情 報 学 基 礎　66－18
デジタル・ドキュメント　32－18
（２００２．３．１５）

特徴要素の重みを考慮に入れたクラスタ代表の洗練による文書クラスタリング

小林 優　吉高 淳夫　平川 正人
広島大学大学院工学研究科

〒 739-8527　東広島市鏡山１丁目４番１号
{koba, yoshi, hirakawa}@isl.hiroshima-u.ac.jp

本論文では、利用者の分類に対する意図をクラスタリング結果に反映させ、かつ、文書クラスタリングを効率よく行うために、適切なクラスタ代表を求める手法を提案する。本手法では、利用者の分類例示に基づいて形成されるクラスタの主題を表す特徴要素を多く含む文書をサンプルとして選出し、そのサンプルをクラスタに追加することで、クラスタ代表の算出を行う。ここで、追加されたサンプルによっては、適切なクラスタ代表を求めることができない場合があるため、サンプルの追加と除去、クラスタ代表の算出を繰り返し行うことによって、クラスタ代表を洗練する。

Refining Cluster Centroids based on Feature Significance for Efficient Document Clustering

Yu KOBAYASHI, Atsuo YOSHITAKA, Masahito HIRAKAWA
Graduate School of Engineering, Hiroshima University

4-1, Kagamiyama 1 chome, Higashi-Hiroshima, 739-8527
{koba, yoshi, hirakawa}@isl.hiroshima-u.ac.jp

In this paper, we propose a method of refining cluster centroids to reflect user's intention to a clustering result and perform document clustering efficiently. In the proposed method, a cluster is formed based on a classification example, and a document which contains many features showing a subject of the cluster is selected as a sample and added to the cluster. Since a sample may be inappropriate for the cluster, addition and removal of a sample are iterated in order to refine the cluster centroid.

1. はじめに

インターネットの普及により、膨大な数の文書データにアクセスできるようになり、情報検索システムは重要な役割を担っている。しかしながら、従来の情報検索システムにおける問題点として、利用者に適切なキーワードの記述力を要求すること、大量の検索結果から必要な情報を探し出すことは利用者にとって大きな負担となることが挙げられる。そのため、大量の情報の中から必要な情報を効率よく探し出すための手法が必要となる。一手法として、対象をクラスタリングすることが挙げられる。

クラスタリングとは、類似しているデータを同じグループに集め、全対象を分類することである。クラスタリングを行うことによって、探索範囲を限定し、効率よく必要な情報を獲得することができる。しかし、利用者の意図するようなクラスタリング結果が得られなければ、探索範囲を限定す

ることができないため、効率的な探索が可能になるとはいえない。そのため、利用者の分類に対する意図を考慮し、対象となる文書をクラスタリングする必要がある。

ここで、ベクトル空間に基づいた文書クラスタリングについて考える。クラスタリング対象となる文書は、特徴要素(文書中に出現する単語)に対する重み(単語の出現回数)を並べた特徴ベクトルとして表現され、そのベクトルの基底となる特徴要素の選び方によって対象間の類似度は変化する。そのため、利用者の意図をクラスタリング結果に反映させるためには、適切な特徴要素の集合を選出することが重要になる。

この視点に立ち、[7]では、システムから与えられたサンプル文書の分類例示に基づいて特徴要素を選出している。そして、選出された特徴要素を用いてクラスタリングを行い、利用者の分類意図をクラスタリング結果に反映させている。しかし、クラスタリング処理において、最も類似している二つの文書(クラスタ)を結合する処理を繰り返しているため、計算量が多いという問題がある。利用者が所望の文書を即時に必要とする状況を想定する場合、クラスタリング処理に多大な時間を割くことができない[1]-[3]。そのため、クラスタリング処理を効率的に行う必要がある。

効率的なクラスタリング処理を行うための一手法として、形成されるクラスタと同数のクラスタ代表を求めることが考えられる。クラスタ代表とは、同じクラスタ内の文書データの重み付き平均ベクトルである。クラスタ代表を求めると、クラスタ代表と文書間の類似度を算出することによってクラスタリング処理を行うことができるので、計算量の削減が可能となる。しかし、クラスタリング精度と処理速度との間にはトレードオフの関係があるため、精度を保ちながら効率的なクラスタリング処理を行うための手法が必要となる。

