2005－DBS－137（Ⅰ）（27）
2005／7／13

差異を意識したクラスタリングとその特徴量集約手法の検討

渡辺 匡[†]　大野 成義[††]　太田 学[†††]　片山 薫[†,††††]　石川 博[†,††††]

[†]首都大学東京大学院工学研究科 〒192-0397 東京都八王子市南大沢 1-1

[††]職業能力開発総合大学校　〒229-1196 神奈川県相模原市橋本台 4-1-1

[†††]岡山大学大学院自然科学研究科　〒700-8530 岡山市津島中 3-1-1

[††††]首都大学東京システムデザイン学部 〒192-0397 東京都八王子市南大沢 1-1

E-mail: †wata300@jcom.home.ne.jp, ††ohno@cs.uitec.ac.jp, †††ohta@suri.it.okayama-u.ac.jp, ††††{katayama,ishikawa}@eei.metro-u.ac.jp

あらまし　膨大な数のデータを扱う際にクラスタリングは非常に有用な手段である。ユーザがクラスタリング結果から求める情報を得るためにはクラスタ間の関係が直観的に理解できることが望ましい。そこで本稿ではクラスタ間の差異を分析して視覚化する方法を提案する。視覚化のための手段としてMDS(MultiDimensional Scaling)に基づく二次元地図表示を用いる。さらにそこで利用する特徴量の集約手法について複数の方法を検討する。すなわち、要素のtf-idf重み付けに基づく集約手法とMEDRANKアルゴリズムに基づくランク集約手法の検討を行う。

キーワード　**クラスタリング，集約化**

On difference-conscious clustering and its feature aggregation methods

Masashi WATANABE[†]　Shigeyoshi OHNO[††]　Manabu OHTA[†††] Kaoru KATAYAMA[†,††††] and Hiroshi ISHIKAWA[†,††††]

[†]Tokyo Metropolitan University, Engineering 1-1 Minami-Ohsawa, Hachioji-shi, Tokyo, 192-0397 Japan

[††]Department of Information and Computer Science, Polytechnic University, 4-1-1 Hashimotodai, Sagamihara, Kanagawa, 229-1196 Japan

[†††]Graduate School of Natural Science and Technology, Okayama University, 3-1-1 Tsushima-naka, Okayama-shi, Okayama, 700-8530 Japan

[††††]Tokyo Metropolitan University, System Design 1-1 Minami-Ohsawa, Hachioji-shi, Tokyo, 192-0397 Japan

E-mail: †wata300@jcom.home.ne.jp, ††ohno@cs.uitec.ac.jp, †††ohta@suri.it.okayama-u.ac.jp, ††††{katayama,ishikawa}@eei.metro-u.ac.jp,

Abstract　Clustering is very useful analyzing a huge amount of data. It is preferable that we can understand relationships between clusters intuitively in order to reach necessary information within the clustering results. Then, we propose a method of analyzing differences between clusters by using visualization. We a two-dimensional map display based on MDS(MultiDimensional Scaling) for visualization in this article. In addition, we examine two methods for aggregating the features used for clustering, one of which is based on the tf-idf weight of feature terms and the other on a rank aggregation method called the MEDRANK.

Keyword　**Clustring, Aggregation**

1. はじめに

大量の文書データを効率的に扱う手段として、クラスタリングは有用な手段である。特に、ユーザが漠然とした検索要求を持ち、様々な話題の中から目的の情報について調べようとしている時、データがクラスタリングによってあるカテゴリーに分類されていれば、欲しい情報を持つクラスタに容易

に問い合わせることが可能となる。

ユーザがクラスタを適切に選択するために、クラスタリング結果の直観的な分析が求められる。この分析を補助する手段としては、一般的に各クラスタの持つ特徴量を集約し中心的な話題を表す手法が用いられる。ただ、これは分割されたクラスタのそれぞれの結果を知ることには有効ではあるが、これは個々のクラスタにのみ注目しているのでクラスタ間の関係までは分からない。各クラスタの関係性を知ることは個々の話題についての類似性や差異性を明らかにすることにもつながり、非常に重要であると考えられる。

そこで本手法では、従来の手法に加えて各クラスタ間の関係について示すことを考える。特にクラスタ間の差異について分析し、これを視覚化する手法を提案する。

すでに Scatter/Gather[5]のような同じカテゴリーに属する文書について差異を比較する方式は研究されているが、本研究ではカテゴリーの違いに関わらず差異の比較検討が可能であるという点で従来の研究と異なっている。

