情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2009-DBS-148 No.7
Vol.2009-FI-95 No.7
2009/7/28

半教師有りクラスタリングを適用した対話型文書分類技術の提案

佐藤祐介† 岩山 真†

ユーザがデータに対してもつ背景知識を利用して精度を向上させる半教師有りクラスタリングが注目されている．本研究では，ユーザによる文書への正解付与と半教師有りクラスタリングを交互に繰り返すことで精度を向上させる対話型文書分類方式において，段階的な正解付与が精度向上にどのような影響を及ぼすのかを評価した．クラスタリング結果の正／誤に関わらず未付与の文書に正解を与えた場合の精度向上効果を検証した．また，対話型文書分類では，ある時点で正解付与件数ぶんしかクラスタリング精度が向上しなくなる状態に達する．この状態は，クラスタが安定状態に達している可能性が高いため，クラスタリングしてもほぼ同じクラスタに収束することになり，クラスタリングによる精度向上は見込めない．本研究では，この状態を検知するための方式についても報告する．

Interactive Document Classification using Semi-supervised Clustering

Yusuke Sato† and Makoto Iwayama†

We discuss an interactive document classification method based on semi-supervised clustering which allows a user to partition a data set into their intended clusters. This method aims to progressively improve accuracy of clustering to repeat both of assigning appropriate clusters to documents and applying semi-supervised clustering. Our goal is to evaluate how well the method accelerates accuracy improvement. Moreover, by repeating both assigning and clustering, it comes to the point at which clustering accuracy improves just only the number of given true documents. We propose an approach to predict such a point based on amount of cluster label changes at K-Means loop.

†（株）日立製作所 中央研究所
Hitachi, Ltd., Central Research Laboratory

1. はじめに

近年，大規模な電子データを高精度に分類する必要性が急速に増している．Web ページや電子メール，学術文献といったテキストデータに加えて，画像，音声，動画などさまざまな大容量データをユーザの意図に合わせて分類する技術が求められている．

こうした大容量データの分類では，クラスタリングのような教師無しのアプローチが広く使われている．K-Means 法[1]に代表されるクラスタリングは，データがもつ特徴量間の類似性を基準として集合を任意のグループに分割する．しかし，データの内的な基準に従って分類するため，ユーザの意図を反映した分類が行えないという問題があった．この問題に対して，半教師有りクラスタリングの一種である制約付きクラスタリングというアプローチが注目を集めている．

制約付きクラスタリングとは，データに対してもっているユーザの背景知識を利用することでユーザの意図に即した分類を行う手法である．背景知識（制約）を与える方法には，「データの正解」情報を与える方法と，「データ間の関係」情報を与える方法の 2 つがある．前者は，ユーザがデータの一部に正解クラスタを与え，それらを教師データとして用いるクラスタリング方式である．後者は，任意の 2 つのデータに対して必ず同じ／異なるクラスタに属するという制約（それぞれ，Must-link／Cannot-link と呼ぶ）を課し，それらを満たすようなクラスタに分類する方式である．

データの正解情報を利用する方式に，Basu らが提案した Seeded-KMeans と Constrained-KMeans がある[2]．Seeded-KMeans は，正解クラスタが与えられたデータを K-Means 法の初期重心に用い，以後の K-Means 法のループ内では正解情報を使用しない方式である．Constrained-KMeans は，同じく正解情報を初期重心に利用するが，以後の K-Means 法のループ内では常に与えた正解クラスタを維持する（つまりは，正解が与えられたデータは類似度による分類を行わない）方式である．

データ間の関係を制約として利用する方式に，Wagstaff らの提案する COP-KMeans がある[3]．この方式は，ユーザが指定した Must-link／Cannot-link の関係を満たすように K-Means 法によるクラスタリングを行う．しかし，制約の数が多くなると，全ての制約を満たす所属クラスタを探索するのが困難になるという短所をもつ．この方式を拡張した提案に，[4]，[5]などがある．

