympic

領の年頭教書)”（42 文書）である．これは，ニュースストーリに年頭教書の内容（いろいろな事柄に触れており，他のイベントと近いものもある）を述べたものが多く，このイベント特有の単語や単語対が少なかったためである．データ 3 では，さらに “Anti-Chinese Violence in Indonesia”（36 文書）が検出に失敗している．これは，同時期に起こった “Anti-Suharto Violence”（324 文書）と内容が近いため，吸収されてしまった結果である．

また，クラスタリング精度についていえば，どのデータにおいても，“Asian Economic Crisis” のうちのインドネシア関連記事の多くが “Anti-Suharto Violence” に誤って帰属するという現象が見られ，誤りの約半数を占めている．“Anti-Suharto Violence” は “Asian Economic Crisis” が引き金になったイベントであり，両者には共通する話題がかなり多く，誤りの多くは納得できるものとなっている．また，誤った記事の多くは “Anti-Suharto Violence” と “Asian Economic Crisis” の両方に対して反応している．この事実は提案手法は非排他的クラスタリング手法としても有効であろうことを示唆している．以上から，提案手法はクラスタリング法として満足すべき能力を有していることが示された．

表 4 は，データセット 3 において 4 つのイベント “Asian Economic Crisis”，“Monica Lewinsky Case”，“Current Conflict with Iraq”，“1998 Winter Olympics” の頻度の高い 10 の単語および単語対を示している．これらの中で特有単語，単語対には先頭に * が付与されている．表から分かるように妥当な単語，単語対が特有と判断されている．特に，特有単語対としては，“economic crisis” や “gold medal” のような名詞句以外に “iraq weapon” や “win gold” のようにそのイベントならではの単語対が表れている．名詞句を抽出して特有性の評価を行っただけでは得られない単語対が特有単語対として求められていることが分かる．

6. 考　　察

6.1 既存手法との比較

提案手法と既存の手法との比較のため，AHC, k-means，kNN について実験を行った．AHC では類似度尺度としてはグループ平均法を用いた．k-means ではイベント数をクラスタ数として指定した．また文書の長さの影響を排除するため各文書ベクトルは正規化して用いた．kNN はクラスタリング法というより文書分類法であるが，文書を仕分けるという点では共通点があるので比較相手とした．kNN では，まず，入力文書はすべての訓練文書との間で類似度（余弦類似

表 5 AHC によるクラスタリング結果
Table 5 Clustering results of AHC algorithm.

データ	クラスタリング精度 (%)	検出されたクラスター数
1	91.57	74
2	92.87	71
3	88.34	95

表 6 k-means によるクラスタリング結果
Table 6 Clustering results of k-means algorithm.

データ	クラスタリング精度 (%)
1	61-67
2	57-60
3	55-57

表 7 kNN による分類結果
Table 7 Classification results of kNN algorithm.

データ	分類精度(%)
1	98.03
2	97.58
3	97.02

度）が求められ，次いで類似度が大きい k 個の文書が選択される．着目クラスと入力文書との類似度は，選択された k 個の文書の中で着目クラスに属する訓練文書の余弦類似度の総和で与えられる[13]．kNN は性能の高さのゆえによく知られている．余弦類似度を求める際には，文書ベクトルの各要素の値は対応する単語の $tf\,idf$（単語出現頻度と文書頻度の逆数の積）とした．ただし tf，idf とも対数を求めている．kNN の実験では，2 重の交差検定で評価した．すなわち，データの半数を訓練用に，残りをテスト用に用いるという実験をデータを回転させて 2 回繰り返した．また，k は 20 とした．

AHC によるクラスタリングでは，多くの微小クラスタ（たとえば文書数 20 以下）が生じ，これら微小クラスタが同じイベントの大きなクラスタにマージされる前に，異なるイベントのクラスタどうしのマージが始まっていた．そのためいつクラスタリングを中断するかの判断が難しく，クラスタの検出能力の評価は困難であったので，クラスタリング精度の最大値とそのときのクラスタ数で評価を行った．**表 5** に結果を示す．表から分かるように，クラスタリング精度において提案手法が優っている．また，表 5 では検出クラスタ数はイベント数に比べ著しく多くなっている．その一方で，クラスタリング精度はさほど低くない．これは，クラスタリングのエラーの多くは，他の大きなクラスタと同じラベルを持つ微小クラスタにおいて生じているためである．このように AHC では微小クラスタの処理が問題となっている．

k-means の結果を**表 6** に示す．表 6 でクラスタリング精度に幅があるのは，クラスタリング結果がクラスタの種の選び方に依存するためである．k-means では，文書数の多いイベントは複数のクラスタに分割され，少ないイベントは他のイベントにマージされる傾向にある．そのため，表 6 に示されるようにクラスタ数が事前に分かっていたとしても高いクラスタリング精度は得られない．

