2003－ＮＬ－158　（７）
2003／11／6

共通性分析による文書クラスタリングの評価

川谷隆彦

日本ヒューレット･パッカード㈱　ヒューレット･パッカード研究所

takahiko.kawatani@hp.com

本報告では、先に提案した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の評価結果について述べる。提案手法では、種として抽出された文書に同じ話題の文書をマージさせつつ順次クラスターを成長させていく。提案手法は、文書・クラスター間の類似度を求める際に、単語共起の情報をも用いた新しい類似度尺度を用いていること、その時点のクラスターの特有単語、単語対を選択的に用いることを特長としている。TDT2 のコーパスから選択した 21 イベント 6788 文書、31 イベント 7306 文書、38 イベント 7546 文書のそれぞれに対し、検出クラスター数 21、30、36、クラスタリング精度 95.17%、95.09%、94.41%を得た。

Evaluation of the Document Clustering Method Based on Commonality Analysis of Multiple Documents

Takahiko KAWATANI

Hewlett-Packard Labs Japan, Hewlett-Packard Japan

takahiko.kawatani@hp.com

This paper describes evaluation of a non-hierarchical clustering method based on multi-document commonality analysis proposed by the author. In the method, a document extracted as a seed grows up to a cluster by iteratively merging documents with the same topic. It features in obtaining document-cluster similarity that it uses a new similarity measure reflecting term co-occur information and that specific terms and term pairs extracted from the current cluster are used. In experiments using 7546 documents extracted from 38 events in TDT2 corpus, 36 events were extracted as the clusters with 94.41% clustering accuracy.

1. まえがき

文書クラスタリングは情報検索やテキストデータマイニングなどの分野で重要な技術のひとつとなっており、これまで様々な方法が研究されてきた[1][2]。現時点では、クラスター間類似度としてグループ平均法を用いる階層的手法、フラット形（非階層形）に属する K-means 法などが実際に多く用いられているようであるが、

- クラスター検出精度(クラスターの数を正しく抽出する)
- クラスタリング精度(文書を正しいクラスターに配置する)

の点で完璧とは云い難い状況にある。

ところで、文書のクラスターとはどのように定義すべきであろうか？　本報告では話題（トピック）を共通にする文書のグループをクラスターと定義する。各話題は話題特有の単語や単語対を有すると考えられるから、特定の単語や単語対を共有する文書集合がクラスターであると定義してもよい。この場合、“特有な単語・単語対”はそのク

ラスターに集中して現れる、即ち、(そのクラスター中の出現頻度／全文書中の出現頻度) の値が大きい単語・単語対を指し、クラスター中の出現頻度が高くない単語・単語対も特有になりうる。このような文書クラスターの定義が常に正しいかについては保証の限りでないが、少なくともトピカルな文書の場合には妥当と思われる。

文書のクラスターをこのように定義する限り、クラスタリング処理においては各クラスターに対して特有な単語・単語対を明らかにすることが求められる。ひとたび各クラスターに特有な単語・単語対が求められれば、各クラスターの話題が明確になり、各クラスターに属すべき文書を正確に決めることができてクラスタリングの精度は向上するであろう。さらにクラスタリング精度が向上すればクラスター検出精度も自ずと高まるであろう。ところが、従来のクラスタリング方法は特有単語・単語対を選択するプロセスを有していない。クラスタリングの結果確かに話題が似たような文書がグループを形成し、同じような単語・単語対を含む文書集合が出来上がる。しかし、この“同じような単語・単語対”はグループ特有であるとは限らず、高頻度でありながら特有でないものも存在する。例えば、TDT2 コーパスにけるイベント“Monica Lewinsky Case”に関連する文書集合で、“Monica Lewinsky”、“President Clinton”は共に高頻度単語・単語対であるが、他のイベントの文書に現れる頻度は前者では少なく、後者では多いであろう。従って前者は特有単語・単語対となるが、後者はならない。従来のクラスタリング法では、クラスター間類似度を求める際に単語・単語対の選択的な使用をしてないため、その値は特有でないにも拘らず頻度の高い単語・単語対の影響を受けていた。その結果、クラスターの中心では同じ話題の文書で占められるにしても、周辺では非特有で頻度の高い単語・単語対を有する他の話題の文書が混入する可能性を排除できなかった。