以上がデータに直接制約を与える代表的な方式であるのに対して，与えられた制約を満たすようにデータ間の関係を表す尺度を変化させる方式がある．Cohn[6]らは Cannot-link 関係にあるデータ間の距離を最大化するように，データがもつ特徴量に重みを課す方式を提案している．データ間の関係を表す尺度を変化させる方式には，他にも[7]，[8]などがある．

以上が，与えた制約データの総数とクラスタリング精度との関係を評価した研究であるのに対し，本研究では制約を段階的に追加しながらデータをクラスタリングする

Vol.2009-DBS-148 No.7
Vol.2009-FI-95 No.7
2009/7/28

場合の制約数と精度との関係について着目した．

このような，段階的な制約の追加とクラスタリングを組み合わせたデータの分類方式として, desJardins らはユーザによるデータのグループ化と制約付きクラスタリングを組み合わせることで精度を向上させる対話型の分類方式を提案している[9]．しかし，[9]では 100 件程度のデータしか評価に用いていない．

本研究ではデータへの制約付与とクラスタリングを交互に繰り返すことで対話的に精度を向上させる分類方式を評価する．クラスタリングの正／誤に関わらず未付与のデータに正解を与えていった場合の精度向上効果について，複数の文書集合を用いて検証する．

さらに，対話的な文書分類の支援方式としてクラスタ安定度の数値化方式を提案する．正解付与とクラスタリングを繰り返していくと，ある時点から正解を与えた件数ぶんしかクラスタリング精度が向上しなくなる．この状態以降はクラスタリングによる精度向上はほぼ見込めない．この状態を検知するための方式を提案し，実験によりその効果を示す．

2 章では，2.1 節にて本研究が対象とする対話型分類方式について述べた後，2.2 節以降で精度向上効果の検証結果を示す．3 章では，提案するクラスタ安定度の数値化方式と評価結果について述べる．

2. 半教師有りクラスタリングを適用した対話型文書分類方式

本研究では，制約付きクラスタリングによる自動分類とユーザによるデータの制約付与操作を交互に繰り返し行う分類方式を対話型文書分類方式と呼ぶこととする．

2.1 対話型文書分類方式

本研究では, Basu らの提案した Constrained-KMeans を制約付きクラスタリングとして採用し, 初期重心計算方法に改良を加えた（後述する）対話型文書分類方式とした．Basu らの方式を採用したのは，正解を指定する制約の方がデータの関係（Must-link／Cannot-link）を与えるよりもユーザにとって与え易い制約であると判断したためである．

Constrained-KMeans は与えられた制約を教師データとして用いる半教師有りの K-Means 法である．K-Means 法は，クラスタ内分散が最小となるように集合を分割する方式であり，以下のステップで実行される．

1.クラスタ数 K を決め，初期重心を生成する．

2.データをそれぞれ最近隣の重心に対応するクラスタに分類する．全てのデータを分類した後，分類したデータに従って重心を更新する．

3.重心に変化が無くなれば処理を終了し，そうでなければ 2 に戻る．

これに対して Constrained-KMeans は正解として与えたデータのみを用いて初期重心を計算し，また，2 において正解として与えたデータの正解クラスタを維持する点が通常の K-Means 法とは異なる．

図 1 に本研究による対話型文書分類方式のフローを示す．入力は文書集合 $X=\{x_1,\cdots,x_N\}$ とクラスタ数 K である．X_h はクラスタ h に属している文書集合を表す．また，クラスタ h が正解クラスタとして与えられた文書集合を S_h とし，正解文書集合全体を $S=\bigcup_{h=1}^{K} S_h$ とする．$\mu_h^{(L,t)}$ を L 回目の Constrained-KMeans における t 回目のループで得られたクラスタ h の重心とし，μ_h^{L} を L 回目の Constrained-KMeans が収束した際に得られたクラスタ h の重心とする．$|\cdot|$ は集合の要素数を表す．

まず初めに重心をランダムに選択した初期分類（通常の K-Means 法）と，正解文書集合の初期化を行う（図１ ②）．次に，文書の正解クラスタ指定と Constrained-KMeans を交互に繰り返す（図１ ③）．③－１では，未付与文書に正解クラスタを与える．③－２において，Constrained-KMeans を実行する．本研究では Constrained-KMeans の i の処理において, 正解が与えられていないクラスタがある場合には, 前回（L－1回目）のクラスタリング結果の対応するクラスタに属する文書を使用して重心を計算することとした．

