情 報 学 基 礎 61－28
自 然 言 語 処 理 142－28
（2001．3．6）

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

則武淳* 吉高淳夫** 平川正人**
*広島大学大学院工学研究科
**広島大学工学部第二類（電気系）

〒739-8527 東広島市鏡山1丁目4番1号
{jun, 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反映した文書クラスタリングを行うために、利用者の分類例示に基づいて特徴要素を選出する手法を提案する。利用者の分類例示に基づいて形成されるクラスタにおいて、同じクラスタに属する文書間の類似度と異なるクラスタに属する文書間の非類似度の総和をクラスタの分離度と定義し、分離度が最大となるように特徴要素を選出する。また、分類例示が不十分な場合、特徴要素を適切に選出できないためクラスタの境界が不明確となり、クラスタリングに曖昧性が生じる。曖昧性が生じた場合、曖昧性の解消に適する文書を選出し、それを利用者に提示して分類例への追加を促すことで、適切な特徴要素を選出するために十分な分類例を獲得する。

Document Clustering by Example-Based Feature Subset Selection

Jun NORITAKE*, Atsuo YOSHITAKA**, and Masahito HIRAKAWA**
*Graduate School of Engineering, Hiroshima University
**Faculty of Engineering, Hiroshima University

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

In this paper, we propose a feature subset selection method based on an example for document clustering reflecting a user's intension. Separability of clusters is defined as the total sum of similarities between documents belonging to the same cluster and dissimilarities between documents belonging to different clusters. Several important features are selected to maximize separability. However insufficient example may causes the boundary of clusters ambiguous. Several documents that disambiguate the boundary of clusters are selected by the system when the boundaries of clusters are ambiguous. Sufficient example is acquired by adding those documents to the example.

1. はじめに

インターネットの普及により、膨大な数のオンライン情報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。従来の検索手法では、利用者が検索要求をキーワードとして入力し、そのキーワードを含む情報を検索結果として出力している。情報検索の専門家であるサーチャは、対象とするデータベースに関する知識

を利用して適切なキーワードを選択し、検索結果を効果的に絞り込むことができる。しかし、対象とするデータベースに関する知識を持っていない一般の利用者にとって、適切なキーワードの選択は容易なことではない。そのため、検索結果を十分に絞り込むことができず、大量の検索結果の中から必要な情報を探すことになる。その結果、利用者に大きな負担を与えることになる。

大量の情報の中から必要な情報を探す場合、それらを分類することで、効率的な探索が期待できる。しかしながら、分類の仕方は様々であり、利用者の意図するような分類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であるとはいえない。そのため、利用者の分類に対する意図を考慮して、対象情報を分類することが重要である。

分類手法のひとつとしてベクトル空間モデルに基づくクラスタリングが挙げられる。クラスタリングとは、データ間の類似性に基づいて、データをいくつかのグループに分類することである。そのため、データ間の類似度を算出する必要がある。ベクトル空間モデルでは、データは各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、その基底となる特徴要素によってデータ間の類似度は変化する。したがって、特徴要素を適切に選出することで、利用者の分類に対する意図をクラスタリングに反映できると考えられる。しかし、利用者は全対象データの内容を把握しているわけではないため、分類に対する意図を特徴要素の集合として、システムに明示的に与えることは困難である。これに対して、システムから与えられたサンプルデータを用いて利用者が分類例示を行うことは容易であり、分類例示の結果に基づいてシステムが適切な特徴要素を選出することで、利用者の分類に対する意図をクラスタリングに反映できると考えられる。

特徴要素は、①ある基準に従って各特徴要素にスコアを与え、選出する順序を決める、②決められた選出順序に従って、ある評価値が最大(最小)となる個数の特徴要素を選出する、という手順で選出される。そのため、スコアの算出法と選出個数に関する評価基準を設定する必要がある。

[1]では、あらかじめ分類されたサンプル文書を表す特徴ベクトルの各要素について、(グループごとの平均値の分散) / (全文書を通じての分散)を求め、その特徴要素に与えるスコアとしている。そして、間違えて分類された文書の数が最小となる個数の特徴要素を選出している。[2]では、特徴ベクトルの各要素を除いたときの、文書間の類似性判定についての不確定さの度合いを、エントロピーの概念に基づいて数量化し、その特徴要素に与えるスコアとしている。そして、クラスタリングを行い、得られたクラスタの分離度が最大となる個数の特徴要素を選出している。これらの手法では、ある選出順序の下での最適な選出個数を決定しているため、選出順序の最適化については考慮されていない。また、実際に分類(クラスタリング)を行った結果に基づいて特徴要素の選出個数を決定しているので、処理に時間がかかるという問題がある。その他にも、特徴要素の選出順序を決めるためのスコア付け手法が提案されているが、特徴要素の選出個数に関する評価基準については考慮されていない[3]-[7]。