さらに、中心的な話題を表すための特徴量の集約手法についても複数の方法を検討する。一つは要素の tf-idf[1][2]重み付けに基づく集約手法、もう一つは MEDRANK[3]アルゴリズムに基づくランク集約手法である。

本稿の構成は次の通りである。第 2 章では提案手法の概要と関連研究について述べる。第 3 章では差異分析クラスタリングについて、第 4 章では実験と評価について述べ、最後に第 5 章では考察と今後の課題について述べる。

2. 提案手法の概要と関連研究

2.1. 概要

クラスタの差異を視覚化する手段として、MDS (MultiDimensional Scaling)[4]に基づく二次元地図表示を利用する。これはクラスタ間の距離に基づき多次元データを二次元地図として表現するものである。類似度の高いクラスタ同士は近くに、類似度の低いクラスタ同士は遠く離れた位置に配置され、視覚的理解が容易になるという利点を持つ。

しかし、全てのクラスタについて厳密にクラスタ間の距離関係が保持された二次元地図を作成することは計算量が$O(N^3)$あるいは$O(N^4)$と多く困難である。よって本手法では特に距離が小さく結びつきの強いクラスタ間の距離関係が保持されることを重視した近似的手法を用いる。

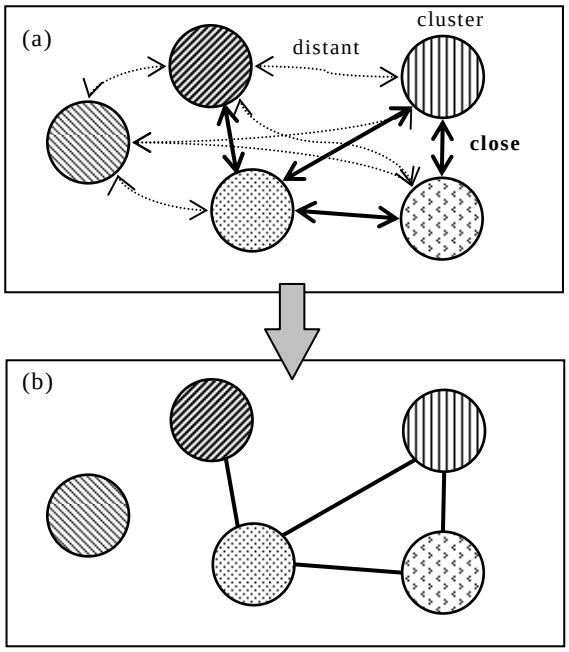

図 1: 本手法における二次元地図表示

これは距離が小さいと判断されるクラスタ同士の地図上の距離の正確さを重視して計算・配置を行う。そして距離が小さいクラスタ同士は直線で結ばれ、これによってクラスタ間の距離や関連性の強さを明示する。つまり、直線で結ばれたクラスタ同士はある程度結びつきが強い。

ただし、この手法では距離が大きく直線によって結ばれていないクラスタ同士については地図上の距離が正確であることを保証しない。

概要を図 1 に示す。円はクラスタを表す。図 1(a)のように 5 つのクラスタが与えられた時、直線で示される close の関係にあるクラスタ間の距離の正確さを重視し、点線で示される distant の関係にあるクラスタ間の距離が不正確であることを表している。そこで、地図表示の出力としては図 1(b)のように与えられ、クラスタ間に引かれた直線の有無によって関連性の有無を示し、直線の長さによって関連性があるとされたクラスタ同士の距離関係を示す。