本章では，以上で示した対話型文書分類方式において，段階的に正解を付与した場合のクラスタリング精度向上効果について検証した．

2.2 実　験

2.2.1 実験に用いたデータ

評価のための文書集合に特許文献を用いた．特許庁[10]にて公開されている“移動体通信システム技術”（Mobile）に関する特許マップ[a]（137 件，3 クラスタ），特許流通促進事業[11]にて公開されている“電子透かしの応用技術”（eMark）に関する特許マップ（484 件, 5 クラスタ），“交流型プラズマディスプレイパネルの技術課題”（PDP）に関する特許マップ（1253 件，9 クラスタ），国際特許分類[b]より“診断；手術；個人識別技術（国際特許分類の A61B に対応）”（A61B）に属している特許文献（2268 件，4 クラスタ）の 4 つのデータを用いた．いずれのデータも公開年が 1993～2002 年

a) 特許マップには，特許文献を書誌情報に従って統計的に解析したグラフで表現したもの（例えば，横軸を出願年，縦軸を出願件数とした棒グラフなど）と，記載された技術内容の違いに従ってカテゴリ分けしたもの（具体例は表 1 を参照）の 2 つがある．本研究では後者を特許マップと呼ぶこととする．

b) 国際特許分類（IPC : International Patent Classification）とは，世界共通に用いられている特許の分類．

Vol.2009-DBS-148 No.7
Vol.2009-FI-95 No.7
2009/7/28

の公開公報を使用し，A61B については，各クラスタが 500 件程度となるように特許文献を抽出した. 文献中の要約部分から全ての単語を用いて特徴ベクトルを生成した. また，各単語の重み付けには TF-IDF 法を用いた.

---

①．文書集合 $X=\{x_1,\cdots,x_N\}$，クラスタ数 K を入力する.

②．初期分類（$L \leftarrow 0$）．集合 X からランダムに選択した文書を初期重心として K-Means 法を実行する．正解文書集合を $S \leftarrow \phi$ とする.

③．以下の操作を繰り返す.

③－１．任意の文書に正解を付与し，正解文書集合 S を更新する.

③－２．Constrained-KMeans の実行.

i. クラスタ h について，$S_h \neq \phi$ の場合は $\mu_h^{(L,0)} \leftarrow \frac{1}{|S_h|}\sum_{x\in S_h} x$，そうでない場合は $\mu_h^{(L,0)} \leftarrow \mu_h^{L-1}$， for $h=1,\ldots,K; t \leftarrow 0$.

ii. 収束するまで以下の処理を繰り返す.

ii-a. $x \in S_h$ であるならば，クラスタ h に分類．$x \notin S$ であるならば，最近隣のクラスタに分類.

ii-b. $\mu_h^{(L,t+1)} \leftarrow \frac{1}{|X_h^{t+1}|}\sum_{x\in X_h^{t+1}} x$.

ii-c. $t \leftarrow (t+1)$.

③－３．$L \leftarrow (L+1)$.

図 1　本研究における対話型分類方式のフロー

2.2.2 実験手順

図１のフローに従って実験を行った.

まず初めに，正解が与えられた文書が無い状態で初期重心をランダムに選択してクラスタリングする. 次に, 得られたクラスタと表 1 に示した正解カテゴリを対応付ける．正解文書の割合が最も多いクラスタにその正解カテゴリを対応付ける．そして，クラスタリング結果の正／誤に関わらず正解が未付与の文書に正解クラスタを与える操作と，Constrained-KMeans（図１中 ②－１）を交互に繰り返す．全文書を正解文書として与えるか, 精度が 100%となった時点で処理を終了する. 以上を 10 回繰り返し，精度の平均を最終的な値とした．なお，クラスタリングの際のクラスタ数 K は表 1 に示したクラスタ数を用いた.