多くの場合文書は各単語の出現頻度を値とする変数の組で表現されており、単語変数は量的な変数とみなされている。しかし、ある単語が何個存在するかよりも存在するか否かの情報が重要となる場合もあり、単語変数は量的な性格だけでなく、質的な性格も有すると考えられる。ある文書集合において共通する単語は存在するか、その中でどの単語・単語対が特有であるかというような議論は単語変数を質的に捉えて初めて自然に行うことができる。一方、従来のグループ平均法を用いる階層的手法、K-means 法などでは上記のような共通性、特有性などの議論には明らかに不向きである。これは、これらの方法が量的な変数を扱うクラスタリング法であるためであり、これらは少なくともトピカルな文書に対しては向いていなかったと云わざるを得ない。

このようなことから、筆者は先に多文書間の共通性分析による文書クラスタリング法を提案している[3]。従来の多くのクラスタリング法で行われているような各クラスターの同時並行的な処理では特有単語・単語対を検出することは難しい。そこで提案手法では、先ずクラスターの種となる文書をひとつ取り出し、ついでその種と同じトピックの文書を検出してマージすることにより種を成長させるというアプローチを採用しており、途中段階のクラスターから検出した特有単語・単語対を文書・クラスター間の類似度算出に選択的に用いている。前報では TDT2 コーパスを用いて高いクラスタリング精度、クラスター検出精度が得られることを示したが、イベント数、文書数が十分多くなく、また、従来手法との比較検討がなされていなかった。そこで、本報告では、TDT2 コーパスにおけるより多くのイベント、文書を用いた場合でも本手法は有効であることを検証し、他のクラスタリング法や分類法との比較により本手法の特長を明確にすることとする。

以下、**2.**では、本報告で用いる文書間類似度、文書・クラスター間類似度について、**3.**ではクラスタリングの流れと前報との変更点について述べる。**4.**では、TDT2 コーパスを用いた実験結果を紹介し、**5.**で考察として、パラメータ依存性について述べ、従来手法との比較を行う。**6.**で全体を纏める。

2. 類似度尺度

2.1 文書間類似度

現れる単語集合が$\{w_1,...,w_M\}$で与えられる N 個の文から成る文書 D を考える。n 番目の文ベクトルを $\boldsymbol{d}_n=(d_1,..., d_M)^T$ として、文書 D の共起行列 $\boldsymbol{S}$ を以下のように求める。

$$\boldsymbol{S} = \sum_{n=1}^{N} \boldsymbol{d}_n \boldsymbol{d}_n{}^T \qquad (1)$$

ここで、T は転置を表す。文をバイナリベクトルで表すと、S_{ii}は単語 i が生起する文の数であり、同じ文に同じ単語は 1 回しか現れない場合は文書中の単語 i の頻度となる。一方、S_{ij}は単語 i、j が共起する文の数となる。

次に、共起行列を $\boldsymbol{S}^P$ とする文書 P を考える。文書 D、P 間の類似度 $sim(D,P)$を以下で定義する[3][4]。

$$sim(D,P) = \frac{\sum_{m=1}^{M} S_{mm} S^P{}_{mm}}{\sqrt{\sum_{m=1}^{M} (S_{mm})^2} \sqrt{\sum_{m=1}^{M} (S^P{}_{mm})^2}} \qquad (2)$$

$$sim(D,P) = \frac{\sum_{m=1}^{M} \sum_{l=1}^{M} S_{ml} S^P{}_{ml}}{\sqrt{\sum_{m=1}^{M} \sum_{l=1}^{M} (S_{ml})^2} \sqrt{\sum_{m=1}^{M} \sum_{l=1}^{M} (S^P{}_{ml})^2}} \qquad (3)$$

これらのうち、式(2)は従来のベクトル空間モデルにおいて単語頻度を文書ベクトルの成分値とする場合の余弦類似度と等価であり、類似度は単語単独の出現回数によって決まる。一方、式(3)では単語間の共起情報も類似度の決定に用いられるため、単語共起の傾向も似ていないと類似度は高くならない。式(2)の分子は文書 D、P 間の全ての文ベクトル同士の組合せの内積の線形和、式(3)の分子はそれらの 2 乗和となるので、式(2)を線形モデル、式(3)を 2 次モデルと呼ぶこととする。