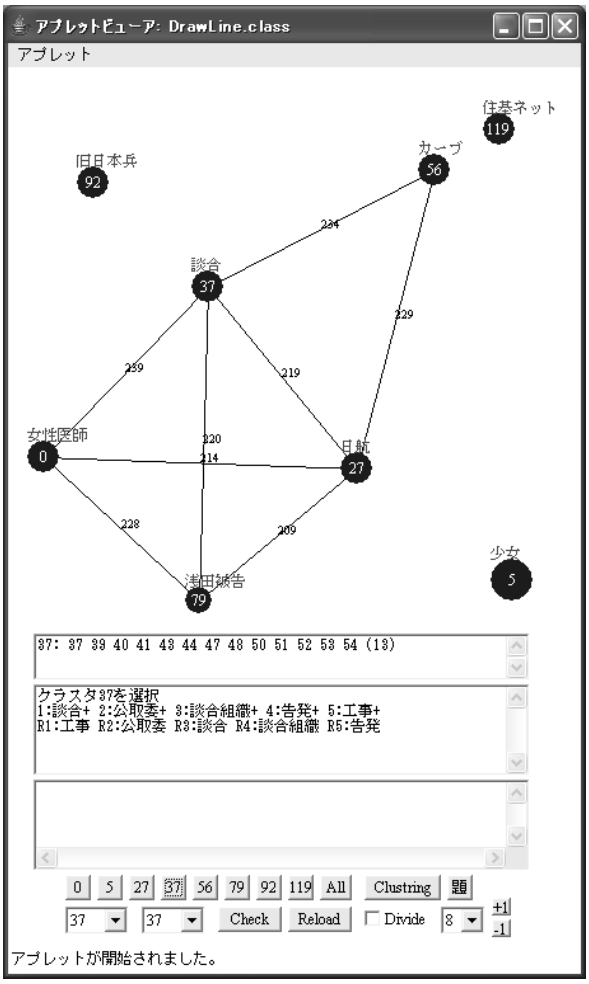

図 2: 本システムの実行例

この具体例について実行例を図 2 に示す。図 2 の画面上部に表示されているのが本手法による二次元地図表示である。図 1(b)と同様、クラスタ間の関係を直線による結びつきと配置によって視覚的に表現している。この例では、クラスタ 5、クラスタ 92、クラスタ 119 は他クラスタとの距離が大きく離れている。この例についての詳細は第 4 章に述べる。

さらに差異の分析のための手段として、ユーザによって選択された任意のクラスタの持つ要素を新たな特徴空間として扱い、再びクラスタを生成する手法を用いる。概要を図 3 に示す。選択されたクラスタが特徴 f_1,f_2,f_3 を持ち、別のクラスタが特徴 f_1,f_2,f_4,f_5 を持つとき、選択されたクラスタに特有の特徴、すなわち特徴 f_1,f_2,f_3 を新たな特徴空間として再クラスタリングをすることで、ユーザにとって興味のある話題に対してより細かな分析が可能となる。

ユーザがクラスタを選択するための手がかりとして、二次元地図表示の他にクラスタの持つ情報からクラスタの典型的な特徴になると考えられる特徴を与える。

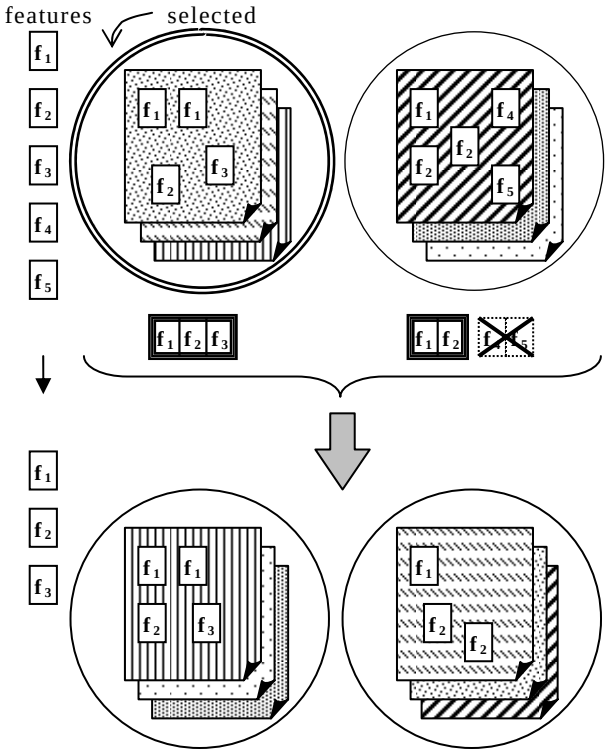

図 3: 新たな特徴空間による再クラスタリング

特徴量集約手法として文献[7]では tf-idf に基づく集約手法を用いていたが、本手法では、tf-idf に基づく集約手法と、MEDRANK アルゴリズムに基づくランク集約手法の 2 つを用い、比較検討する。

2.2. 関連研究

Web 文書の検索結果などの多量な情報を扱う手法として、クラスタリングとユーザによるクラスタの選択の繰り返しによって欲しい情報に誘導する Scatter/Gather[5]がある。これは、入力された情報の集合を 10 のクラスタに分類し、各クラスタの持つ文書に頻出する単語を話題単語として各クラスタに付加し、表示する。ユーザは興味のあるクラスタを選択すると、今度は選択されたクラスタに分類された情報のみを対象として再びクラスタを生成する。これを繰り返すことでユーザは漠然とした興味からでも得たい情報にたどり着くことができるという手法である。