表 1　評価に用いた文書集合

移動体通信システム(Mobile)	
1. CDMA方式	36件
2. 移動端末	48件
3. サービス	43件
計	137件

交流型ﾌﾟﾗｽﾞﾏﾃﾞｨｽﾌﾟﾚｲﾊﾟﾈﾙの技術課題(PDP)	
1. 階調表示技術の改善	274件
2. 高解像度化・大容量化	46件
3. 表示品質の改善	297件
4. 各種映像信号・画像ﾌｫｰﾏｯﾄへの対応	47件
5. 動作特性・装置特性の改善	289件
6. 低消費電力化	121件
7. 低コスト化・小型化	84件
8. 高信頼性化	66件
9. 応用技術の改善	29件
計	1253件

電子透かしの応用技術(eMark)	
1. 暗号技術	54件
2. 印刷表示技術	121件
3. 記録技術	110件
4. 通信技術	138件
5. 機器組込技術	61件
計	484件

診断；手術；個人識別技術(A61B)	
1. 診断のための検出，測定，記録	534件
2. 内視鏡，そのための照明装置	544件
3. 超音波，音波による診断	594件
4. 放射線診断用機器	596件
計	2268件

2.3 評価に用いた尺度

精度の尺度には正解率を用いた．正解率は，対応付けたクラスタに分類された正解文書数の総和を全文書数で割った値としている.

2.4 結　果

正解が付与されていない文書の中からクラスタリングの正／誤に関わらずランダムに文書を選択し正解を与えて制約付きクラスタリングを行う操作を繰り返した場合の精度向上効果の結果を示す．このような正解付与方法としたのはユーザの分類操作を再現するためである.

図 2 のグラフの横軸は全文書に占める与えた正解文書の割合，縦軸は正解率を示す．正解文書の割合とは，例えば，移動体通信システム技術の場合，横軸が 20%の点は 27（≒137×0.2）件の文献の正解を指定したことを意味している．より小さい正解文書の割合で高い正解率を示すほど良い結果であることを意味する.

いずれの文書集合においても, 前半 10%程度の正解付与では大幅な精度向上となり，以降では緩やかな向上を示す結果となった．正解付与の後半では，精度向上に寄与しない文書（例えば，重心に非常に近い文書）にも正解を付与していることが精度向上が小さくなった原因と考えられる.

特に，精度曲線がほぼ直線に近い状態になっている区間では，与えた正解文書のぶ

Vol.2009-DBS-148 No.7
Vol.2009-FI-95 No.7
2009/7/28

んしか精度が向上していなかった．この区間では，外れ値といったクラスタリングによる分類が困難な文書だけが残った状態であることが予想される．したがって，この区間以降は文書の特徴語の頻度情報だけでは正解クラスタの判別が困難であるため，別の方法による仕分けが必要と思われる．

本検証から，最も良い場合（A61B）で全件の 20%程度の文書に正解を付与すれば，90%程度のクラスタリング精度が得られることがわかった．つまり，20%程度の分類作業で 90%の文書を仕分けたのと同じ効果が得られる．