このように，提案手法は従来法に比べ，クラスタリング精度，クラスタ検出精度ともに優れている．特に，クラスタ検出能力は従来法に比べ著しく改善されている．

また，2.6 GHz の Pentium4 プロセッサを用いた場合，文ベクトル算出後の処理時間は，データ 3 に対して，AHC では約 7 時間 47 分，k-means では約 8 分 50 秒，提案手法では約 1 時間 13 分であった．AHC ではすべての文書対についての類似度算出の処理時間が全体の 56%を占めていた．提案手法では処理量は単語種類数の 2 乗に比例するため処理時間が問題となる恐れがあったが，5.2 節で述べた実験条件の下では提案手法の処理時間は k-means に比べれば長いものの，AHC よりは短くなっている．さらに，前述のように高頻度の単語の組合せも特有単語対になりうるので，より文書頻度の高い単語のみに絞っても有効な特有単語対が残されている可能性もある．そうなれば処理量はさらに少なくなる．この確認は今後の課題である．

表 7 は kNN による文書分類結果を示す．たとえ半分とはいえ，各クラスの文書が例示されている文書分類に比べ，クラスタリングでは各クラスの情報について事前にまったく分かっていない．そのため，クラスタリング精度については，文書分類の結果よりも劣るのは当然である．しかし，実験結果では文書分類に非常に近い値のクラスタリング精度が得られている．

6.2 提案類似度の評価

提案クラスタリング法が正しく動作するか否かは，本論文で提案した類似度が文書と文書集合との内容の近さをどの程度正確に表すかにかかっている．そこで提案類似度を文書分類に適用することにより，その評価を行った．用いたデータは上記データ 3 であり，文書クラスは 38 である．分類精度は 2 重の交差検定により評価した．文書・文書集合類似度としては式 (15)（線形類似度），および式 (16)（2 次類似度）を用い，特有単語・単語対を選択した/しない場合，さらに式 (16)

表 8 提案類似度を用いた文書分類の結果
Table 8 Precision of text classification using the proposed similarity measure.

特有単語・単語対選択	線形類似度	2次類似度	
		非対角成分重みなし	非対角成分重みあり
なし	94.85	95.59	95.98 (z_{mn}=2.0)
あり	95.57 (α=6.0)	96.38 (α=2.2)	97.01 (z_{mn}=3.0, α=1.4)

に対しては非対角項に重みを付与した/しない場合について実験を行った．**表 8** は，4.3 節における α，式 (16) における z_{mn} の値を振った場合の最も良い結果とそのときの α，z_{mn} を示している．表 8 から分かるように，精度向上を図るうえで特有単語・単語対の選択的使用，非対角項の重み付けはともに有効である．特に，特有単語・単語対の選択的使用の効果は著しい．2 次類似度で特有単語・単語対の選択的使用，非対角項の重み付けを行って得られる精度 97.01%は表 7 における kNN の分類精度 97.02%とほぼ同じである．この事実は，提案手法では kNN に匹敵する正確さで文書・文書集合間の類似度が求められることを示している．文書総数を R，文書内の平均単語数を M_{ave} とすると，kNN の処理量は $O(R^2)$，提案手法では $O(M_{ave}R)$ である．平均単語数が文書総数よりも少ないときには提案手法の方が処理量が少ない．提案手法は文書分類にも有効と考えられる．ちなみに，データ 3 の場合には，$R = 7546$，$M_{ave} = 98.1$ である．