[1]では、クラスタ代表を求めるための手法を二つ提案している。まず、Buckshot 法では、クラスタリング対象となる文書集合をランダムにサンプリングして文書数を減らし、この部分集合に対して、群平均法のクラスタリングアルゴリズムを用いることでクラスタ代表を求めている。次に、Fractionation 法では、文書集合をグループに分割し、群平均法を用いて Buckshot 法より精度の高いクラスタ代表を求めている。[2]では、K-平均法を拡張している。まず、K 個のクラスタ代表をランダムに決め、クラスタリング対象となる文書データがクラスタに割り当てられるたびに、クラスタ代表の算出を繰り返し行うことで、クラスタ代表を求めている。これらの手法では、求められたクラスタ代表が適切であるかどうかは評価されていない。[4]では、対象データの部分集合をランダムに複数回サンプリングし、それぞれのサンプルに対して、K-平均法を用いてクラスタ代表を求めている。しかし、ランダムなサンプルによって求められたクラスタ代表の妥当性が問題となる。

本研究では、利用者の分類例示に基づいて選出された特徴要素を用いることで、利用者の分類意図をクラスタリング結果に反映させ、かつ、効率的なクラスタリング処理を行うための手法を提案する。提案手法では、適切なクラスタ代表を求めるために、分類例に基づいて形成されたクラスタの主題を表す特徴要素を多く含む文書をサンプルとして用いることで、クラスタ代表の算出を行う。クラスタ代表は、特徴要素に対する重みを並べた特徴ベクトルとして表現されるため、サンプルの追加と除去、クラスタ代表の算出を繰り返し行い、特徴要素に対して適切な重みを与えることによってクラスタ代表を洗練する。

2. 処理概要

クラスタリング処理の流れを図 1 に示す。ここで、利用者の分類例示に基づいて形成されるクラスタを例示クラスタと呼ぶことにする。

まず、クラスタリング対象となる文書集合から、ある部分集合をサンプル文書として取り出し、利用者に提示して、主題が似ていると判断した文書が同じグループに入るように分類させる。そして、分類例に基づき、[7]で提案された手法を用いて、同じグループに属する文書間の類似度を高く、異なるグループに属する文書間の類似度を低くするような特徴要素の集合を選出する。次に、選出された特徴要素によって形成される多次元空間内に、クラスタリング対象となる文書を配置する。

初期状態においては、分類例示したグループと同じ数の例示クラスタ、クラスタリング処理が行われていない文書が存在する。まず、例示クラスタに含まれる文書の特徴ベクトルの平均ベクトル(例示クラスタの重心ベクトル)を、初期のクラスタ代表とする。そして、クラスタの主題を表す特徴要素を多く含む文書をサンプルとして選出し、例示クラスタに追加することで、再びクラスタ代表の算出を行う。算出されたクラスタ代表が適切であるかどうかを判定するために、クラスタ代表とそのクラスタに属する文書との平均類似度を算出する。平均類似度がサンプルを追加する前の値よりも低くなる場合は、そのサンプルを除き、クラスタ代表を再計算する前の状態に戻す。このように、サンプルの追加と除去、クラスタ代表の算出を繰り返すことで、クラスタ代表を洗練する。

最後に、洗練されたクラスタ代表とクラスタリング対象となる文書間の類似度を算出し、最も類似度が高いクラスタに文書を割り当て、全ての文書がいずれかのクラスタに割り当てられたら処理を終了する。

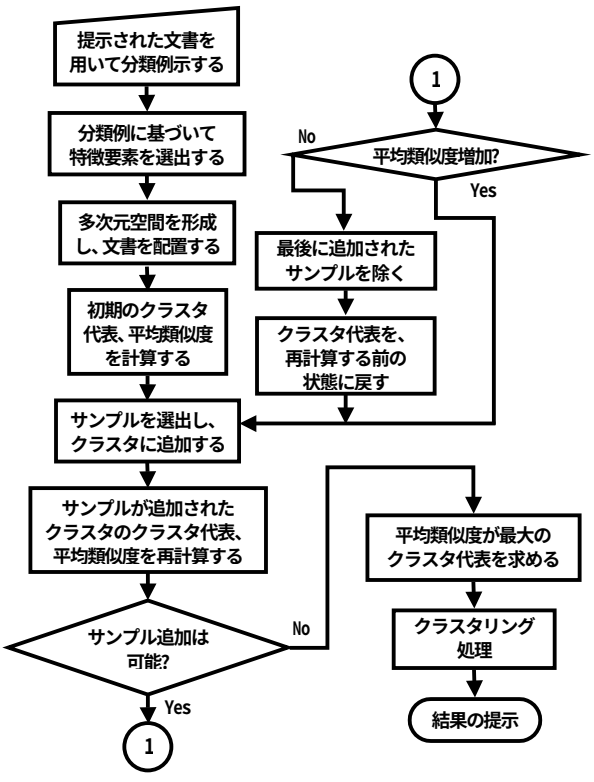

図 1: クラスタリング処理の流れ

3. 特徴要素の選出

3.1 分離度の定義

クラスタリング対象となる文書 $d_i(i=1,\ldots,n)$の特徴要素 $f_k(k=1,\ldots,m)$に対する重み(単語の出現回数)を $w_i(f_k)$とし、文書 d_i の特徴ベクトル $\mathbf{d}_i$ を以下のように表す。

