The 27th Symposium on Information Theory
and Its Applications (SITA2004)
Gero, Gifu, Japan, Dec. 14-17, 2004

ドキュメントの特徴を考慮した非階層的クラスタリング
Nonhierarchical Clustering Based on Features of Documents

渡辺 祐介* 細谷 剛* 平澤 茂一*
Yusuke WATANABE Gou HOSOYA Shigeichi HIRASAWA

Abstract— Document clustering is a method for improving efficiency and effectiveness for information retrieval and text mining. As growing of importance of electronic media for dealing large textual databases, the document clustering becomes more significant. A method of hierarchical document clustering is inadequate for large document databases in terms of its time complexity. Besides when each document is characterized by only several terms or keywords, clustering algorithms often produce poor results. So a method of nonhierarchical document clustering based on tolerance rough set models have been proposed to adapt these problems. Though it is more suitable for documents represented by keywords whose weights are the same degree, it is inadequate for general documents consisted of terms with a wide range of weight.

We propose a nonhierarchical document clustering algorithm for general documents. We show the validity of the proposed clustering algorithm by simulation rusults with the test collection.

Keywords— Clustering, Text Mining, Rough Set Theory

1 はじめに

ドキュメントクラスタリングはドキュメントをいくつかのクラスタにグルーピングする技術であり，検索やテキストマイニングにおいての検索効率や分類精度を向上させるものである．そのため近年，大規模な電子媒体の保管や取引の重要性が高まる中でドキュメントクラスタリングの重要性も増している．

ドキュメントの処理には，階層的ドキュメントクラスタリングと呼ばれるドキュメントを階層的にクラスタリングする手法が存在する．しかし計算量の観点で非効率的なため，大規模なドキュメントデータには不適であると言われている．また，ドキュメントを表現する方法はタームを要素として用いるのが一般的であるが，タームの出現頻度がドキュメントの要素の次元に対して非常に小さい場合，ドキュメント同士やドキュメントとクラスタ間の類似度の値がゼロに近づき，適切な処理を行えない可能性もある．そのため，計算量やドキュメントのデータスパースネスの問題に効果があるラフ集合理論を応用したトレランスラフ集合モデル (TRSM) に基づく非階層的クラスタリング手法 [1] が提案されている．しかし，その有効性はドキュメントのタームがそれぞれの重みに大きな差がない場合で，タームの重みに差がある一般的な文章に応用した場合，各クラスタの代表元が中心に集まってしまい，クラスタリングがうまく行われない可能性が生じる．本研究では，各クラスタが中心に集まるのを防ぐため，クラスタ内の特徴的なドキュメントに注目したクラスタ代表元を構築する手法を提案する．テストコレクションを用いてシミュレーションを行い，提案手法のアルゴリズムの有効性を示す．

*〒 169-8555 東京都新宿区大久保 3-4-1 早稲田大学理工学部経営システム工学科 School of Science and Engineering,Waseda University, Okubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 Japan. E-mail: watanabe@hirasa.mgmt.waseda.ac.jp

2 準備

2.1 ドキュメント [1]

本節では，ドキュメントの定義について述べる．ドキュメント数 M のドキュメント集合は，$\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ で表される．また，各ドキュメント d_j は，重み $w_{ij} \in [0,1]$ で特徴づけられたターム t_i で $d_j = (t_1, w_{1j}; t_2, w_{2j}; \ldots; t_N, w_{Nj})$ とされ，D からの全 N 個のターム集合は，$\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$ と定義される．

2.2 ラフ集合理論

2.2.1 同値関係ラフ集合モデル (ERSM) [4][6]

全体集合 U 上の $x \in U$ に対してある識別不能関係（同値関係）$\mathcal{R}$ が与えられたとき，x と関係 $\mathcal{R}$ である U 上の要素の集合を，

$$[x]_{\mathcal{R}} = \{y \in U \mid x\mathcal{R}y\} \tag{1}$$

と定義する ($x\mathcal{R}y$ は x と y との間に関係 $\mathcal{R}$ があることを意味する)．このとき，$X \subseteq U$ について，

$$\mathcal{L}(\mathcal{R}, X) = \{x \in U \mid [x]_{\mathcal{R}} \subseteq X\} \tag{2}$$

$$\mathcal{U}(\mathcal{R}, X) = \{x \in U \mid [x]_{\mathcal{R}} \cap X \neq \emptyset\} \tag{3}$$