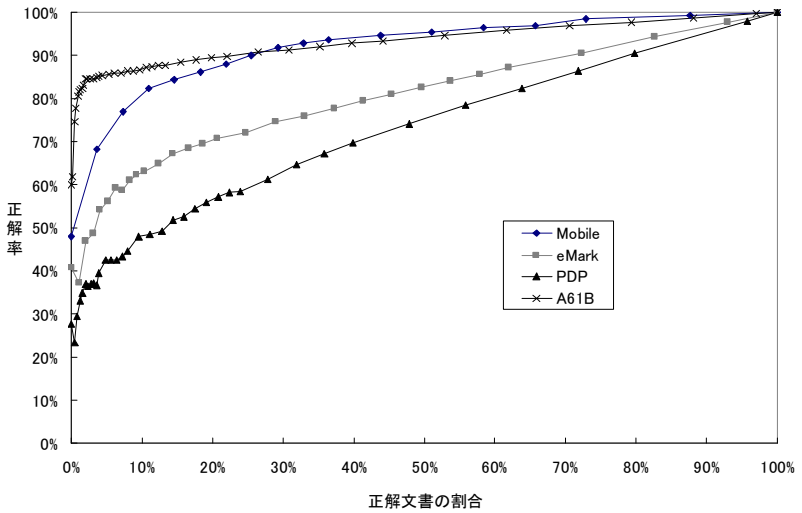

図 2　制約未付与の文書からランダムに選択して正解を与えた場合の精度曲線

用いた文書集合による精度向上の差には 2 つの理由が考えられる．1 つはクラスタ数の多さであり，もう 1 つはクラスタ間での特徴語の出現傾向である．

クラスタ数が多いほど，より多くの正解文書数を必要とする．一般的に，所望の分類に達するためにはある一定数の正解が必要となる．仮に 1 クラスタ当り 5 件の正解を必要とした場合, 3 クラスタでは 15 件, 10 クラスタでは 50 件の正解が必要となる．したがって，クラスタ数が多いデータセットほど精度向上が遅くなるものと考えられる．このことから，Mobile，eMark，PDP という順の精度結果となった．

これに対して，A61B はクラスタ数が 3 である Mobile よりも良い結果となった．これは，A61B の正解クラスタ毎の特徴語に重複がほとんど無かったことが原因の 1 つと考えられる．特徴語に重複が無ければ，境界が明確であるため精度向上が早い．一方で，例えば Mobile では，端末によってサービスの形態が異なるため（携帯電話なのか，PDA のようなモバイル PC なのか），サービスに含まれる文書の中には端末に関する特徴語をもつ文書が散見された．また，PDP は文書の意味にまで踏み込んだ仕分けがされているため，特徴語の頻度情報だけではクラスタリングが難しい文書集合であった．そのため，精度向上が遅かったものと考えられる．

2.5 考　察

本検証では，クラスタリング結果の正／誤に関わらず文書に正解を付与した．この付与操作には，正しく分類された文書に正解であることを教える操作と，誤って分類された文書に正しいクラスタを与える操作の 2 つがある．今後は，この 2 つの付与操作のどちらがより精度向上に与える影響が大きいのかを検証する必要がある．また，どういった文書に優先的に正解を与えていくと効果が高いのかも検証の余地がある．例えば，クラスタ境界に位置する文書やクラスタ重心から遠い文書を優先させる方法が考えられる．

3. クラスタ安定度の数値化

3.1 背　景

本章では，2 章で述べた方式を実際の対話型分類に応用した場合の分類支援として，クラスタ安定度の数値化方式を提案する．2.4 節でも述べたとおり，精度曲線がほぼ直線の区間は，与えた正解文書の件数ぶんしか精度が向上していない．与えた正解文書ぶんしか精度が向上していないということは，それ以外の文書のクラスタラベルがほぼ変化していない（K-Means 法の入力と出力でクラスタラベルが変化していないという意味であり，ループ中にクラスタ間を行き来している可能性はある）ということであり，つまりは，ほとんど同じクラスタに収束していると言える．この状態に達すると，誤分類している文書は文書ベクトルに用いた特徴語では正解クラスタを判別できない文書であり，K-Means 法での分類が困難な状態にある可能性が高い．つまりはクラスタリングによる自動分類の精度向上が見込めない．本章の目的は，このような対話型の分類方式における自動分類の止め時を検知することにある．

3.2 提案方式

精度曲線がほぼ直線を描く区間ではクラスタリングによる精度向上効果がほぼ無いに等しい．したがって，精度曲線を近似するような指標によりこの区間を特定できるような方式を検討した．本研究では，このような区間ではほぼ同じクラスタに収束していることから，この指標をクラスタ安定度と呼ぶこととする．

クラスタ安定度を，1 度の Constrained-KMeans（図 1　③－２の処理）が収束するまでにクラスタラベルが変わった文書数の総和とした．クラスタ数を K，クラスタリン

Vol.2009-DBS-148 No.7
Vol.2009-FI-95 No.7
2009/7/28

グが収束までに要したループ数をT，i回目のループにおいて，クラスタjにクラスタラベルが変わった文書数を $change\_num_{ij}$ とすると，クラスタ安定度 I_{conv} を以下の式で表す．

$$I_{conv} = \sum_{i=0}^{T} \sum_{j=1}^{K} change\_num_{ij}$$

表 2に従って，クラスタ安定度の計算例を示す．表 2は10件の文書集合のL回目のクラスタリングのクラスタラベルの変化を示した例である．表内の値は各文書のクラスタラベルを表す．この例では，初期分類（i = 0）と1回目のループ（i = 1）の計2回でクラスタリングが収束したものとしている．(L-1)回目の結果は，前回のクラスタリングにより得られた各文書のクラスタラベルを指す．