$$\mathbf{d}_i = (w_i(f_1), w_i(f_2), \ldots, w_i(f_m)) \quad (1)$$

文書間の類似度は、余弦尺度を用いて算出する。特徴要素の集合 $F_x=\{f_1,\ldots,f_m\}$を用いて形成される多次元空間内において、文書 d_i と d_j 間の類似度 $Sim^{F_x}(d_i, d_j)$を以下のように算出する。

$$Sim^{F_x}(d_i, d_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{\|\mathbf{d}_i\|\|\mathbf{d}_j\|} \quad (2)$$

この式は、大きさ 1 に正規化した 2 つの特徴ベクトル $\mathbf{d}_i$ と $\mathbf{d}_j$ の内積であり、2 つのベクトルが重なり合うときに類似度は 1、直交するときに類似度は 0 となる。

特徴要素の選出基準として、例示クラスタの分離度 Sep^{F_x} を以下のように定義する。

$$Sep^{F_x} = \sum_{d_i \in G_a} \sum_{\substack{d_j \in G_a \\ d_i \neq d_j}} Sim^{F_x}(d_i, d_j) + \sum_{d_i \in G_a} \sum_{d_j \notin G_a} \{1 - Sim^{F_x}(d_i, d_j)\} \quad (3)$$

この式は、分類例示に用いた文書集合のうち、同じクラスタ G_a に属する文書 d_i と d_j の類似度が高く、異なるクラスタに属する文書 d_i と d_j の類似度が低くなれば、例示クラスタの分離度が高くなることを表している。

3.2 特徴要素の選出手法

特徴要素の選出順序を決めるために、各特徴要素にスコアを与える。多次元空間を形成する特徴要素の集合 F_x に、ある特徴要素 f_k を加えたときの分離度の増加量を特徴要素 f_k に与えるスコアとする。ただし、特徴要素 f_k に与えるスコアは、既に選出されている特徴要素の集合 F_x' に依存するため、以下のような手順で、各特徴要素に与えるスコアの算出と選出順序の並べ替えを繰り返し行う。

1) 各特徴要素に与える初期スコア $V(f_k)$を計算する。

$$V(f_k) = Sep^{f_k} \qquad (f_k \in F_x) \quad (4)$$

2) 初期スコアの高い順に特徴要素を並べる。

3) 各特徴要素に与えるスコア$V(f_k)$を再計算する。
$$V(f_k) = Sep^{F_x' \cup f_k} - Sep^{F_x'} \quad (F_x' \subset F_x) \qquad (5)$$
ここで、F_x'は再計算する前の段階で与えられたスコアが、特徴要素f_kのスコアよりも高い特徴要素の集合である。

4) 再計算によって与えられたスコアの高い順に特徴要素を並べ替える。再計算する前の段階と比べて、特徴要素の選出順序が変わらない場合は処理を終了する。選出順序が変わる場合は3)へ戻る。

最終的に与えられたスコアの高い順に特徴要素を選出する。多次元空間を形成する特徴要素の集合F_xに、ある特徴要素f_kを加えたときの分離度の増加量を特徴要素f_kに与えるスコアとして求めたので、最終的に正のスコアが与えられた特徴要素を全て選出したときに、分離度が最大となる。

4. クラスタ代表の洗練

4.1 サンプルの追加

クラスタリング処理を行うために、例示クラスタに含まれる文書の特徴ベクトルの平均ベクトルをクラスタ代表とする方法が考えられるが、それが適切なクラスタ代表であるとは限らない。そこで、文書を分類するにあたって、多くの文書に出現する特徴要素は重要であるという考えに基づき[5]、このような特徴要素を多く含む文書をサンプルとして選出し、クラスタに追加する。利用者の分類例示に基づいて選出された特徴要素の集合をF_b、クラスタリング対象となる文書d_iに出現する特徴要素f_kに対する重みを$w_i(f_k)$、全文書の中で特徴要素f_kが出現する文書の数を$DF(f_k)$として、文書d_iの重要度$Eff(d_i)$を以下のように表す。

$$Eff(d_i) = \frac{\sum_{f_k \in F_b} w_i(f_k) DF(f_k)}{\sqrt{\sum_{f_k \in F_b} \{w_i(f_k)\}^2}} \qquad (6)$$