このような文書・文書集合間の類似度の正確さは，ステップ 3 においてカレントクラスタに帰属する文書を選択するときに効果を発揮する．**図 4** は，表 3 のデータ 3 に対して求められた 36 個のクラスタ（最終状態）の各々に帰属する文書を決定したときの再現率，精度の関係を表す図である．再現率は各クラスタに正しく帰属した文書数の各クラスタに帰属すべき文書数に対する割合，精度は各クラスタに正しく帰属した文書数の各クラスタに実際に帰属した文書数に対する割合である．これらは，1 つのクラスタと全文書との間で求められた類似度をもとに閾値を振りながらそのクラスタに正しく帰属した文書数，誤って帰属した文書数を算出した後，全クラスタに対するそれらの和を求めて算出した．また，図 4 では類似度は，クラスタリングで実際に用いた 2 次類似度（式 (16) で非対角成分の重み付け，特有単語・単語対の選択を行ったもの）と線形類似度（式 (15) 使用，特有単語の選択は

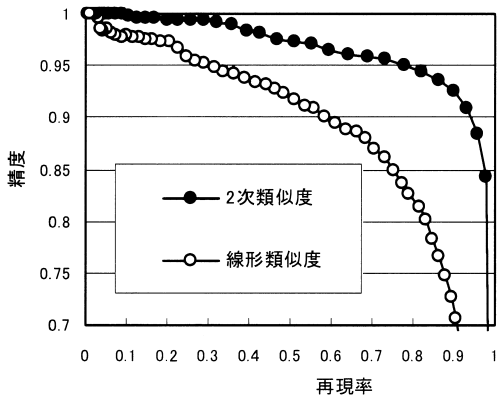

図 4 カレントクラスタへの帰属決定時の精度と再現率
Fig. 4 Precision and recall for documents belonging to current clusters.

行っていない）の両方について求めている．線形類似度は従来のベクトル空間モデルで求めた類似度と等価である．図から分かるように，2 次類似度の場合には，90%以上の再現率，精度を同時に達成しており，線形類似度に比べて著しく性能が高い．この効果は次のように表れ，クラスタを 1 つずつ検出するという戦略を有効にする．

(1) 再現率が高いがゆえに検出済みのクラスタに対応する話題の文書は残存文書集合に残されることが少なくなる．同時に，精度が高いがためにクラスタとして未検出の話題の文書は残存文書に多く残ることになる．そのため，検出済みの話題とは異なる話題の種文書の検出が容易になる．
(2) 精度が高いのは成長途中のクラスタについても同様である．したがって，カレントクラスタでは他の話題の文書が紛れ込むことが少ないので，式 (16) の $\boldsymbol{Q}^A$ を正確に求めることができる．そのため次にクラスタを成長させるときにも，そのクラスタに属すべき文書を正確に抽出できる．

6.3 パラメータの影響

提案手法では設定すべきパラメータが多く，その 1 つ 1 つが結果に影響を及ぼしうる．各パラメータがどのように，またどの程度結果に影響するかを把握することは重要である．そこで，以下の 6 種類のパラメータをとりあげ，その影響を調べた．

(a) 式 (13)，(14) における閾値 A．A はカレントクラスタのメンバ数の β 倍に設定することにする．
(b) ステップ 2 において，各文書と種文書との類似度と比較し，近隣文書を選択するための閾値 nei_{th}．
(c) ステップ 3 において，各文書とカレントクラス

表 9 各パラメータのクラスタリング性能に及ぼす影響
Table 9 Influences of parameters on clustering performance.