これは本手法の考え方と同様の考え方で

はあるが、Scatter/Gather では選択されたクラスタのサブクラスタを扱うのに対し、我々の手法では特徴空間を変化させることで異なるクラスタの情報をも含んだ新たなクラスタの生成を行う点が異なっている。

また、Scatter/Gather では各クラスタの持つ話題単語だけを手がかりとしてクラスタの選択を行うが、我々の手法ではクラスタ間の差異の二次元地図表示を手がかりとすることでクラスタの選択のためのより直観的な理解が期待できる。

3. 差異分析クラスタリング

3.1. 手順

本手法は、N件の文書集合$D:\{d_1,d_2,...,d_N\}$が与えられた時、次に示す**ステップ 1〜6** のプロセスで動的にクラスタリングし、二次元地図表現を行う。

(ステップ 1)

Dに対し、形態素解析を行い、特徴語を得る。M個の特徴語からなる特徴語集合$F:\{f_1,f_2,...,f_M\}$を抽出する。

(ステップ 2)

特徴語f_iの重要度を求め、Dを特徴ベクトルとして表現する。

(ステップ 3)

クラスタリングを行い、階層構造を獲得する。

(ステップ 4)

得られた階層構造を元にクラスタを取得し、MDS に基づく二次元表示を行う。各クラスタの距離を元に二次元配置し、表示するものとする。

(ステップ 5)

全てのクラスタについてクラスタ間の差異となる特徴を計算する。

(ステップ 6)

任意のクラスタを選択し、選択されたクラスタの持つ特徴空間を用いて再クラスタリングを行い、上記の**ステップ 3〜6** を繰り返す。

以下の節ではこの手順についての詳細を述べる。

3.2. 特徴語の抽出

ステップ 1 では「茶筌[9]」を利用して文書の形態素解析を行い、得られた品詞情報を利用して各文書に含まれる名詞・未知語を特徴語候補として抽出する。

ここで、必要に応じて数詞・代名詞などの不要語の除去を行った上で特徴語とする。名詞・未知語が連続して出現する時には語の連結を行いひとつの特徴語として扱う。このM個の特徴語の集合を$F:\{f_1,f_2,...,f_M\}$とする。

3.3. 特徴語の重み付け

ステップ 2 では文書d_jを特徴ベクトルとして表現するため、タームf_iの文書d_jにおける重要度$w_{i,j}$をtf-idfによる重み付けで表す。ここで、ターム頻度$tf_{i,j}$は文書d_jにおけるf_iの出現頻度、文書頻度df_iはf_iが出現する文書数である。逆文書頻度idf_iはdf_iの逆数であり、$\log \frac{N}{df_i}$で表される。このとき$w_{i,j}$は、

$$w_{i,j} = tf_{i,j} \cdot idf_i$$

で表される。

これにより文書d_jのベクトル$\vec{d}_j$は

$$\vec{d}_j = \left(w_{1,j}, w_{2,j}, ..., w_{M,j}\right)$$

と表される。

3.4. クラスタリング

ステップ 3 では得られた特徴ベクトルを元にクラスタリングを行う。

クラスタリングの方法として、階層的クラスタリング(hierarchical clustering)と非階層的クラスタリング(non-hierarchical clustering)に大きく分類することができる。本手法では階層的クラスタリングの一種である階層的凝縮型クラスタリング(HAC, hierarchical Agglomerative Clustering)[6]を利用する。

はじめに 1 文書だけを含むN個のクラスタがある初期状態を作る（Nは文書数）。特徴ベクトルより文書d_iとd_jのユークリッド平方距離$D(d_i, d_j)$が次の式で表される。

$$D\left(d_i, d_j\right) = \sum_{k=1}^{M} \left(w_{k,i} - w_{k,j}\right)^2$$

これを初期状態におけるクラスタ間の距離として与える。

クラスタが 1 つになるまで以下を繰り返

す。

(1) 距離の近い二つのクラスタを逐次的に併合

(2) 併合によってできたクラスタと他のクラスタ間の距離$D(C_i, C_j)$を再計算する

これによって階層構造を獲得する。

クラスタ間の距離の計算方法として Ward 法を用いる。Ward 法による距離の再計算式は次のように与えられる。

$$D(C_p, C_q) = \frac{n_p n_q}{n_p + n_q} \sum_{j=1}^{M} \left(x_j^{(p)} - x_j^{(q)} \right)^2$$

n_pはクラスタC_pの要素数、n_qはクラスタC_q の要素数を、ベクトル$x_j^{(p)} = \{x_1^{(p)}, x_2^{(p)}, \ldots, x_M^{(p)}\}$は$C_p$の重心、ベクトル$x_j^{(q)} = \{x_1^{(q)}, x_2^{(q)}, \ldots, x_M^{(q)}\}$は$C_q$の重心を表す。