各文書について$Eff(d_i)$の値を求めることによって、文書d_iの重要度を推定することができる。あるクラスタに追加される重要度$Eff(d_i)$の値が高い文書d_iは、そのクラスタの主題を表す特徴要素を多く含むと考えられる。したがって、$Eff(d_i)$の値が高い順にサンプルとして用いることによって、より適切なクラスタ代表を求めることができる。

4.2 Rocchio の式のクラスタリングへの適用

サンプルが追加されたクラスタのクラスタ代表を再び算出するため、本研究では、Rocchio の式について考える。Rocchio の式は、情報検索において、利用者が入力した検索語を修正することによって、検索精度を高めるために用いられてきた。Rocchio の式を以下のように表す。

$$\mathbf{q}^{new} = \frac{\mathbf{q}^{old}}{\left\|\mathbf{q}^{old}\right\|} + \frac{\alpha}{|R|} \sum_{d_r \in R} \frac{\mathbf{d}_r}{\left\|\mathbf{d}_r\right\|} - \frac{\beta}{|D-R|} \sum_{d_s \in D-R} \frac{\mathbf{d}_s}{\left\|\mathbf{d}_s\right\|} \qquad (7)$$

$$\mathbf{q} = (w(f_1), \ldots, w(f_l)) \qquad (8)$$

$$\mathbf{d}_r = (w_r(f_1), \ldots, w_r(f_l)) \qquad (9)$$

$$\mathbf{d}_s = (w_s(f_1), \ldots, w_s(f_l)) \qquad (10)$$

ここで、ベクトル$\mathbf{q}^{old}, \mathbf{q}^{new}$はそれぞれ、最初に与えた検索語、修正された検索語をベクトル表現したものである。また、集合$R, D\text{-}R$はそれぞれ、検索対象となる文書集合Dに含まれる適合(利用者が必要であると判断した)文書、非適合(利用者が必要でないと判断した)文書の集合である。α, βはそれぞれ、適合文書に出現する特徴要素に対する重みを大きく、非適合文書に出現する特徴要素に対する重みを小さくするためのパラメータであり、ともに正の値をとる。ベクトル表現された適合文書、非適合文書を(7)式に代入することで、適合文書間の類似度を高く、非適合文書間の類似度を低くするようなベクトル$\mathbf{q}^{new}$が新たに算出される。

本研究では、利用者の意図をクラスタリング結果に反映させながら、クラスタリング処理を効率よく行うことを目的としている。したがって、本研究で想定するようなクラスタリングに Rocchio の式を適用すると、分類例示に用いられた文書、追加されたサンプルより、利用者の意図をクラスタリング結果に反映できるようなクラスタ代表を求めることができると考えられる。そこで、Rocchio の式を以下のように変形する。

$$\mathbf{c}(G_a^{new}) = \frac{\mathbf{c}(G_a^{old})}{\left\|\mathbf{c}(G_a^{old})\right\|} + \frac{\alpha}{|G_a|} \sum_{d_t \in G_a} \frac{\mathbf{d}_t}{\left\|\mathbf{d}_t\right\|} \qquad (11)$$

$$\mathbf{c}(G_a) = (w_a(f_1), \ldots, w_a(f_l)) \qquad (12)$$

ベクトル$\mathbf{c}(G_a^{old})$は最初に算出されたクラスタG_a

のクラスタ代表、ベクトル $\mathbf{c}(G_a^{new})$はサンプルが追加された後に算出されたクラスタ代表である。

(11)式では、負の項を除いている。なぜなら、[7]で提案された特徴要素の選出手法を利用する場合、クラスタに出現する特徴要素の集合は、クラスタ毎に異なるためである。