と定めたとき，$\mathcal{L}(\mathcal{R}, X)$ を X の下近似，$\mathcal{U}(\mathcal{R}, X)$ を X の上近似といい，これらをまとめた $(\mathcal{L}(\mathcal{R}, X), \mathcal{U}(\mathcal{R}, X))$ を X のラフ集合という．

本研究では，自然言語を扱うため，タームの同値関係における推移律が成り立たず，ERSM を定義できない可能性がある．そこで，同値関係の反射律，対称律のみを示すトレランスラフ集合モデルを用いる．

2.2.2 トレランスラフ集合モデル (TRSM) [1][2]

x の同値関係 $[x]_{\mathcal{R}}$ に代わるトレランスクラスを $I(x)$ とする．また，関数 ν を $\nu(X, Y) = |X \cap Y|/|X|$ のように定める．トレランスラフ集合モデルは，

$$\mathcal{L}(\mathcal{R}, X) = \{x \in U \mid \nu(I(x), X) = 1\} \quad (4)$$

$$\mathcal{U}(\mathcal{R}, X) = \{x \in U \mid \nu(I(x), X) > 0\} \quad (5)$$

と定義し，これを集合 X のトレランスラフ集合と呼ぶ．

3 TRSM を用いた非階層的クラスタリング（従来手法）[1]

3.1 タームのトレランス空間

タームのトレランスクラスを用いて、ドキュメントを表現する．

- $f_{d_j}(t_i)$： ドキュメント d_j 中のターム t_i の出現回数
- $f_{\mathcal{D}}(t_i)$： ターム t_i の出現するドキュメント数
- $f_{\mathcal{D}}(t_i, t_j)$： ターム t_i と t_j が共起するドキュメント数

と定めたとき，タームのトレランスクラスは閾値 θ を用いて，

$$I_\theta(t_i) = \{t_j | 100 \times (f_{\mathcal{D}}(t_i, t_j)/M) \geq \theta\} \cup \{t_i\} \quad (6)$$

と定義する．

このときターム集合 $\mathcal{T}$ のトレランスラフ集合モデルは，

$$\mathcal{L}(\mathcal{R}, T) = \{t_i \in \mathcal{T} \mid \nu(I_\theta(t_i), T) = 1\} \quad (7)$$

$$\mathcal{U}(\mathcal{R}, T) = \{t_i \in \mathcal{T} \mid \nu(I_\theta(t_i), T) > 0\} \quad (8)$$

となる．以降のドキュメント d_j は上近似を用いて，

$$\mathcal{U}(\mathcal{R}, d_j) = \{t_i \in \mathcal{T} \mid \nu(I_\theta(t_i), d_j) > 0\} \quad (9)$$

と表現する．このときドキュメント d_j を構成するターム t_i の重み w_{ij} は，

$$w_{ij} = \begin{cases} (1 + \log(f_{d_i}(t_i))) \times \log \frac{M}{f_{\mathcal{D}}(t_i)}, \\ \qquad \text{if } t_i \in d_j; \\ \min_{t_h \in d_j} w_{hj} \times \frac{\log(M/f_{\mathcal{D}}(t_i))}{1+\log(M/f_{\mathcal{D}}(t_i))}, \\ \qquad \text{if } t_i \in \mathcal{U}(I_\theta(t_i), d_j) \backslash d_j; \\ 0, \qquad \text{if } t_i \notin \mathcal{U}(I_\theta(t_i), d_j). \end{cases} \quad (10)$$

とする．

3.2 TRSM 非階層的クラスタリングアルゴリズム

従来手法である k-means 法 [5] を利用したクラスタリングアルゴリズムを以下に示す

[TRSM 非階層的クラスタリングアルゴリズム]

ドキュメント集合 $\mathcal{D}$ とクラスタ数 K を入力する．

1) ドキュメント集合 $\mathcal{D}$ からランダムに K 個のドキュメントを選択し，クラスタ $C_1, C_2, \ldots, C_K$ の初期代表元を $R_1, R_2, \ldots, R_K$ とする．
2) 各ドキュメント $d_j \in \mathcal{D}$ に関して，上近似 $\mathcal{U}(\mathcal{R}, d_j)$ とクラスタ代表元 $R_k, (k = 1, 2, \ldots, K)$ の類似度 $S(\mathcal{U}(\mathcal{R}, d_j), R_k)$ を計算する．類似度が一定以上になった d_j を C_k にマージする．
3) 各クラスタ C_k の代表元 R_k を更新する．
4) 2), 3) を類似度が変化しなくなるまで繰り返す．
5) どのクラスタにも属していないドキュメントを，それと最も類似したドキュメントの所属クラスタに格納する． □