	β						
	0.00	0.05	0.10	0.15	0.20	0.25	0.30
nei_{th}=0.05 cl_{th}=0.01	36 34 2 0	36 34 2 0	36 34 2 0	36 34 2 0	36 34 2 0	36 34 2 0	40 32 2 6
res_{th}=0.005 α=1.2 z_{mn}=5.0	94.79 / 94.79	94.82 / 94.82	94.60 / 94.60	94.53 / 94.53	94.26 / 94.26	94.51 / 94.51	89.91 / 94.35
nei_{th}=0.03	35 32 3 0	35 32 3 0	35 32 3 0	35 32 3 0	35 30 4 0	35 32 3 0	36 32 3 1
	94.34 / 94.34	94.35 / 94.35	94.31 / 94.31	94.21 / 94.21	93.32 / 93.32	93.25 / 93.25	86.77 / 92.76
nei_{th}=0.07	39 32 2 5	39 32 2 5	38 33 2 3	38 30 3 5	39 32 2 5	43 34 1 8	43 34 1 6
	90.37 / 93.52	90.41 / 93.61	91.17 / 94.25	89.31 / 93.47	90.39 / 94.04	93.23 / 93.23	85.78 / 93.77
cl_{th}=0.005	33 28 5 0	33 28 5 0	35 32 3 0	34 31 3 0	35 32 3 0	35 32 3 0	36 31 3 2
	91.84 / 91.84	91.70 / 91.70	93.39 / 93.39	91.88 / 91.88	93.19 / 93.19	92.53 / 92.53	90.31 / 92.65
cl_{th}=0.02	38 33 2 3	36 31 3 2	35 32 3 0	37 33 2 2	36 31 3 2	37 33 2 2	40 34 1 5
	91.20 / 93.93	90.03 / 93.47	93.40 / 93.40	90.46 / 93.86	89.78 / 93.20	90.31 93.53	87.94 / 93.47
res_{th}=0.003	36 31 3 2	36 31 3 2	37 33 2 2	35 33 2 0	35 32 3 0	35 32 3 0	40 32 2 6
	89.84 / 92.68	90.05 / 92.80	90.31 / 93.07	91.97 / 91.97	94.17 / 94.17	93.92 / 93.92	87.53 / 93.53
res_{th}=0.01	36 31 3 2	37 33 2 2	37 33 2 2	37 33 2 2	36 34 2 0	35 32 3 0	37 33 2 2
	91.04 / 93.15	91.52 / 93.75	91.12 / 93.52	90.74 / 93.36	94.29 / 94.29	93.44 / 93.44	91.60 / 93.75
α=1.1	38 32 2 4	38 32 2 4	37 33 2 2	39 31 2 6	37 33 2 2	41 32 2 7	40 34 1 5
	90.35 / 94.17	90.78 / 94.34	91.48 / 94.95	87.53 / 93.45	91.25 / 94.67	85.22 / 93.67	84.93 / 94.17
α=1.3	34 30 4 0	34 30 4 0	34 30 4 0	35 32 3 0	35 32 3 0	35 32 3 0	36 34 2 0
	92.59 / 92.59	92.56 / 92.56	92.59 / 92.59	93.82 / 93.82	94.33 / 94.33	93.18 / 93.18	93.81 / 93.81
z_{mn}=1.0	31 24 7 0	31 24 7 0	32 27 5 0	33 29 5 0	32 26 5 1	34 24 6 4	33 28 5 0
	88.46 / 88.46	/ 88.40 / 88.40	90.49 / 90.49	91.03 / 91.03	88.39 / 89.11	84.14 / 89.19	89.57 / 89.57
z_{mn}=8.0	37 33 2 2	35 32 3 0	36 34 2 0	36 34 2 0	36 34 2 0	36 34 2 0	38 32 2 4
	92.51 / 93.88	93.59 / 93.59	94.38 / 94.38	94.45 / 94.45	94.14 / 94.14	93.61 / 93.61	90.70 / 93.90

タとの類似度と比較し，カレントクラスタに帰属する文書を選択するための閾値 cl_{th}.

(d) ステップ 1，4 において，各文書と各クラスタとの類似度と比較し，残存文書を選択するための閾値 res_{th}.

(e) 4.3 節において特有単語・単語対を選択するためのパラメータ α.

(f) 式 (16) における非対角成分に対する重み z_{mn}.

表 3 の最良の結果は $\beta = 0.05$, $nei_{th} = 0.05$, $cl_{th} = 0.01$, $res_{th} = 0.005$, $\alpha = 1.2$, $z_{mn} = 5.0$ のときに得られている．そこで，β については，0.0 から 0.30 まで 0.05 刻みに値を与え，その他のパラメータについては，1 つのパラメータのみ最良の結果を与える値とは異なる値を与えてクラスタリングを行い，パラメータ依存性を観察することとした．**表 9** に結果を示す．表 9 の最上段の結果では，β 以外のパラメータは最良の結果を与える値に固定されている．2 段目以下の欄では，最も左の欄に示されたパラメータ以外は最上段のパラメータ値をそのまま採用している．各欄の結果は，上段の 4 組の数字は表 3 と同じである．また，下段の 2 組の数字のうち，左側はクラスタリング精度，右側はクラスタ純度である．クラスタ純度は，クラスタラベルと同じイベントラベルを持つ文書の総数の各クラスタのメンバ数の総和に対する比である．したがって，イベント分割クラスタの中で，メンバ数が最大でないクラスタにおいてもクラスタラベルと同じイベントラベルの文書は正解となる．1 つのイベントが複数のクラスタに分割されなければ，クラスタリング精度はクラスタ純度と同じである．表 9 から以下をいうことができる．