本研究では、利用者の分類に対する意図を反映した文書クラスタリングを行うために、利用者の分類例示に基づいて、適切な特徴要素を選出する手法を提案する。本手法では、分類例示に基づいて形成されるクラスタの分離度を高める効果の大きい順に、分離度が最大となる特徴要素まで選出することで、特徴要素の選出個数に対する分離度の最大化を目指す。ただし、各特徴要素の分離度を高める効果は、特徴要素の選出順序に依存するため、分離度の算出と選出順序の並べ替えを繰り返し行い、結果を改善していく。

2. 分類例示に基づいた特徴要素の選出

クラスタリング対象の文書集合から、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。その分類例示に

基づいて、同じグループに属する文書間の類似度をできる限り高く、異なるグループに属する文書間の類似度をできる限り低くするように特徴要素集合を選出する。そして、選出された特徴要素集合によって多次元ベクトル空間を構成し、その多次元ベクトル空間内に配置した各文書の特徴ベクトルに対してクラスタリングを行うことで、利用者の分類に対する意図をクラスタリングの結果に反映する。

多次元ベクトル空間を構成する特徴要素は、サンプル文書中に出現する特徴要素の中から選出される。分類例示が不十分な場合、特徴要素を適切に選出できないため、クラスタリングの結果が曖昧性を含んだものになる。適切な特徴要素を選出するためには、十分な数のサンプル文書を用いて分類例示する必要がある。しかし、多数の文書を利用者に提示すると、分類例示の際に利用者にかかる負担が大きくなる。

本研究では、最初に与えるサンプル文書の数は、利用者に過度の負担をかけない程度に抑える。そして、クラスタリング処理過程において曖昧性が生じた場合に、曖昧性を解消するために適した文書を選出し、それを利用者に提示して分類例への追加を促すことで、十分な分類例を獲得する。

3. 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、クラスタリング処理過程に自動的に形成されるクラスタを提案クラスタと呼ぶことにする。

利用者が与えられたサンプル文書を用いて分類例示すると、その分類例示に基づいて特徴要素が選出される。各文書は、選出された特徴要素集合によって構成される多次元ベクトル空間内に配置され、それらの文書に対してクラスタリングが行われる。初期状態においては、分類例示されたグループと同数の例示クラスタと、文書が1つだけ属している多数の提案クラスタが存在する。そして、最も類似度が高いクラスタのペアが結合の候補に挙げられる。クラスタ間の類似度は、各クラスタに含まれる文書の特徴ベクトルの平均をクラスタの代表ベクトルとして、代表ベクトル間の類似度(余弦尺度)として求められる。

結合の候補に挙げられたクラスタのペアが共に例示クラスタの場合、それらのクラスタは結合されず、次に類似度が高いクラスタのペアが結合の候補に挙げられる。結合の候補に挙げられたクラスタのペアが共に提案クラスタの場合、それらのクラスタは結合され、次に類似度が高いクラスタのペアが、次の結合の候補に挙げられる。結合の候補に挙げられたクラスタのペアの一方が例示クラスタの場合、それらの結合に曖昧性が生じているか判定される。曖昧でないと判定された場合、それらのクラスタは結合され、次に類似度が高いクラスタのペアが、次の結合の候補に挙げられる。曖昧であると判定された場合、曖昧性の解消に適した文書が選出され、利用者に提示される。利用者が分類例を修正すると、再び特徴要素が選出され、各文書は新たに選出された特徴要素集合によって構成される多次元ベクトル空間内に配置される。

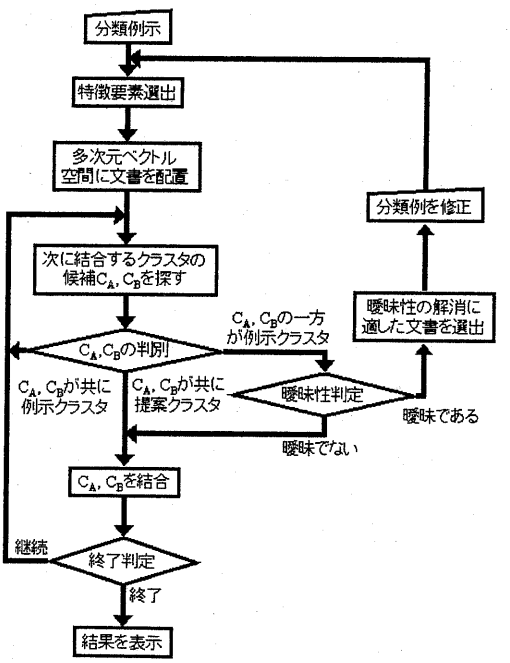

図1 処理の流れ

以上のような処理が繰り返され、次第にクラスタの個数は減少していく。そして、結合できるクラスタのペアが無くなると終了する。クラスタリング処理の終了時には、全ての文書がいずれかの例示クラスタに属している。