こうして得られた階層構造を元に k 個のクラスタを取得する。また k は事前に与える。

3.5. 二次元地図表示

ステップ 4 では MDS(MultiDimensional Scaling)に基づく二次元表示を行う。まず二次元の初期配置として各クラスタを任意に定める。これを適切な配置におくため、クラスタ間の距離の大小関係保持を考慮した目的関数を次のように与える。

$$S = \sqrt{\frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{i<j} d_{ij}^2}}$$

対象i,j間のユークリッド距離をd_{ij}とし、$\hat{d}_{ij}$ を距離の大小関係が保持されていない時に補正される距離とする。S=0 の時、クラスタ間の距離の大小関係が完全に保持されていることになる。しかし解の存在は保証されておらず、計算量も$O(N^3)$あるいは$O(N^4)$となり困難であるため、Sが最小となるような近似解を求める。

本手法では全てのd_{ij}のうち任意の閾値λよりも小さい値を持つものを重視して計算する。d_{ij}がλよりも小さいような距離をd'_{ij}, それ以外のd_{ij}を$\overline{d'_{ij}}$とした時、目的関数を次のように与える。

$$S' = \sqrt{\frac{\sum_{i<j} \left(d'_{ij} - \hat{d}'_{ij} \right)^2}{\sum_{i<j} d'^2_{ij}}} + \alpha \sqrt{\frac{\sum_{i<j} \left(\overline{d}'_{ij} - \overline{\hat{d}}'_{ij} \right)^2}{\sum_{i<j} \overline{d}'^2_{ij}}}$$

α は 1 未満の十分小さい数を与える。クラスタの座標配置を逐次更新することによりS'を最小化する近似解を求める。

3.6. 特徴量集約

ステップ 5 では得られたクラスタの中心的な話題を示す手段として特徴量集約手法を用いる。本手法では要素の tf-idf 重み付けに基づく集約手法と MEDRANK アルゴリズムに基づくランク集約手法の 2 つの手法を用いる。

tf-idf重み付けに基づく集約手法では各クラスタの特徴量の重要度を各文書の特徴量の重要度の平均から求める。すなわちクラスタC_pの特徴量の重要度は次式で与えられる。

$$w_{i,C_p} = \frac{1}{n_p} \sum_{j \in C_p} w_{i,j}$$

クラスタごとに重要度の高いいくつかの特徴語を表示する。

一方、MEDRANKアルゴリズムに基づくランク集約手法では、まず各文書d_jでの特徴量f_iのランクを重要度$w_{i,j}$によって定める。d_jによってf_iに与えられたランクを$v_{i,j}$と表す。次に、与えられたランク$v_{i,j}$に基づき特徴量f_iのクラスタC_pにおけるランク$v_{i,Cp}$を決定する。$v_{i,Cp}$はクラスタC_p内の全ての文書から与えられたランク全体の中間ランク(Median rank)として集約され、クラスタの特徴語のランクとして与えられる。すなわちクラスタC_pでの特徴量f_iのランク$v_{i,Cp}$は次式で与えられる。

$$v_{i,c_p} = median(v_{i,p_1}, v_{i,p_2}, \ldots, v_{i,p_{n_p}})$$

(ただし $d_{p_j} \in C_p$, n_pはC_pの要素数)

上位に位置するいくつかの特徴語を表示する。

MEDRANK では重要度ではなくその順位を利用するので、効率的に良い結果が得られると考えられる。そこで本手法ではこれを特徴量集約手法として利用し、実験によって tf-idf 手法と MEDRANK 手法の有効性について評価した。結果は第 4 章に与える。

表 1：tf-idf と MEDRANK の特徴量集約手法の比較

	クラスタ 0			m=10					
	tf-idf			MED RANK			見出し		
順位	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m
1	女性医師	3	30.0%	女性医師	3	30.0%	書類送検	7	70.0%
2	人工呼吸器	2	20.0%	患者	1	10.0%	医師	4	40.0%
3	医師	4	40.0%	人工呼吸器	2	20.0%	殺人容疑	4	40.0%
4	男性患者	0	0.0%	男性患者	0	0.0%	女性医師	3	30.0%
5	道警	1	10.0%	家族	0	0.0%	殺人	3	30.0%
	Average		20.0%			12.0%			42.0%