2.2 文書・クラスター間類似度

提案手法では文書集合中の各文書の共通情報と各文書との類似度を文書・クラスター間類似度として求める。

文ベクトルがバイナリとして、各文書から一つずつ取り出して構築した文ベクトルの集合において、全ての文ベクトルの同じ成分の積の値を対応する成分の値とするベクトルを共通文ベクトルとする。全ての文に含まれる単語が存在すれば、共通文ベクトルの対応成分は 1 となる。従って、このような共通文ベクトルの集合は文書集合の共通情報を反映すると考えられる。文書 r の共起行列を $\boldsymbol{S}^r$、R 個の文書から成る文書集合Ωに対する共通文ベクトル集合の共起行列を $\boldsymbol{S}^C$ とすると、$\boldsymbol{S}^C$ の ml 成分は次式で与えられる。

$$S^C{}_{ml} = \prod_{r=1}^{R} S^r{}_{ml} \qquad (4)$$

$S^C{}_{ml}$は単語 m、l が共起する共通文の数となり、$S^C{}_{ml}$ が 0 でないときには全ての文書に単語 m、l が共起する文が含まれることを意味する。式(2)、(3)においてこのような $\boldsymbol{S}^C$ を $\boldsymbol{S}$ として用いれば、文書 P と文書集合Ωの共通情報との類似度が求められることになる。しかし、実際には、サイズの大きい文書集合の場合、

(1) 話題が同じであっても全ての文書で出現する単語が存在するとは限らない。
(2) $S^C{}_{ml}$が 0 でない場合、極端に大きな値をとる場合がある。

などの問題があり、以下のように対処する。

(1)の問題に対しては、各文書から一つずつ取り出した R 個の文のうちの A 個の文に現れる単語を共通単語とすることで対処する。先ず、各単語、単語共起の文書頻度を格納する行列 $\boldsymbol{U}$ を求める。U_{mm}、U_{ml}はそれぞれ単語 m の出現する文書数、単語 m、l の共起する文書数となる。次に、ml 成分が式(4)において値が 0 でない $S^r{}_{ml}$ の積として与えられる行列 $\boldsymbol{T}$ を求める。つまり、$S^r{}_{ml}$ が 0 の場合はその値を一時的に 1 としたうえで式(4)を用いて $S^C{}_{ml}$ の値を求め、$T_{ml} = S^C{}_{ml}$ とする。但し、全ての $S^r{}_{ml}$ が 0 の場合は T_{ml} は 0 とする。次ぎに、A 個以上の文書に現れる単語を用いた共通文ベクトルの共起行列 $\boldsymbol{T}^A$ を定義する。先ず、y_m を

$$y_m = 1 \quad \text{if } U_{mm} \geq A,$$
$$y_m = 0 \quad \text{otherwise.} \qquad (5)$$

とし、行列 $\boldsymbol{T}^A$ の ml 成分を

$$T^A{}_{ml} = T_{ml} y_m y_l \qquad (6)$$

により与える。$T^A{}_{mm}$ は A 個以上の文書に現れる単語 m の出現する共通文の数、$T^A{}_{ml}$ は A 個以上の文書に現れる単語 m、l の共起する共通文の数となる。

次に、問題(2)に対処するため、行列 $\boldsymbol{Q}^A$ を

$$Q^A{}_{ml} = \log(T^A{}_{ml}) \quad \text{if } T^A{}_{ml} > 1$$
$$= 0 \quad \text{otherwise} \qquad (7)$$

により定義する。結局、文書 P の文書・クラスター間類似度は

$$sim(P,\Omega) = \frac{\sum_{m=1}^{M} Q^A{}_{mm} S^P{}_{mm}}{\sqrt{\sum_{m=1}^{M} (Q^A{}_{mm})^2} \sqrt{\sum_{m=1}^{M} (S^P{}_{mn})^2}} \qquad (7)$$

$$sim(P,\Omega) = \frac{\sum_{m=1}^{M} \sum_{l=1}^{M} Q^A{}_{ml} S^P{}_{ml}}{\sqrt{\sum_{m=1}^{M} \sum_{l=1}^{M} (Q^A{}_{ml})^2} \sqrt{\sum_{m=1}^{M} \sum_{l=1}^{M} (S^P{}_{ml})^2}} \qquad (8)$$