4. 特徴要素の選出手法

4.1 分離度の定義

多次元ベクトル空間を構成する特徴要素 f_k の集合を $F_x=\{f_1, \cdots, f_m\}$ とすると、文書 d_i の特徴ベクトル $\mathbf{d}_i^{F_x}$ は以下のように表現できる。

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

ここで、$w_i(f_k)$は文書 $d_i(i=1, \cdots, n)$の特徴要素 f_k $(k=1, \cdots, m)$に対する重みである。

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 によって構成される多次元ベクトル空間内において、文書 d_i と d_j の類似度 $SIM^{F_x}(d_i, d_j)$は、以下のようにして求められる。

$$SIM^{F_x}(d_i, d_j)=\frac{\mathbf{d}_i^{F_x}\cdot\mathbf{d}_j^{F_x}}{\left\|\mathbf{d}_i^{F_x}\right\|\left\|\mathbf{d}_j^{F_x}\right\|}$$

これは、2 つの特徴ベクトル $\mathbf{d}_i^{F_x}$ と $\mathbf{d}_j^{F_x}$ を大きさ 1 に正規化し、その内積を求めたものであり、2 つのベクトルが直交するとき最小値 0、重なり合うとき最大値 1 となる。$\|\mathbf{d}_i^{F_x}\|=0$、または、$\|\mathbf{d}_j^{F_x}\|=0$ の場合、上式では文書間の類似度を算出することができない。このとき、文書 d_i と文書 d_j に共通して出現する特徴要素がないので、それらの文書は類似していないと判断し、類似度を 0 とする。

特徴要素の選出基準として、特徴要素の集合 F_x によって構成される多次元ベクトル空間内において、利用者の分類例示に基づいて形成されるクラスタの分離度 SEP^{F_x} を以下のように定義する。

$$SEP^{F_x}=\sum_{(d_i,d_j)\in R_s} SIM^{F_x}(d_i, d_j)+\sum_{(d_i,d_j)\in R_d}\{1-SIM^{F_x}(d_i, d_j)\}$$

ここで、R_s は例示に用いた全文書から 2 つの文書を取り出す場合、取り出すことのできる全ての組み合わせ(d_i, d_j)の集合 R_0 において、d_i と d_j が同じクラスタに属する組み合わせの集合である。また、R_d は集合 R_0 において、d_i と d_j が異なるクラスタに属する組み合わせの集合である。この式は、同じクラスタに属する文書間の類似度が高く、異なるクラスタに属する文書間の類似度が低くなれば、クラスタの分離度が高くなることを示している。

4.2 スコア算出法

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。多次元ベクトル空間を構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 SEP^{F_x} 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 に依存するので、各特徴要素に与えるスコアの算出と特徴要素の選出順序の並べ替えを繰り返し行い、結果を改善していく。各特徴要素のスコアの算出法を以下に示す。

(1) 次式に従って各特徴要素に与える初期スコア $V(f_k)$を計算する。

$$V(f_k)=SEP^{\{f_k\}}$$

(2) 与えられた初期スコアの高い順に特徴要素を並べる。

(3) 次式に従って各特徴要素に与えるスコア $V(f_k)$ を再計算する。

$$V(f_k)=SEP^{F_k\cup\{f_k\}}-SEP^{F_k}$$

ここで、F_k は、前回の計算によって与えられたスコアが、特徴要素 f_k より高い特徴要素の集合である。

(4) 新たに与えられたスコアの高い順に特徴要素を並べ替える。前段階と比較して順序の変更がない場合は処理を終了する。順序の変更がある場合、(3)へ戻る。

4.3 特徴要素の選出

最終的に与えられたスコアの高い順に、分離度が最大となる個数の特徴要素を選出する。多次元

ベクトル空間を構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 SEP^{F_x} 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、与えられたスコアが正の特徴要素を全て選出したとき、分離度が最大となる。