4.3 特徴要素に対する重みの調整

Rocchio の式では、パラメータαが固定値であるため、全ての特徴要素を同等に扱うという問題がある。そこで、特定のクラスタに出現する特徴要素はクラスタの主題を表し、重要であるため、重みを大きくするべきだという考えに基づき[6]、特徴要素に対する重みを変化させる。そのため、特徴要素f_kが特定のクラスタに依存する度合を算出する必要がある。まず、クラスタの代表ベクトル$\mathbf{c}(G_a)$ (a=1,…,M)の基底となる、k番目の特徴要素f_kに対する重み$w_a(f_k)$から導かれるベクトルを$\mathbf{T}_k$、ベクトル $\mathbf{T}_k$ の 1-ノルムで正規化したベクトルを$\mathbf{T}_k$'とし、それぞれ以下のように表す。

$$\mathbf{T}_k = (w_1(f_k),\ldots,w_M(f_k)) \tag{13}$$

$$\mathbf{T}_k' = \frac{\mathbf{T}_k}{\|\mathbf{T}_k\|_1} = (w_1(f_k)',\ldots,w_M(f_k)') \tag{14}$$

$$\|\mathbf{T}_k\|_1 = \sum_{a=1}^{M} |w_a(f_k)| \tag{15}$$

さらに、特徴要素f_kが特定のクラスタに依存する度合P_kを以下のようにして算出する。

$$P_k = \sum_{a=1}^{M} \{w_a(f_k)'\}^2 \tag{16}$$

P_kの値は、全てのクラスタに同じ特徴要素f_kが出現するときに最小値1/M、特定のクラスタのみに特徴要素f_kが出現するときに最大値1となる。(16)式は、複数のクラスタに出現する特徴要素に対する重みを小さくすることを目的としているが、一度の洗練処理で、適切なクラスタ代表が求められるとは限らない。そのため、サンプルの追加や除去を繰り返す過程で、特徴要素に対して適切な重みを与える必要がある。そこで、P_k を $P_k^{\lambda/\delta}$(δ>1)と修正する。ここで、λはサンプルを追加する回数である。λの値は、サンプルを追加するたびに1ずつ増えるので、複数のクラスタに出現する特徴要素に対する重みが段階的に小さくなる。従って、特定のクラスタに出現する特徴要素に対する重みを大きく、複数のクラスタに出現する特徴要素に対する重みを小さくすることが可能となる。これらを踏まえて、分類例示に用いられた文書、クラスタ代表を洗練するために用いられる文書d_tの特徴ベクトル$\mathbf{d}_t$を$\mathbf{d}_h$と変形し、以下のように表す。

$$\mathbf{d}_h = (P_1^{\lambda/\delta} w_h(f_1), P_2^{\lambda/\delta} w_h(f_2),\ldots,P_l^{\lambda/\delta} w_h(f_l)) \tag{17}$$

ベクトル$\mathbf{d}_h$を用いることで、特徴要素に対する重みを変化させることができる。以上より、Rocchio の式を以下のように変形する。

$$\mathbf{c}(G_a^{new}) = \frac{\mathbf{c}(G_a^{old})}{\|\mathbf{c}(G_a^{old})\|} + \frac{\alpha}{|G_a|} \sum_{d_h \in G_a} \frac{\mathbf{d}_h}{\|\mathbf{d}_h\|} \tag{18}$$

4.4 クラスタ代表の洗練

クラスタ代表が洗練されたかどうかを判定するための基準として、クラスタ代表と同じクラスタに属する文書間の平均類似度を用いる。これは、同じクラスタに属する文書間の類似度を高くするという、分離度の定義を考慮に入れた基準である。算出されたクラスタ代表を $c(G_a)$、クラスタ代表を洗練するために用いられた文書をd_hとして、平均類似度$Avg^{F_x}(c(G_a), d_h)$を以下のように表す。

$$Avg^{F_x}(c(G_a), d_h) = \frac{1}{|G_a|} \sum_{d_h \in G_a} \frac{\mathbf{c}(G_a) \cdot \mathbf{d}_h}{\|\mathbf{c}(G_a)\| \|\mathbf{d}_h\|} \tag{19}$$

平均類似度が最大となるように、以下の手順でクラスタ代表を洗練する。

1) 例示クラスタの重心ベクトルを算出し、ベクトル$\mathbf{c}(G_a^{old})$とする。
2) クラスタの平均類似度$Avg^{F_x}(c(G_a^{old}), d_h)$を算出する。ここでは、例示クラスタの重心ベクトルを算出しているので、$\mathbf{d}_h$=$\mathbf{d}_i$である。
3) ベクトル $\mathbf{c}(G_a^{old})$より、$P_k^{\lambda/\delta}$の値を算出する。ここでは、λの値は0である。
4) 文書d_iの重要度$Eff(d_i)$の値が最も高い文書をサンプルとしてクラスタに追加する。
5) (18)式を用いて、新しいクラスタ代表ベクトル$\mathbf{c}(G_a^{new})$を求める。
6) クラスタの平均類似度 $Avg^{F_x}(c(G_a^{new}), d_h)$を算出する。
7) いずれかの条件を満たすクラスタ代表を求め、