① 全体的な傾向として，β が小さいときは複数イベントクラスタ数が多くなっており，イベントがマージしやすくなることを示している．反対に，β が大きいときにはイベント分割クラスタ数が多く，イベントが分割されやすくなっている．$\beta = 0.3$ のときには特に著しい．このような結果は 3.6 節で述べたように，A の値によりクラスタの話題の揃い方が規定されたことによっている．表 9 では 0.1〜0.2 の β で良好な結果が得られることが多い．

② nei_{th} を大きくとるとイベント分割クラスタ数が多くなり，小さくとると複数イベントクラスタ数が多くなっている．前者は，nei_{th} が大きい場合，種文書との類似度が大きい文書のみ初期クラスタに含まれるので初期クラスタが小さくなってその話題が過度に狭くなり，クラスタの成長が阻害される場合があることを示している．後者は，nei_{th} を小さくとれば初期クラスタは大きくなり，複数

の話題の文書を含む可能性が大きくなるため，複数のイベントが 1 つのクラスタに吸収されやすくなることを示している．

③ cl_{th} についても，分かりやすい影響が現れており，cl_{th} に小さな値を設定すると複数イベントクラスタ数が多くなっている．これは，カレントクラスタと類似度の低い文書も取り込むため他のイベントの文書を吸収しやすくなった結果である．反対に cl_{th} に大きな値を設定するとイベント分割クラスタ数が多くなる．これは，カレントクラスタと同じイベントの文書でもカレントクラスタに取り込まれにくくなるために成長が十分に進まず，イベントが分割されやすくなって生じた結果である．

④ res_{th} については大きな差はないが，β が中間的な値 (0.05–0.20) のときに，$res_{th} = 0.003$ とすると得られるクラスタ数が少なめになり，$res_{th} = 0.01$ とすると多めになるようである．前者の場合には，res_{th} が小さくなることにより残るべき文書が残存文書に残らなくなり，最後まで成長しうるような種文書が抽出されなくなった結果と考えられる．また，後者では，抽出されたクラスタから取り残される文書が多くなり，残存文書が少なくなったときにいろいろな話題の文書が混在するようになって良好な種文書の抽出が阻害された結果と思われる．

⑤ α についても分かりやすい結果が現れている．α が小さいときには特有単語・単語対が十分に抽出されず，イベントが分割されやすくなって，イベント分割クラスタ数が多くなる．反対に α が大きいと特有単語・単語対が他のイベントと共有されやすくなり，複数のイベントがマージされやすくなって，複数イベントクラスタ数が多くなる．

⑥ z_{mn} については，$z_{mn} = 1.0$ のときに大きな影響が現れている．$z_{mn} = 1.0$ では式 (16) において非対角成分に重みは付与されない．その結果異なるイベントの文書との類似度を低めることができず，他のイベントの文書を取り込みやすくなって，複数イベントクラスタが増えたものと思われる．

⑦ 表 9 の実験結果において，複数イベントクラスタ数が 4 以下の場合には，3 個以上のイベントが 1 つのクラスタにマージされた例は皆無であった．したがって表 9 の大半の例では，複数イベントクラスタ数は 3 以内であり，他にマージされて検出できなかったイベントは 3 個以内に抑えられている．

⑧ 表 9 では，少なからぬ例でイベント分割クラスタが生じている．1 つのイベントが複数のクラスタに分割されていたにしても，内容が把握できればユーザには受け入れられると考えられる．複数の話題が 1 つのクラスタにマージされているよりはよい．したがって，クラスタリング純度もクラスタリングの性能を規定するうえで重要な指標となると考えられる．表 9 の結果では精度重視の観点からすれば，パラメータへの依存性は小さいとはいえない．しかし，純度を重視し純度 93%を目安とすると，多くの例で 93%を超えており，パラメータ依存性は緩やかといってよいであろう．

さて，残された問題はここで求められた最適パラメータが他の文書集合にも有効かどうかである．上述のように本論文では TDT2 を用い，内容の揃ったクラスタを得ることに成功している．したがって他の文書集合でも同じパラメータで内容は揃ったクラスタが得られると考えられる．クラスタリング精度についてはどのような基準で話題が分類されているかにもよるので議論することは難しい．しかし，クラスタの内容は揃うと予想されるので，純度については満足のいく結果が得られると信ずる．