3.2.1 クラスタ代表元構築法

クラスタリングアルゴリズムのステップ 3) におけるクラスタ C_k の代表元 R_k の構築法を以下に示す．

[クラスタ代表元構築法]

i) $R_k = \phi$ とする．
ii) すべての $d_j \in C_k$ と，すべての $t_i \in d_j$ について，$f_{C_k}(t_i)/|C_k| > \sigma$ であれば，$R_k = R_k \cup t_i$ とする．
iii) もし，$d_j \in C_k, d_j \cap R_k = \phi$ であれば，$R_k = R_k \cup \arg\max_{t_i \in d_j} w_{ij}$ とする． □

また，クラスタ C_k の代表元のターム t_i の重みは

$w_{ik} = (\sum_{d_j \in c_k} w_{ij}) / |\{d_j : t_i \in d_j\}|$

として代表元の次元数で標準化する．

3.2.2 クラスタ代表元とドキュメントの類似度

ドキュメントとクラスタ代表元の類似度を以下のように定義する．

$$S(\mathcal{U}(\mathcal{R}, d_j), R_k) = \frac{2 \times \sum_{l=1}^{N} (w_{l\mathcal{U}} \times w_{lR_k})}{\sum_{l=1}^{N} w_{l\mathcal{U}} + \sum_{l=1}^{N} w_{lR_k}} \quad (11)$$

なお，$w_{l\mathcal{U}}$ はドキュメント d_j の上近似 $\mathcal{U}(\mathcal{R}, d_j)$ のターム t_l の重み，w_{lR_k} はクラスタ C_k の代表元 R_k におけるターム t_l の重みを表す．

3.3 従来手法の問題点

各タームとその出現ドキュメント数を図 1 に示す．

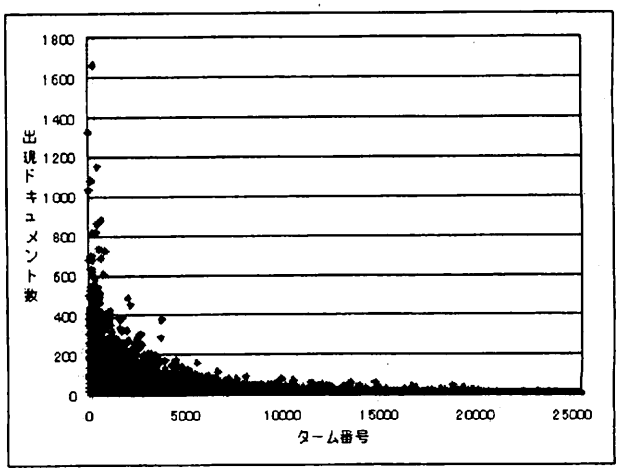

図 1: タームの出現ドキュメント数

従来手法はクラスタ内の出現頻度の高いタームを選択し，そのクラスタの代表元とする．そのため，各クラス

タとも図1に示される出現頻度の高いタームが主に抽出され，またその重みが非常に大きい代表元が生成される．再帰的にクラスタの代表元を構築していく過程の中で，各クラスタの代表元が非常に似通ったものになる．結果として各クラスタが全体集合の中心に集まる傾向があり，効果的なクラスタリングが行われない可能性がある．

4 提案手法

4.1 提案手法の方針

以上のような従来手法の問題に対処するため，クラスタの代表元の構築にドキュメントを要素とするクラスタの下近似を利用したクラスタリング手法を提案する．

提案手法ではクラスタ内の要素をタームの集合であるドキュメントと考え，そのクラスタの下近似をクラスタの代表元の抽出に用いる．ドキュメントのクラスタへマージする基準を $\mathcal{R}_C$ とすると，クラスタ C_k のラフ集合は，

$$\mathcal{L}(\mathcal{R}_C, C_k) = \{d_j \in \mathcal{D} \mid [d_j]_{R_C} \subseteq C_k\} \quad (12)$$

$$\mathcal{U}(\mathcal{R}_C, C_k) = \{d_j \in \mathcal{D} \mid [d_j]_{R_C} \cap C_k \neq \emptyset\} \quad (13)$$

と表せる．

クラスタの代表元の構築には以上のクラスタの要素をドキュメントとしたときの下近似 $\mathcal{L}(\mathcal{R}_C, C_k)$ を用いる．また，クラスタの下近似が存在しないときは最もそれに近いドキュメント集合，つまり所属クラスタが最も少ないドキュメントの集合を用いる．これによって，出現頻度の高いタームのみがクラスタの代表元として抽出される可能性を抑え，クラスタ C_k 内にマージされたドキュメントの中で C_k に特徴的なタームも同時に抽出することができると考えられる．