次に重要度 $Eff(d_i)$の値が高い文書をサンプルとして追加する。

7-1) $Avg^{F_x}(c(G_a^{new}), d_h) > Avg^{F_x}(c(G_a^{old}), d_h)$であれば、$\mathbf{c}(G_a^{new})$をクラスタ代表ベクトルとし、再びサンプルを追加する。クラスタ代表が新しく求められたため、λの値を 1 増やし、$P_k^{\lambda/\delta}$の値を再び算出する。

7-2) $Avg^{F_x}(c(G_a^{new}), d_h) \leq Avg^{F_x}(c(G_a^{old}), d_h)$であれば、$\mathbf{c}(G_a^{old})$をクラスタ代表ベクトルとし、最後に追加されたサンプルを除く。

8) 5)から 7)を繰り返し、平均類似度が最大となるクラスタ代表ベクトル $\mathbf{c}(G_a)$を求める。

一度除かれたサンプルは、クラスタ代表と文書(サンプル)間の平均類似度を高める効果を持たないので、サンプルとして用いない。

4.5 クラスタリング処理

洗練されたクラスタ代表とクラスタリング対象となる文書間の類似度を以下の式で算出し、クラスタリング処理を行う。

$$Sim^{F_x}(c(G_a), d_j) = \frac{\mathbf{c}(G_a) \cdot \mathbf{d}_j}{\|\mathbf{c}(G_a)\| \|\mathbf{d}_j\|} \qquad (20)$$

クラスタリング対象となる文書 d_j を最も類似度が高いクラスタ G_a に割り当て、全ての文書がいずれかのクラスタに割り当てられたら、クラスタリング処理を終了する。

5. 評価実験

5.1 実験に用いた文書集合

コンピュータ関連のニュースを扱ったサイト[9]に掲載された 822 の記事、電子情報通信学会のサイト[10]で公開されている 836 の論文の概要文を用いて、2 種類のデータベースを作成した。ニュース記事については、HTML 文書からタグを取り除き、残った文書を対象に形態素解析を行い、名詞と未知語と判定された単語を特徴要素の候補として抽出した。論文については、タイトル、あらまし、キーワード部分の文書を取り出し、その文書に対して形態素解析を行い、名詞と未知語と判定された単語を特徴要素の候補として抽出した。ただし、名詞の中で“ 接尾辞 ”、“ 接尾 ”、“ 非自立 ”、“ 代名詞 ”、“ 接頭辞 ”、“ 数 ”、“ 形容動詞語幹 ”、“ 副詞可能 ” と判定されたものは、特徴要素の候補から除いた。形態素解析には『茶筌』[8]を用いた。

キーワードの出現回数を重みとして、重みの高い順に 15 個のキーワードを各文書の特徴要素とした結果、それぞれのデータベースにおけるベクトル空間の次元数は 2544、2542 となった。

5.2 評価方法

(1) クラスタ境界の明確さ

クラスタ代表を洗練してクラスタリングを行う場合、洗練しないでクラスタリングを行う場合について、クラスタ境界の明確さを比較するため、以下の式を用いて評価を行った。

$$H = \frac{\sum_{a=1}^{M} \sum_{d_i \in G_a} Sim^{F_x}(c(G_a), d_i)}{\sum_{a=1}^{M} |G_a| Sim^{F_x}(c(G_a), c(D))}, \quad \mathbf{c}(D) = \frac{1}{|D|} \sum_{d_i \in D} \mathbf{d}_i$$

この式は、同じクラスタ内の文書群をできるだけ近づけ、クラスタに含まれる文書群を全対象からできるだけ分離することを目的としている。すなわち、H の値が大きいほど、クラスタ境界が明確であると言える。そこで、クラスタ境界の明確さを、(洗練する場合の H の値) / (洗練しない場合の H の値)を計算し、洗練する場合としない場合との比率として求めた。

(2) 他の手法との比較評価

データベース作成に用いた文書群をあらかじめ表 1、2 のように分類しておき、その分類に従って各文書が正しいクラスタに配置されているかを判断し、以下の式を用いて評価を行った。

$$Precision = \frac{\text{正しいクラスタに配置された文書数}}{\text{全文書数}}$$

本研究で提案したクラスタ代表の洗練手法と他の手法との比較実験を行った。さらに、クラスタリング処理にかかった時間についても比較を行った(Pentium IV, 1.4GHz 使用時)。

比較対象として、[1]で提案された Fractionation 法(F 法)および Buckshot 法(B 法)、[2]で提案された K-平均法を拡張した手法(K 法)、さらに、例示クラスタの重心ベクトルをクラスタ代表とした場合(C 法)を取り上げ、用意した 2 つのデータベース