7. ま と め

以上，本論文ではクラスタを 1 つずつ検出するという戦略のもとに，新しい非階層的クラスタリング法を提案した．提案手法で用いる文書・クラスタ間類似度は以下のような特長を有している．

- 多文書間の共通性分析に基づき，文書・クラスタ間類似度を対象文書とクラスタ内の文書集合の共通情報との間で求める．
- 文書・クラスタ間類似度の定義には，単語単独の出現情報ばかりでなく，単語共起の情報も考慮に入れる．
- その算出には，各クラスタの特有単語，単語対を検出し，類似度算出に選択的に用いる．

評価実験の結果，話題に応じて正しい数のクラスタを得る能力，各文書を正しいクラスタに配置する能力において，提案手法は AHC や k-means などの既存手法に比べ優っていることが確認された．提案手法は教師なしの分類手法と見なすことができるが，分類法としても教師ありの分類手法の kNN に近い精度を有することが確認された．これらの結果はクラスタリング戦略の妥当性を証明するものである．提案手法はクラスタを 1 つずつ正しく抽出するので，より多数の話題を有する文書集合に対しても有効に動作すると考えら

れる．実験結果は提案手法はオフライン処理によって文書集合に含まれる話題を精度良く抽出できることを示している．TDT はオンライン処理であるので，提案手法をオンラインのクラスタリングにも適用できるように改善することは今後の重要な課題である．

参 考 文 献

1) van Rijsbergen, C.J.: *Information Retrieval*, Buttersworth, London (1979).
2) Harady, H., Shimizu, N., Strzalkowski, T. and Ting, L.: Cross-document summarization by concept classification, *Proc. 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.121–128, Tampere, Finland (Aug. 2002).
3) Zha, H.: Generic Summarization and Keyphrase extraction using mutual reinforcement principle and sentence clustering, *Proc. 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.113–120, Tampere, Finland (Aug. 2002).
4) Zamir, O. and Etzioni, O.: Web document clustering: A feasibility demonstration, *Proc. 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.46–54, Melbourne, Australia (Aug. 1998).
5) Hearst, M.A. and Pedersen, J.O.: Reexamining the cluster hypothesis: scatter/gather on retrieval results, *Proc. 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.76–84, Zurich, Switzerland (Aug. 1996).
6) Allan, J. (Ed.): *Topic Detection and Tracking Event-based Information Organization*, Kluwer Academic Publishers, Boston (2002).
7) http://www.ldc.upenn.edu/Projects/TDT2/
8) Willet, P.: Recent trends in hierarchical document clustering: A critical review, *Information Processing and Management*, Vol.24, pp.577–597 (1988).
9) Willet, P.: Document clustering using an inverted file approach, *Journal of Information Science*, Vol.2, pp.223–231 (1990).
10) Milligan, G.W. and Cooper, M.C.: An examination of procedures for detecting the number f clusters in a data set, *Psychometrika*, Vol.50, pp.159–179 (1985).
11) Liu, X., Gong, Y., Xu, W. and Zhu, S.: Document Clustering with Cluster Refinement and Model Selection Capabilities, *Proc. 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.191–198, Tampere, Finland (Aug. 2002).
12) Dharanipragada, S., Franz, M., McCarley, J.S., Roukos, S. and Ward, T.: Story Segmentation and Topic Detection in the Broadcast News Domain, *Proc. DARPA Broadcast News Workshop* (1999).
13) Yang, Y. and Liu, X.: Re-examination of Text Categorization, *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'99*), pp.43–49 (1999).
14) Manning, C.D. and Schutze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press (1999).

(平成 17 年 9 月 8 日受付)
(平成 18 年 3 月 2 日採録)

川谷 隆彦（正会員）

1944 年生．1967 年東京大学工学部物理工学科卒業．同年日本電信電話公社（現 NTT）電気通信研究所入所．ホログラム，文字認識の研究実用化に従事．1993 年ヒューレット・パッカード日本研究所入所．引き続き文字認識，テキスト処理の研究実用化に従事．2004 年メディアドライブ（株）入社．工学博士．電子情報通信学会，IEEE Computer Society，ACM 各会員．