表 2　クラスタラベルの変化の例

		文書1	文書2	文書3	文書4	文書5	文書6	文書7	文書8	文書9	文書10	変化量
(L-1)回目の結果		1	1	2	2	2	2	2	2	2	3	—
L回目	i = 0	3	1	1	2	1	2	2	2	3	3	4
	i = 1	1	1	1	2	2	3	3	3	3	3	3

まず，各ループiにおけるクラスタラベルの変化量はi-1回目のループとは異なるラベルをもつ文書数である．(i = 0)の時は前回（L-1回目）のクラスタリング結果からの変化をカウントする．したがって，表 2の例の場合では，(i = 0)での変化量が4，(i = 1)での変化量が3となり，クラスタ安定度は7と計算される．なお，(L = 0)回目のクラスタリング時は前回のクラスタリング結果が存在しないので，クラスタ安定度を計算しない．

3.3 実　験

提案方式と精度曲線の関係を検証した．図 2に示した精度曲線のそれぞれの点で正解率とクラスタ安定度との相関を計算し，精度曲線と似た挙動を示すかを検証した．

精度曲線とクラスタ安定度の関係を図 3に示す．グラフの横軸は与えた正解文書数，縦軸は左がクラスタ安定度，右が正解率を表す．また，表 3に精度曲線とクラスタ安定度との間の相関係数の絶対値を示す．

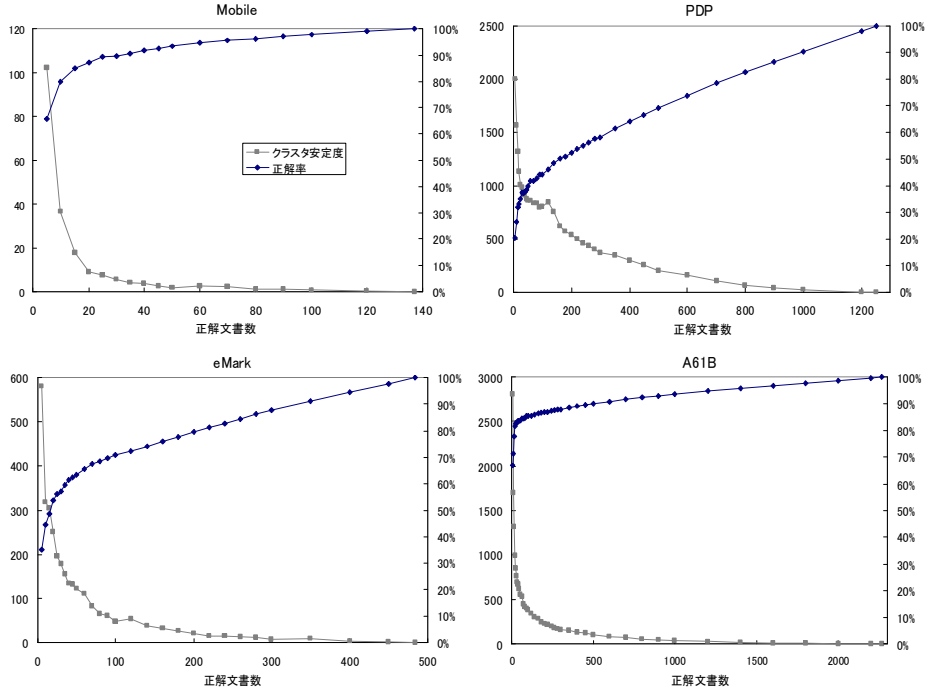

図 3　精度（正解率）曲線とクラスタ安定度の関係

表 3　精度曲線とクラスタ安定度の相関係数の絶対値

Mobile	eMark	PDP	A61B
0.904	0.883	0.922	0.883

3.4 結果と考察

図 3，表 3に示すとおり，提案するクラスタ安定度は精度曲線とある程度似た挙動を示していることがうかがえる．表 3に示した相関係数の絶対値からもそのことがわかる．このことから，実際の対話型文書分類方式において，クラスタ安定度を提示することでクラスタ状態の目安を知ることができ，クラスタリングによる精度向上が見込めないような操作を回避できる可能性がある．