で与えられる。式(7)は線形モデル、式(8)は 2 次モデルである。

上記 A の値は得られるクラスターの話題の揃い方を規定するパラメータになる。A の値を大きくとるほど得られたクラスターの話題は揃うことになる。また、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、クラスターが順調に成長せず、局所的に話題の揃ったサブクラスターに分割されてしまうことも考えられる。従って、A の値は、入力母集団の各話題の広がりの程度に応じて適切な値を決定する必要がある。

3. クラスタリングの方法

3.1 手順の概略

処理の流れの概略は以下の通りである。

ステップ 1：初めての処理の時には全文書から、2 回目以降の繰返しの時にはその時点のどのクラスター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、クラスターの種の候補となる文書を複数検出する。

具体的には、異なる話題の文書は同じ単語対を共有しない、同じ話題の文書集合の中で中心的な文書は式(8)の値は大きいとの仮定のもとに、その時点の残存文書から求めた行列 $\boldsymbol{Q}^A$ を用いて各残存文書の式(8)の値を求め、値の大きな一定数の文書を種文書候補とする。

ステップ 2：先ず、クラスターの種の各候補文書につき全ての文書との類似度（式(2)もしくは式(3)）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。近隣文書数が最も多くなる文書をクラスターの種とし、その近隣文書集合により初期クラスターを形成する。

ステップ 3：その時点でのクラスター文書集合から特有単語・単語対を求めた後、各文書との間で類似度（式(7)もしくは式(8)）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ーに仮に帰属させることによりクラスターを成長させる。クラスターに仮に帰属する文書数が一定になればステップ 4 へ。そうでなければ本ステップを繰返す。

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：各文書について各クラスターとの類似度を求め、類似度の最も高くなるクラスターに帰属させる。

ステップ 6：クラスター間の重なりをチェックし、重なりの大きなクラスターがあれば、各クラスターの重要度（そのクラスターに対してのみ類似度が一定値以上の文書の数）を求め、重要度の小さなクラスターを冗長なクラスターとして削除し、各文書の帰属するクラスターを求めなおす。

3.2 前報との違い

クラスタリングの各ステップの処理は基本的に前報と同じであるが、異なっている部分は次の通りである。

3.2.1 終了条件（ステップ 4）

残存文書数が一定値以下（例えば 50）になった時、残存文書同士で各文書の近隣文書を求め、最大の近隣文書数が一定値以下（例えば 5）の場合には検出すべきクラスターは最早残っていないと判断し、終了する。

3.2.2 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ー文書集合から求められた文書頻度行列とする（前報では $\boldsymbol{U}^0$ と $\boldsymbol{U}$ とを取違えていた）。U_{mm}/U^0_{mm} もしくは U_{ml}/U^0_{ml} の値は単語 m、もしくは単語 m、l の対の着目クラスター文書集合への集中の度合い（集中度）を表す。集中度の大きい単語はそのクラスターに特有と考えられるので、集中度が閾値以上の単語 m、もしくは単語 m、l の対は着目クラスターに特有とする。前報では単語も単語対も同じ閾値を用いて特有性を判断していたが、本報告では異なる値を用いた。単語の場合は、着目クラスター文書集合における最も頻度の高い単語 30 個のうち、U_{mm}/U^0_{mm} の値の大きな 5 個の平均値の 1/1.2 倍、単語対の場合は、最も頻度の高い単語対 100 個のうちの U_{ml}/U^0_{ml} の値の大きな 50 個の平均値の 1/1.2 倍とした。

特有と判断されなかった単語 m、単語 m、l の対に対して、式(7)、(8)で $Q^A_{mm}=0$、$Q^A_{ml}=0$ とされる。

4. クラスタリングの実験結果

4.1 実験データ

用いたコーパスは TDT2 である[5]。TDT2 は 1998 年の 1 月から 6 月の間の 100 個のイベント（例えば、“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”、及び電子

ニュース系の“NYT”、“APW”、“PRI”から採取されている。本報告では、表1に示すような3種類のデータを作成し、実験に用いた。データ1、2、3はそれぞれ文書数が70、40、30以上の21、31、38イベントを選択したものである。各イベントの文書数のうち最大は1484で、データ3では最小文書数の約50倍である。従って、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、文書当りの文の数は、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。