	クラスタ 5			m=194					
	tf-idf			MED RANK			見出し		
順位	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m
1	少女	8	4.1%	調べ	1	0.5%	小林容疑者	17	8.8%
2	小林容疑者	17	8.8%	事件	2	1.0%	少女監禁	12	6.2%
3	監禁	5	2.6%	疑い	1	0.5%	書類送検	11	5.7%
4	金属片	11	5.7%	逮捕	9	4.6%	金属片	11	5.7%
5	事故	1	0.5%	監禁	5	2.6%	ガードレール	8	4.1%
			4.3%			1.9%			6.1%

	クラスタ 27			m=15					
	tf-idf			MED RANK			見出し		
順位	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m
1	日航	7	46.7%	認可	0	0.0%	日航	7	46.7%
2	整備	3	20.0%	グループ会社	1	6.7%	旅客機	5	33.3%
3	認可	0	0.0%	整備	3	20.0%	無認可整備	5	33.3%
4	無認可整備	5	33.3%	旅客機	3	20.0%	経産省	4	26.7%
5	経産省	4	26.7%	日航	7	46.7%	牛肉偽装事件	4	26.7%
			25.3%			18.7%			33.3%

	クラスタ 37			m=13					
	tf-idf			MED RANK			見出し		
順位	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m
1	談合	2	15.4%	工事	0	0.0%	橋梁談合	9	69.2%
2	公取委	1	7.7%	公取委	1	7.7%	着手	4	30.8%
3	談合組織	0	0.0%	談合	2	15.4%	談合	2	15.4%
4	告発	2	15.4%	談合組織	0	0.0%	捜索	2	15.4%
5	工事	0	0.0%	告発	2	15.4%	強制捜査	2	15.4%
			7.7%			7.7%			29.2%

	クラスタ 56			m=25					
	tf-idf			MEDRANK			見出し		
順位	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m
1	カーブ	3	12.0%	軽装	2	8.0%	ATS	7	28.0%
2	ノーネクタイ	4	16.0%	ノーネクタイ	4	16.0%	クールビズ	7	28.0%
3	ネクタイ	1	4.0%	職員	1	4.0%	ノーネクタイ	4	16.0%
4	軽装	2	8.0%	上着	0	0.0%	カーブ	3	12.0%
5	職員	1	4.0%	全国	1	4.0%	回転扉事故	2	8.0%
			8.8%			6.4%			18.4%

	クラスタ 79			m=4					
	tf-idf			MEDRANK			見出し		
順位	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m
1	浅田被告	1	25.0%	浅田被告	1	25.0%	浅田元会長	2	50.0%
2	農水省	0	0.0%	農水省	0	0.0%	牛肉偽装	2	50.0%
3	公判	0	0.0%	犯行	0	0.0%	ドン	2	50.0%
4	反省	1	25.0%	公判	0	0.0%	浅田被告	1	25.0%
5	ハンナン	1	25.0%	懲役	1	25.0%	ハンナン	1	25.0%
			15.0%			10.0%			40.0%

	クラスタ 92			m=17					
	tf-idf			MEDRANK			見出し		
順位	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m
1	旧日本兵	10	58.8%	フィリピン	1	5.9%	旧日本兵	10	58.8%
2	フィリピン	1	5.9%	帰国	0	0.0%	生存	5	29.4%
3	帰国	0	0.0%	旧日本兵	10	58.8%	比	5	29.4%
4	師団	1	5.9%	師団	1	5.9%	元日本兵	4	23.5%
5	ミンダナオ島	1	5.9%	ミンダナオ島	1	5.9%	面会	4	23.5%
			15.3%			15.3%			32.9%

	クラスタ 119			m=27					
	tf-idf			MEDRANK			見出し		
順位	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m	単語	c(f)	c(f)/m
1	住基ネット	26	96.3%	住基ネット	26	96.3%	住基ネット	26	96.3%
2	原告	2	7.4%	原告	2	7.4%	金沢地裁	8	29.6%
3	個人情報	4	14.8%	離脱	1	3.7%	削除	7	25.9%
4	判決	6	22.2%	個人情報	4	14.8%	違憲	7	25.9%
5	削除	7	25.9%	判決	6	22.2%	訴訟	7	25.9%
			33.3%			28.9%			40.7%
	All Average		16.2%			12.6%			31.8%

3.7. 差異分析

全てのクラスタについてクラスタ間の差異となる特徴を計算する。各クラスタの持つ特徴量 w の差の絶対値に関して大きな値を取る上位数個を表示する。選択された任意のクラスタの持つ特徴語を用いて再クラスタリングを行い 3.1 節の**ステップ 3〜6** を繰り返す。