なお，クラスタ安定度の可視化の例として，図 4に示したような棒グラフを提示

Vol.2009-DBS-148 No.7
Vol.2009-FI-95 No.7
2009/7/28

する方式が考えられる．変化量の総和に加えて，クラスタ毎の値を示すことで変化量の大きいクラスタから重点的に正解を与えていく等の効率的な分類が可能となる可能性がある．

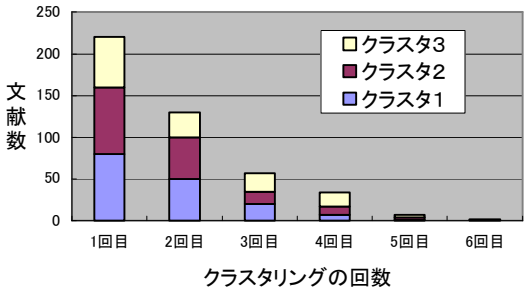

図 4　クラスタ安定度の可視化方式の例

4. おわりに

本研究では，半教師有りクラスタリングを用いた対話型の文書分類方式における精度向上効果を検証した．また，対話的に分類を行う際のクラスタリング精度向上が見込めなくなる状態を検知する方式（クラスタ安定度の数値化）を提案した．

対話型文書分類方式の精度検証では，クラスタリング結果の正／誤に関わらずランダムに選択した文書への正解付与と制約付きクラスタリングを繰り返すことで，最大で20%程度の正解を与えれば約90%のクラスタリング精度が得られることがわかった．これにより，大幅に文書分類の作業量を減らすことができる可能性がある．

また，クラスタ安定度の数値化方式は，正解率との相関係数との絶対値が平均で約0.9 という高い値を得ることができた．本方式により，クラスタリングによる精度向上効果が見込めなくなる状態を検知できる可能性がある．

本研究の提案方式を応用することで，対話的な文書分類作業を従来よりも少ない作業量で効率的に行えるようなツールの実現が期待できる．

参考文献

1) J. B. MacQueen: Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297 (1967)
2) S. Basu, A. Banerjee & R. Mooney: Semi-supervised Clustering by Seeding, Proceedings of the 19th International Conference on Machine Learning, pp.19-26 (2002).
3) K. Wagstaff, C. Cardie, S. Rogers & S. Schroedl: Constrained K-means Clustering with Background Knowledge, Proceedings of the 18th International Conference on Machine Learning, pp.577-584 (2001).
4) H. Yang & J. Callan: Near-Duplicate Detection by Instance-level Constrained Clustering, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (2006).
5) 新納浩幸，佐々木稔，村上浩司: 制約を修正に用いた半教師有りクラスタリング，2006 年情報論的学習理論ワークショップ (2006).
6) D. Cohn, R. Caruana & A. McCallum: Semi-supervised Clustering with User Feedback, Technical Report TR2003-1892, Cornell University (2003).
7) E. Xing, A. Ng, M, Jordan & S. Russell: Distance Metric Learning, with Application to Clustering with Side-Information, NIPS 15 (2003).
8) S. Basu, M. Bilenko & R. Mooney: A Probabilistic Framework for Semi-Supervised Clustering, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.59-68 (2004).
9) M. desJardins, J. MacGlashan & J. Ferraioli: Interactive Visual Clustering for Relational Data, Constrained Clustering: Advances in Algorithms, Theory, and Applications, Chapman & Hall, pp.329-374 (2008).
10) 特許庁 資料室 「技術分野別特許マップについて」 http://www.jpo.go.jp/shiryou/s_sonota/tokumap.htm.
11) 特許流通促進事業 「特許流通支援チャート」 http://www.ryutu.inpit.go.jp/chart/tokumapf.htm.