4.2 提案手法のアルゴリズム

クラスタリングアルゴリズムは従来手法と同様である．

$k = (1, 2, \ldots, K)$ におけるクラスタ C_k の代表元 R_k の構築法を以下に示す．

[提案手法のアルゴリズム]

i) $R_k = \phi$ とする．
ii) $\forall d_j \in \mathcal{L}(\mathcal{R}_C, C_k)$ となる $[d_j]_{R_C}$ を求める．
iii) $R_k = \{t_i | t_i \in d_j\}$ とする．
iv) $R_k = \emptyset$ であれば，$[d_j]_{R_C} \cap C_k \neq \emptyset$ を満たす C_k の数が最も小さい $[d_j]_{R_C}$ を抽出し，iii) へ行く．
v) もし，$d_j \in C_k, d_j \cap R_k = \emptyset$ であれば，$R_k = R_k \cup \arg\max_{t_i \in d_j} w_{ij}$ とする． □

なお，クラスタ C_k の代表元のターム t_i の重みは

$w_{ik} = (\sum_{d_j \in c_k} w_{ij}) / |\{d_j : t_i \in d_j\}|$

として代表元の次元数で標準化する．

5 提案手法の評価と考察

5.1 条件

テストデータに毎日新聞1994[7]を元にしたBMIR-J2テストコレクション[8]を用いてクラスタリングのシミュレーションを行った．また，評価対象20クエリに対して正解ドキュメントはBMIR-J2が提供するものとする．実験で用いるテストコレクションBMIR-J2の構造を以下に示す．

・ドキュメント数	5080
・全ターム数	25446
・ドキュメント当りのターム数 (平均)	64.9

5.2 評価指標

本研究ではクラスタリングの傾向と評価について，クラスタベース検索を用いる．クエリと類似度が最も高いクラスタを対象としてクエリに対して，

・a :対象クラスタ内の正解ドキュメント数
・b :対象クラスタ内の不正解ドキュメント数
・c :対象クラスタ外の正解ドキュメント数

を定める．また，

$$\text{F 値} = \frac{2 \times Pre \times Rec}{Pre + Rec} \quad (14)$$

を定義し，この値を評価尺度とする．ただし，適合率 Pre と再現率 Rec の値は，

$$Pre = \frac{a}{a+b} \quad (15)$$

$$Rec = \frac{a}{a+c} \quad (16)$$

とする．

5.3 結果

$\sigma = 0.035, K = 15$ として，$\theta = \{1, 2, 3, 4, 5\}$ について10回シミュレーションを行い，その平均値を求めた．($\theta > 5$ では変化は見られなかった．) その結果を表1に示す．なお $terms$ はクエリのターム数，$documents$ はクエリに対する正解ドキュメント数とする．また，クラスタ内のドキュメント数の例を以下の表2に示した．

全20クエリの中でほぼ全てのクエリで改善が見られた．また，従来手法はクラスタ中のドキュメント数が非常に大きく，クラスタ間のドキュメントの重複が頻繁に起きたため再現率は高かった．一方，適合率は，クラスタ内ドキュメント数を大幅に減少させた提案手法にメリットがあり，そのためF値が向上したと考える．

また，クエリごとのF値の大きさの違いは，クエリに対する正解ドキュメント数に大きく依存するといえる．

表 1: 実験結果 (F 値 $\times 10^{-2}$)

θ	手法	検索課題 (クエリ) 1	2	3	4	5	6	7	8	9	10
1	従来	0.486	3.22	2.86	1.29	5.16	1.87	2.89	3.64	23.8	0.442
	提案	0.911	3.83	14.8	4.11	11.7	4.28	7.29	7.23	36.3	2.65
2	従来	0.529	2.72	2.02	1.23	4.99	1.02	2.12	3.70	26.1	0.419
	提案	1.21	3.43	11.5	4.10	9.79	4.02	6.10	10.4	35.8	3.78
3	従来	0.626	3.93	3.78	2.68	5.11	2.07	2.65	3.46	26.7	0.421
	提案	1.26	4.26	12.0	4.15	14.2	3.92	6.11	8.27	31.6	2.16
4	従来	0.527	3.93	3.08	1.34	5.16	2.25	3.01	4.15	25.1	0.467
	提案	0.596	2.93	20.4	4.10	10.5	3.81	5.95	8.91	33.6	2.88
5	従来	0.235	2.26	2.64	1.29	8.32	1.05	3.04	4.17	21.9	0.432
	提案	0.994	3.72	16.6	4.64	12.3	4.77	7.45	7.63	35.5	2.61
terms		4	2	6	5	2	3	8	2	4	1
documents		5	37	34	26	110	22	47	61	564	9