表1　実験に用いたデータ

データ	イベント数	文書数			
		総数	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

4.2 実験条件

前処理としては、文切り出しの後、品詞付け、lemmatizingを行い、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、出現頻度がデータ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、4552、5769、5914となった。

また、文書間類似度、文書・クラスター間類似度については予備実験の結果線形モデルはクラスター検出能力の点で劣ることが確認されたので、2次モデルのみを用いた。文書間もしくは文書・文書集合間の類似性を判断するうえで単語共起の情報は単語単独の生起情報よりも重要と考えられる。そこで、2次モデルで類似度を算出するときは行列$\boldsymbol{S}$、$\boldsymbol{S}^{P}$、$\boldsymbol{Q}^{A}$の非対角成分の値を一定倍（例えば5倍）して用いた。

また、クラスタリング精度は以下のように定義した。先ず、各クラスター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ーのラベルとする。次いで、各クラスターにおいてクラスター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。複数のクラスター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ー以外のクラスターに帰属する文書は全て誤りとする。

表2　クラスタリングの結果

データ	イベント数	検出クラスター数	クラスタリング精度
1	21	21	95.17%
2	31	30	95.09%
3	38	36	94.41%

4.3 実験結果

表2に実験結果を示す。クラスター検出能力に関して云えば、データ1では全てのイベントが、データ2では31イベント中30個が、データ3では38個中36がクラスターとして正しく抽出されている。なお、どのデータでも2〜3個の冗長クラスターが発生している。データ2で検出に失敗したイベントは、“State of the Union Address(大統領の年頭教書)”(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（その時期の情勢を反映し、色々な事柄に触れている）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、更に“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、パラメータの値によっては、“John Glenn”（36文書）が“Rats in Space!“（60文書）に吸収され、検出クラスター数が35になってしまう場合もある。これも内容の近いイベント同士で起こった現象である。

また、クラスタリング精度について云えば、どのデータにおいても、“Asian Economic Crisis”の内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”が引金になったイベントであり、共通する話題がかなり含まれている。表2の結果は完璧ではないが、トンデモナイ誤りも見受けられない。

表3は、データ3において“Asian Economic Crisis”に対応するクラスターのステップ3での最後の繰返しのときの（文書数1382）特有単語抽出結果、表4は同じく特有単語対の抽出結果を示し

表3　特有単語の抽出結果

	U_{mm}	U^0_{mm}	ratio	Spe	term
1	929	1416	0.66		economic
2	820	1562	0.52		crisis
3	744	892	0.83	○	economy
4	708	2077	0.34		government
5	706	2436	0.29		year
6	702	2091	0.34		country
7	677	794	0.85	○	finacial
8	662	1372	0.48		international
9	646	792	0.82	○	asia
10	644	721	0.89	○	asian
11	622	793	0.78	○	market
12	592	616	0.96	○	currency
13	590	3276	0.18		president
14	582	842	0.69		indonesia
15	559	655	0.85	○	fund
16	547	2841	0.19		make
17	546	2305	0.24		u.s.
18	544	562	0.97	○	monetary
19	528	812	0.65		billion
20	508	743	0.68		bank

表4　特有単語対の抽出結果

	U_{ml}	U^0_{ml}	ratio	Spe	term pair	
1	505	512	0.99	○	international	monetary
2	499	507	0.98	○	international	fund
3	496	503	0.99	○	fund	monetary
4	475	572	0.83	○	crisis	economic
5	359	366	0.98	○	crisis	financial
6	348	1782	0.20		state	united
7	345	349	0.99	○	crisis	asian
8	301	506	0.59		president	suharto
9	294	296	0.99	○	asia	economic
10	290	294	0.99	○	asia	crisis
11	284	288	0.99	○	korea	south
12	270	324	0.83	○	indonesia	economic
13	249	340	0.73		economic	country
14	238	970	0.25		minister	prime
15	227	232	0.98	○	market	stock
16	224	224	1.00	○	asia	financial
17	218	269	0.81		government	economic
18	214	222	0.96	○	financial	asian
19	212	242	0.88	○	economic	reform
20	212	213	1.00	○	currency	dollar

ている。表3、4でSpeの欄に○が付与された単語、単語対が特有と判断されている。例えば、単語対“international”、“monetary”は全文書7546中の512文書で共起し、その内の505文書がこのクラスターに含まれていて、集中度が高いので特有と判断された。表3、4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。