4. 実験と評価

4.1. データセット

本実験では、Google News 日本版[8]から収集したニュース記事 305 件を用いた。ニュース記事は 1 件あたり平均 1229Byte（約 614 文字）のテキストファイルである。

4.2. 特徴量集約手法の評価

まず、特徴量集約手法の評価を与える。上記の 305 件のニュース記事を 8 つのクラスタに分類し、tf-idf、MEDRANK の 2 つの手法によってそれぞれ上位 5 語の単語をそのクラスタの話題単語として抽出した。

ここで、各クラスタ内のニュース記事の見出しに登場する単語を評価尺度として利用する。話題単語 f と一致する語を見出しに持つ文書の数を c(f)、クラスタ内の全文書数を m とする時、再現率を c(f)/m で示す。

なお各クラスタには便宜的にクラスタの持つ文書のうちもっとも小さい記事番号が名前として与えられている。

表 1 に tf-idf、MEDRANK で得られたクラスタの話題単語の上位 5 つと、クラスタ内のニュース記事で話題単語と一致する語を見出しに持つような文書の数を c(f)、ここから算出される再現率 c(f)/m を示した。また、見出しに頻出する単語を「仮想的正解」とみなし、見出しでの登場頻度の高い上位 5 つの単語とその再現率を併記した。