θ	手法	検索課題 (クエリ) 11	12	13	14	15	16	17	18	19	20
1	従来	2.63	9.53	3.95	1.49	4.47	1.70	1.74	0.962	0.581	14.8
	提案	3.60	15.7	25.8	2.36	6.09	2.37	2.81	3.01	3.16	25.7
2	従来	2.73	9.23	4.62	2.47	4.54	1.36	1.44	1.22	0.922	15.2
	提案	3.28	13.9	19.9	3.28	10.5	2.39	2.12	2.68	3.57	25.9
3	従来	2.89	9.53	4.35	1.59	4.75	1.29	2.08	1.27	1.39	15.6
	提案	3.25	14.4	18.5	2.94	6.99	1.84	3.82	2.93	2.21	20.2
4	従来	2.90	9.68	4.05	1.74	4.66	2.66	2.26	1.38	0.889	15.2
	提案	3.61	11.6	33.4	2.47	9.48	1.62	3.22	4.32	3.17	22.1
5	従来	4.56	9.77	2.70	1.43	4.62	3.29	1.43	1.35	1.25	15.3
	提案	4.03	15.1	28.4	2.98	6.19	1.73	2.28	3.30	2.89	26.7
terms		1	2	1	2	3	7	3	6	4	2
documents		29	168	51	18	88	19	17	22	12	303

5.4 考察

代表元構築の際に用いる閾値を削減するクラスタの代表元の構築方法の提案により，F 値を向上させることができた．これはドキュメント内のタームにおいて，より特徴的なものをクラスタの代表元に組み込むことができたためと考える．

クラスタ内に含まれるドキュメント数が削減されたことで適合率が上昇したため F 値が高くなったと考えられる．また，クラスタをマージする類似度の閾値やクラスタの代表元を構築する際に用いる閾値 σ などはクラスタの粒度に大きく依存するため，予備実験において最も適した値に設定して実験を行った．従って，それらの値によってはアルゴリズムが機能しない場合が従来手法・提案手法共に存在する．

また，タームのトレランスクラスは全ドキュメント数における出現ドキュメント数を基準に構成しているため，必ずしもトレランスクラスが類似したタームから構成されていないとことが問題点に挙げられる．

6 まとめと今後の課題

今回の研究では，ドキュメントの特徴を考慮し，クラスタ代表元の構築方法を新たに提案した．その結果，非階層的ドキュメントクラスタリングをドキュメントの性質に適応させ，評価値を向上させることができた．

また，本手法以外にクラスタ内のタームの重みの分散や全体平均との差を用いること，類似しているタームのクラスを構築することで，クラスタの凝集具合の改善やドキュメント表現の効率化が可能であると考えられ，評価値の更なる向上の可能性があるといえる．それらの検討と本手法を検索や分類へ適用することが課題として挙げられる．

表 2: クラスタ格納ドキュメント数 (例)

手法	格納ドキュメント数 平均	最大値	最小値
従来手法	3568	4845	1595
提案手法	740	1604	118

7 謝辞

著者の一人である渡辺は，本研究を行うにあたり，数多くのご助言，ご支援を賜りました早稲田大学平澤研究室の各氏に感謝いたします．

参考文献

[1] T. B. Ho, N. B. Nguyen: "Nonhierarchical Document Clustering Based on Tolerance Rough Set Model," *International Journal of Intelligent Systems*, Vol.17, Issue 2, pp.199-212, (2002).

[2] S. Kawasaki, N. B. Nguyen, and T. B. Ho: "Hierarchical Document Clustering Based on Tolerance Rough Set Model," *Lecture Notes In Artificial Intelligence*,Vol.1910, pp. 67-82, (2000).

[3] D. S.Modha, W. S. Spangler: "Feature Weighting in k-Means Clustering," *Machine Learning*, Vol.52 , pp.217-237, (2003).

[4] Pawlak.Z., "Rough sets: Theoretical aspects of reasoning about data," Kluwer Academic Publishers, (1991).

[5] 宮本定明, "クラスター分析入門," 森北出版, (1999).

[6] H. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, (2000).

[7] 毎日新聞社, "CD-毎日新聞'94 データ集," 日外アソシエーツ.

[8] 情報処理学会データベースシステム研究会, BMIR-J2 テストコレクション, 新情報処理開発機構 (1998).