5. 考察

5.1　パラメータの影響

本手法で、性能に影響を与えうる要因には

① どのような種文書が求められたか？

② クラスターの大きさ、成長の速さを左右するパラメータの選び方

③ ラスターの共通単語の選び方

がある。

これらの内、①の影響は次のように現れる。

(A) クラスターの発育不全。即ち、種が局所的にしか成長せず、イベント全体に広がらない。

(B) 2つの内容の近いイベントのマージ。種文書が内容の近い他のイベントに近い場所で検出された場合に起きる。

(A)については実際に“Asian Economic Crisis”で起きた現象であるが、後続の処理において同じイベントの残された文書から検出された種文書がイベント全体をカバーするように成長し、発育不全のクラスターは冗長と判定されたため、性能には影響しなかった。また、(B)については、**ステップ2**において近隣文書数が2番目に多い文書を選択したときに実際に起きた現象である。“John Glenn”が“Rats in Space”に吸収されたのがその例である。しかし、種文書の違いの影響が現れたのは2、3のイベントに止まっており、大部分のイベントは影響を受けなかった。

②については、**ステップ2**において選択された種文書の近隣文書数を多くする、もしくは、**ステップ3**においてその時点でのクラスターに多くの文書を（仮に）帰属させるようにすると、クラスターの成長が促進されるようになり、反対にすると成長は抑えられる。従って、パラメータの値が適切でない場合には、過度の成長により複数のイベントがマージされやすくなるか、もしくは成長が抑制されてクラスターが発育不全を起こし、幾つかの部分クラスターに分かれてしまうことが起きる。そのためこれらのパラメータは適切に設定する必要があったが、精度がこれらのパラメータに敏感に反応することはなく、容易に決定できた。③は**2.2**における閾値Aの問題である。前述のよ

うに、閾値 A に適切な値を設定することにより、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。表5は、データ3において閾値 A を各時点でのクラスターのメンバー数のα倍に設定したときのαとクラスタリングの精度の関係を示す。表から分かるように A を適切な値に設定することによりクラスタリング精度は向上する。なお、α=0.3 とした場合には発育不全を起こし、イベント“Asian Economic Crisis”は3つの、“1998 Winter Olympics”は2つのサブクラスターに分割されてしまう。

5.2 従来手法との比較

比較相手とする従来法は、クラスター間類似度としてグループ平均法を用いたボトムアップの階層的手法、k-means 法、及びkNN法である。

ボトムアップの階層的手法では初期状態として各文書をひとつのクラスターとし、最も近いクラスター同士をマージするという処理をクラスターが一つになるまで繰り返す。k-means 法では、1)指定されたクラスター数の文書をランダムに選択して各クラスターの種とし、2)各文書について各クラスター中心との距離を求め、各文書を最も近いクラスターに帰属させ、3)各クラスターに帰属する文書ベクトルの平均により各クラスターの中心を求め、4)ステップ2)に戻って処理を続行し、各文書の帰属するクラスターに変化がなければ終了、そうでなければ 3)に戻るという方法である。kNN 法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。ここでのkNN 法は各文書につき他の全文書との類似度を求めたうえで最も類似度の大きいk個の文書を求め、k個の各文書が属するクラス毎にその文書の類似度を加算していき、最も類似度の総和が大きくなったクラスを選択するというものである[6]。

ボトムアップの階層的手法では、理想的には、

- 同じ話題のクラスターがマージされた後、異なるイベントのクラスターのマージが始まる。
- 同じイベントのクラスターがマージされる時にはそれらのクラスター間の類似度は大きく、異なるイベントの場合には小さくなって、次々にマージされるクラスター間の類似度の系列の解析からイベントの数を推測できる。

となることが望ましい。しかし、実際には多くの微小クラスター（例えば文書数20以下）が生じ、これら微小クラスターが同じイベントの大きなクラスターにマージされる前に異なるイベントのクラスターのマージが始まっていた。また、次々にマージされるクラスター間の類似度の系列の解析からイベントの数を推測することも不可能であった。そのため、ここではクラスタリング精度が最大になるときのクラスタリング精度とクラスター数、及びクラスター数が既知とした場合、即、クラスター数としてイベント数を採用した場合の精度で評価する。表6はその結果を示す。表から分かるように、クラスタリング精度が最大になる場合でも提案手法のそれよりも低く、またクラスター数も異常に多くクラスタリング検出精度は 非