に対して評価を行った。

あらかじめ分類した各グループから無作為に 2 つ(4 つ)の文書を取り出し、それを分類例示とした。それぞれの分類例示に基づいて特徴要素を選出し、特徴要素の選出個数を 10, 20, 30, …, 70 と変化させた場合に対してクラスタ代表の洗練を行い、クラスタ境界の明確さ、Precision と処理時間を求めた。ただし、求めた値はいずれも、10 回の分類例示より得た値の平均である。

表 1　分類結果(ニュース記事)

トピック	文書数
ノート型パソコン	106
マザーボード	69
メモリ	70
デジタルカメラ	100
携帯端末	65
業界動向	90
アプリケーションソフト	68
入力デバイス	90
インターネット接続	95
セキュリティ	67
合計	822

表 2　分類結果(論文の概要文)

トピック	文書数
音声	78
通信網	81
バイオサイバネティックス	157
アンテナ	131
非線形問題	82
ディジタル信号処理	95
電磁環境	67
画像処理	145
合計	836

5.3　実験結果

(1)　クラスタ境界の明確さ

実験結果を図 2～3 に示す。クラスタ代表を洗練すると、クラスタ代表を洗練しない場合と比べてクラスタの境界がより明確になっている。このことは、分類が明確になることを意味するので、例示クラスタの分離度の定義に従ってクラスタリング処理を行うことができていると言える。

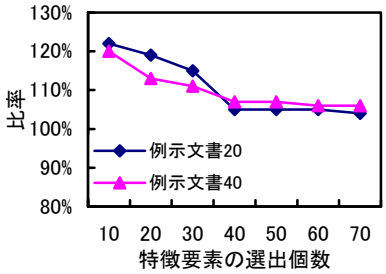

図 2: クラスタ境界の明確さ(ニュース記事)

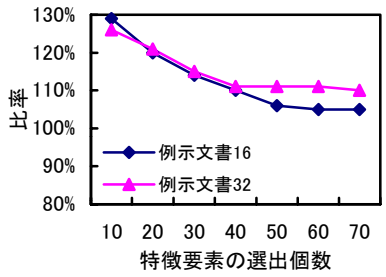

図 3: クラスタ境界の明確さ(論文の概要文)

(2)　他の手法との比較評価

実験結果を図 4～7 に示す。比較対象に取り上げた C 法は、他の手法と比べて悪い Precision を示している。これは、クラスタ代表を洗練することが、クラスタリングの精度を保つために有効であることを意味する。F 法、B 法、K 法との比較では、特徴要素の選出個数が少ない部分において、本手法が良い結果を残している。このことから、処理時間も考慮に入れると、本手法は精度を保ちながら効率的なクラスタリング処理を行うために有効であると言える。

6.　まとめ

本研究では、利用者の分類に対する意図をクラスタリング結果に反映させ、かつ、効率的な文書クラスタリングを行うために、適切なクラスタ代表を求める手法を提案した。評価実験により、①クラスタ代表を洗練すると、よりクラスタの境界が明確になる、②特徴要素の少ない範囲において高い分類精度が得られることから、処理時間を削減すると同時に、精度を保つことができることを示した。適切なサンプルが見つからない場合に対処できるようなクラスタ代表の洗練手法を確立することが今後の課題として挙げられる。

参考文献

[1] D.R. Cutting, D.R. Karger, J.O. Pedersen and J.W. Tukey, “Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections,” In Proceedings of SIGIR ’92, pp. 318-329, 1992.

[2] B. Larsen and C. Aone, “Fast and Effective Text Mining using Linear-time Document Clustering,” In Proceedings of KDD-99, pp. 16-22, 1999.

[3] H. Schütze and C. Silverstein, “Projections for Efficient Document Clustering,” In Proceedings of SIGIR ’97, pp. 74-81, 1997.

[4] P.S. Bradley and U.M. Fayyad, “Refining Initial Points for K-Means Clustering,” Proceedings of ICML ’98, pp. 91-99, 1998.

[5] Y. Yang and J.O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” Proceedings of ICML ’97, pp. 412-420, 1997.

[6] S. Shankar and G. Karypis, “A Feature Weight Adjustment Algorithm for Document Categorization,” In KDD-Workshop on Text Mining, 2000.

[7] 則武淳, 吉高淳夫, 平川正人, “利用者の分類例示に基づいて選出された特徴要素を用いた文書クラスタリング,” 情報処理学会自然言語処理研究会(2001-NL-142), pp. 205-212, 2001.