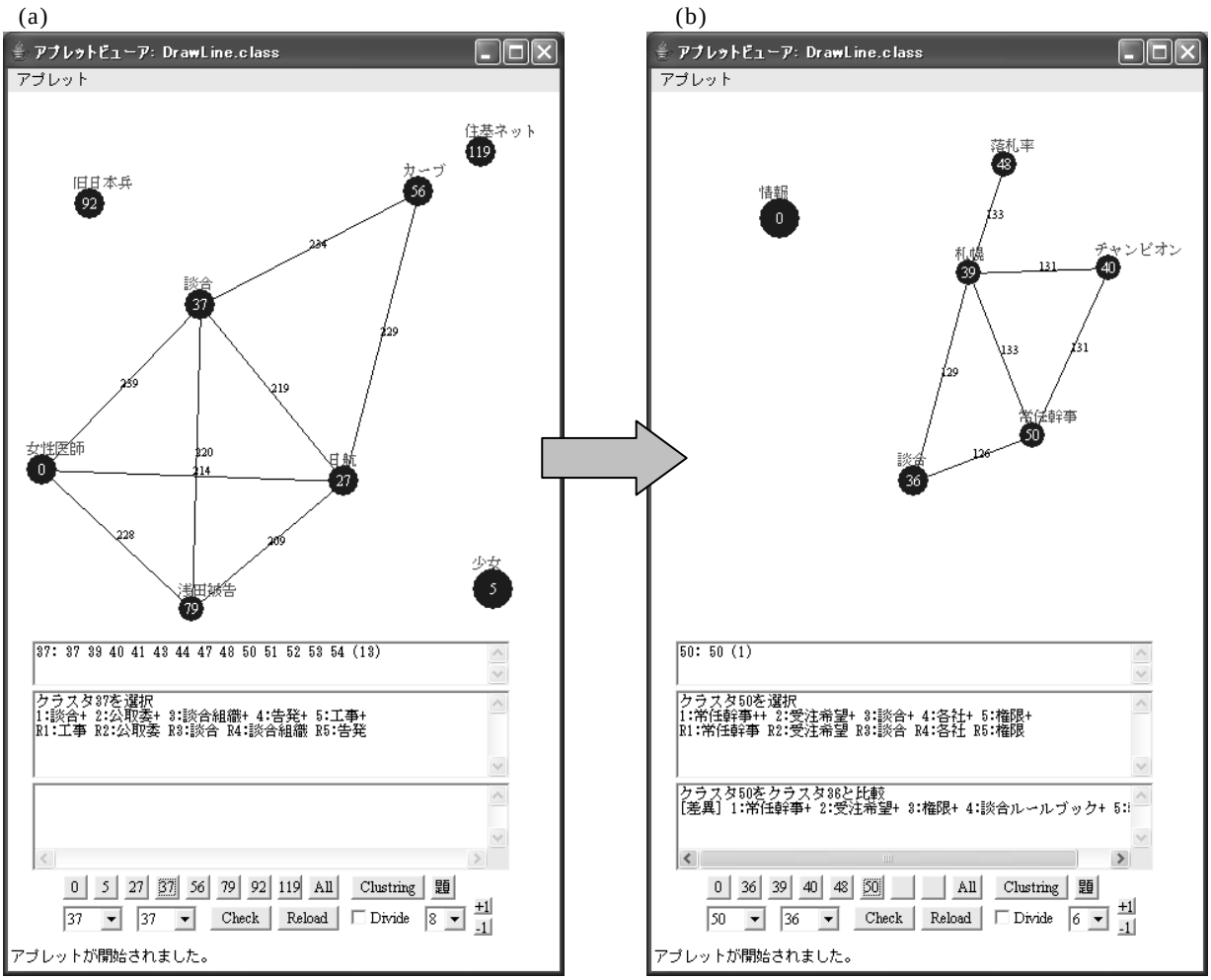

図 4: 再クラスタリング実行例

この評価尺度で全体の平均再現率は、tf-idf では 16.2%、MEDRANK では 12.6%、見出しによる「仮想的正解」では 31.8%という結果となった。

MEDRANK が tf-idf よりも悪い結果になった原因のひとつとして、本実験で用いた評価尺度である「見出し」の抱える問題点が挙げられる。見出しはしばしば簡略化された語が用いられるため、話題単語と一致しないことがある。たとえば、表 1 のクラスタ 92 では「見出し」の第 3 位として「比」が登場している。これは「フィリピン」と同じ意味であるが、見出しでは文字数の多い「フィリピン」を避けて「比」という一字で代用していたものである。このような見出し特有の語は文中ではあまり用いられないため、適切な話題単語としては「比」よりも「フィリピン」の方が良いと考えられる。

また、MEDRANK と tf-idf を比較すると、クラスタ 5 では「調べ」を第 1 位とし、クラスタ 27 では見出しに登場する「無認可整備」ではなく「認可」を第 1 位としているなど、見出しには登場しない抽象的な語を集約する傾向が見られる。これは見出しにはない曖昧な語によって話題単語を示したい場合には有効だと考えられる。

4.3. 再クラスタリングの実験

次に再クラスタリングの実験例を図 4 に示す。図 4(a)は先ほど図 3 にも示した図である。今、ユーザは談合に関連する記事を要求しており、橋梁談合疑惑に関する記事を持つクラスタ 37 を選択したとする。そこで、本プログラムはユーザが選択したクラスタに特有の空間で再クラスタリングを行う。この選択されたクラスタ 37 に特有の空間を用いて再クラスタリングを実行した結果が図 4(b)である。新たにクラスタ 50 という単一の文書からなるクラスタが生成され、このクラスタがクラスタ 36 と比較して「常任幹事」「受注希望」「権限」「談合ルールブック」などの差異を持つことが分かる。このように再クラスタリングによってある話題についてより詳細に差異を発見できることが示された。

5. おわりに

本研究では差異を意識したクラスタリングとしてクラスタリングの二次元地図表示による視覚化表現の方式を提案した。提案手法によってユーザが求める情報を抽出するための手助けになることを示した。他手法との比較については今後の課題である。

また、本研究では tf-idf,MEDRANK の 2 つの手法をクラスタの特徴量集約手法として利用した。MEDRANK は tf-idf に比べ再現率がやや低い結果となったが、その有効性は十分高いと思われる。MEDRANK はあらゆる特徴をランクとして扱うため、異なる指標の比較も可能であるなど汎用性の高い手法でもある。他分野への MEDRANK の有効性の調査を今後の課題として検討したい。

文　献

[1] Salton, G. and Buckley, C.: “Term-weighting approaches in automatic text retrieval,” Information Processing and Management, 24, pp.513-523, 1988d.

[2] Salton, G. and Buckley, C.: “Improving retrieval performance by relevance feedback,” Journal of the American Society for Information Science, 41(4), pp.288-297, 1990.

[3] Ronald Fagin, Ravi Kumar, D. Sivakumar: “Efficient similarity search and classification via rank aggregation”, SIGMOD 2003: pp. 301-312, 2003.

[4] Young, F.W., Hamer, R.M., “Theory and applications of multidimensional scaling.”, Hillsdale, NJ: Erlbaum, 1994.

[5] Marti A. Hearst, Jan O. Pedersen, “Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results”, in Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (SIGIR'96), pp. 76-84, 1996.

[6] Lance, G. N. and Williams, W. T., “A general theory of classificatory sorting strategies, I. Hierarchical systems”, The Computer Journal, 9, 373-380, 1967.

[7] 渡辺匡, ”文書間の差異に着目したクラスタリング手法”, 第 67 回情報処理学会全国大会,2005.

[8] Google News : http://news.google.co.jp/ Accessed 2005.

[9] 茶筌 : http://chasen.naist.jp/hiki/ChaSen/ Accessed 2005.