表5　共通単語選択閾値とクラスタリング精度の関係

α	0	0.05	0.1	0.15	0.2	0.25
クラスタリング精度(%)	93.88	94.17	94.21	94.41	93.68	93.53

表6　階層形クラスタリング法の結果

データ	精度を最大とする場合		クラスター数=イベント数の場合のクラスタリング精度
	クラスタリング精度	クラスター数	
1	91.57%	74	78.56%
2	92.87%	71	85.63%
3	88.34%	95	85.26%

表7　k-means法の結果

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

表8　kNN法の分類結果

データ	分類精度
1	98.03%
2	97.58%
3	97.02%

常に低い。また、正確なクラスター数が分かっているとした場合には、クラスタリング精度は提案手法よりも著しく劣ってしまう。

kの値をイベント数とした場合のk-means法の結果を表7に示す。表7でクラスタリング精度に幅があるのは、クラスタリング精度がクラスターの種の選び方に依存するためである。k-means法では、文書数の多いイベントは複数のクラスターに分割され、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7に示されるようにクラスター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、クラスター検出精度共優れている。特に、クラスター検出能力は従来法に比べ著しく改善されている。この結果は、クラスターの種文書は殆どの場合同じ話題の文書の範囲内で過不足無く成長を終えるため、後続の処理で次のクラスターの検出が容易になったことを物語っている。

表8は2重交差検定による文書分類結果である。即ち、データの半数を訓練用に、残りをテスト用に用いた実験をデータを入れ替えて2回行っている。たとえ半分とはいえ、各クラスの文書が例示されている状態で行う文書分類に比べ、クラスタリングではデータの各クラスの情報について事前に全く分かっていない。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、実験結果では著しく劣るということはなく、むしろ善戦していると云える。

6. まとめ

本報告では、TDT2のコーパスから選択した21～38イベント6788～7546文書を用いたクラスタリング実験、従来手法との比較を行い、先に提案した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の評価とした。確認された主な点は以下の通りである。

- 従来文書クラスタリングに多く用いられているグループ平均法を用いた階層的手法、k-meansはクラスターの数を正しく検出する能力に大きな問題を抱えているのに対し、提案手法では高い能力を示している。21イベントの場合は全て正しく、31イベントの場合は30、38イベントの場合は36イベントがクラスターとして検出されている。
- クラスタリング精度としては94～95%が得られ、提案手法は従来のクラスタリング法に比べて著しく高いほか、訓練文書が与えられる文書分類に比べても殆ど遜色がない。

提案手法のこのような特長は、単語共起の情報も用いた文書・クラスター間類似度尺度を用いたこと、クラスタリングの過程でその時点での特有単語、単語対を選択的に用い、他の話題の文書が混入することを防いだことによってもたらされたものである。

今後の課題は以下の通りである。提案手法で正しくクラスタリングされたとは云え、例えば“Current Conflict with Iraq”に対応するクラスターでは1500前後の文書を含んでいる。これだけの数の文書集合では、単に一つのトピック“Current Conflict with Iraq”ではくくれない様々なサブトピックが存在するものと考えられる。これらサブトピックを正しく抽出することは今後の重要な課題である。また、文書は多くの場合ストリームの形で入力されるのでオンラインクラスタリングは必須である。提案手法をオンラインクラスタリングに適用することも重要な課題である。

参考文献

[1] C. D Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[2] 徳永健伸. 情報検索と言語処理. 東京大学出版会(1999).

[3] 川谷隆彦. 多文書間の共通性分析による文書クラスタリング. 情報処理学会自然言語処理研究報告, 2003-NL-154, pp.93-100(2003).

[4] 川谷隆彦. 多文書間の共通性の分析. 情報処理学会自然言語処理研究報告, 2002-NL-152, pp.85-92(2002).

[5]http://www.ldc.upenn.edu/Projects/TDT3/TDT3_Overview.html

[6] Y. Yang and X. Liu. Re-examination of Text Categorization. In *Proceedings of the 22nd Annual International ACM SIGIR Conference.on Research and Development in Information Retrieval (SIGIR'99)*, pp.43-49, 1999.