[8] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸, “日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書,” 2000.

[9] http://ascii24.com/

[10] http://search.ieice.or.jp/

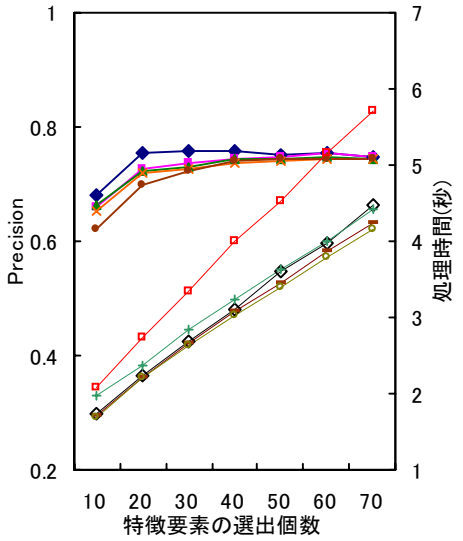

図 4: 他の手法との比較
(ニュース記事、例示文書 20)

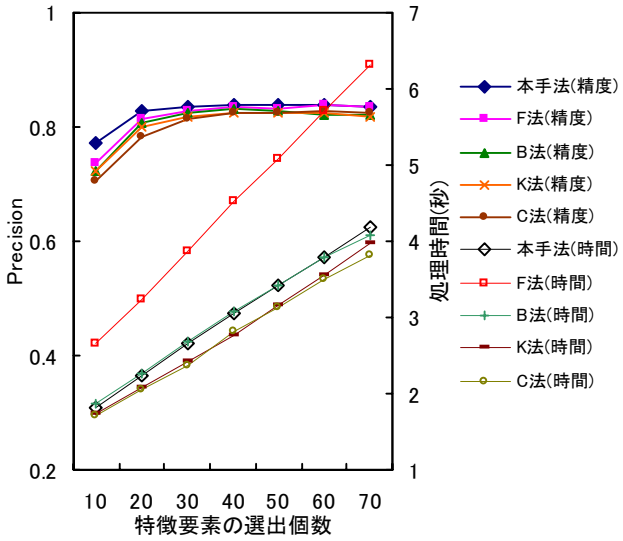

図 5: 他の手法との比較
(ニュース記事、例示文書 40)

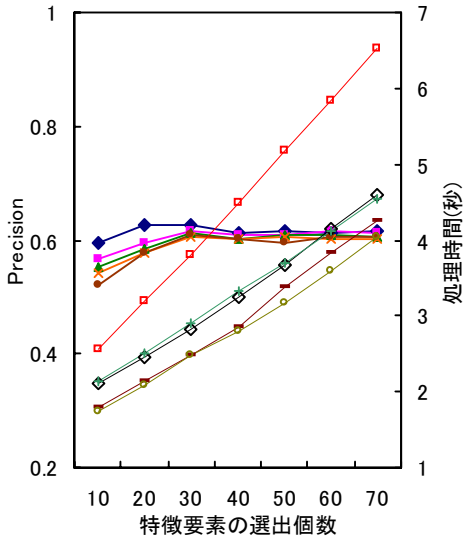

図 6: 他の手法との比較
(論文の概要文、例示文書 16)

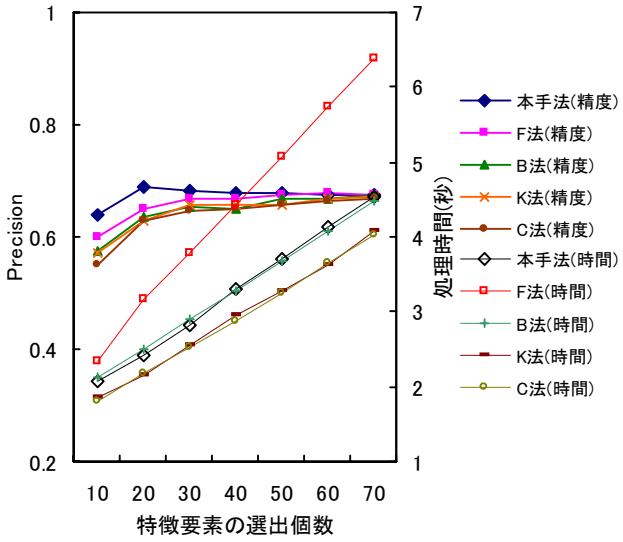

図 7: 他の手法との比較
(論文の概要文、例示文書 32)