5. 曖昧性の解消

5.1 クラスタリングに生じる曖昧性

本研究では、以下に述べる2つの状態をクラスタリングに生じる曖昧性として、これらの解消を目指す。

(1) 結合候補のクラスタ間の距離が遠い場合

図 2(a)のような文書の分布状態を考える。結合しないという条件が与えられている例示クラスタ C_{E1} と C_{E2} の距離よりも、結合の候補に挙げられた提案クラスタ C_x と例示クラスタ C_{E1} の距離が遠いため、これらのクラスタを結合するべきか、結合するべきでないかという曖昧性が生じている。

(2) クラスタの結合候補が複数ある場合

図 2(b)のような文書の分布状態を考える。提案クラスタ C_x と結合する例示クラスタの候補が C_{E1}、C_{E2} と複数存在するため、提案クラスタ C_x をどちらの例示クラスタと結合するべきかという曖昧性が生じている。

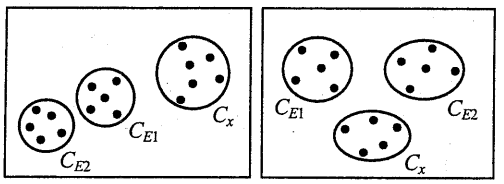

図2 文書の分布の例

これらのような状態の場合、①C_x 中の文書と C_{E1} (あるいは C_{E2})中の文書の類似度を高くするような特徴要素を選出し、C_x と C_{E1} (C_{E2})を近づける、あるいは、②C_x 中の文書と C_{E1}(C_{E2})中の文書の類似度を低くするような特徴要素を選出し、さらに C_x 中の文書を新たに定義した例示クラスタ C_{E3} に配置する、ことで C_x と C_{E1}(C_{E2})の結合に対する曖昧性が解消される。

5.2 曖昧性の判定

クラスタの代表ベクトルを用いて算出した類似度を単純にクラスタ間の距離の尺度とすると、各クラスタ内の文書の分布状態を無視しているため、クラスタ間の正確な距離を表現できない。そこで、各クラスタ内の文書の平均類似度を文書の分布具合の尺度として考慮に入れ、クラスタ間の距離を求める。

特徴要素の集合 F_x によって構成される多次元ベクトル空間内において、クラスタ C_A と C_B の類似度を $SIM^{F_x}(C_A, C_B)$、クラスタ C_A 内の文書の平均類似度を $AVS^{F_x}(C_A)$ とし、C_A と C_B の距離 $DIS^{F_x}(C_A, C_B)$ を以下のように定義する。

$$DIS^{F_x}(C_A, C_B) = \frac{AVS^{F_x}(C_A)AVS^{F_x}(C_B)}{SIM^{F_x}(C_A, C_B)}$$

$$AVS^{F_x}(C_A) = \frac{1}{n(R_A)} \sum_{(d_i, d_j) \in R_A} SIM^{F_x}(d_i, d_j)$$

ここで、R_A は C_A に属する文書から2つの文書を取り出す場合に、取り出すことのできる全ての組み合わせ(d_i, d_j)の集合、$n(R_A)$は集合 R_A に属している要素の数である。そして、以下に示す条件に基づいて、曖昧性を判定する。

(1) 曖昧性の判定条件1

結合の候補に挙げられた提案クラスタ C_x と例示クラスタ C_{E1} の距離 $DIS^{F_x}(C_x, C_{E1})$が閾値 γ 以上の場合、C_x と C_{E1} の結合に対して曖昧性が生じていると判定する。

(2) 曖昧性の判定条件2

結合の候補として提案クラスタ C_x と例示クラスタ C_{E1} が挙げられたとする。また、C_x との類似度が2番目に高い例示クラスタを C_{E2} とする。そして、$DIS^{F_x}(C_x, C_{E1})$ と $DIS^{F_x}(C_x, C_{E2})$の差の絶対値が閾値 δ 以下の場合、C_x と C_{E1} の結合に対して曖昧性が生じていると判定する。

5.3 利用者に提示する文書の選出

曖昧性判定の対象になった提案クラスタ C_x から文書を選出し、利用者に提示して分類例への追加を促す。提示した文書が、どの例示クラスタに属するかを利用者が示すことで、提案クラスタ C_x

の結合に対する曖昧性が解消される。

クラスタの結合に対して曖昧性が生じる原因は、そのクラスタに属する文書の主題を表現した特徴要素が選出されていないためであると考えられる。したがって、曖昧性判定の対象になった提案クラスタ C_x 内の多くの文書に出現する特徴要素を含む文書を選出することが重要である。

特徴要素の全体集合 F_0 から、利用者の分類例示に基づいて選出された特徴要素集合 F_s を除いた特徴要素集合を F_b とする。また、文書 d_i の特徴要素 f_k に対する重みを $w_i(f_k)$、提案クラスタ C_x に属する文書の中で特徴要素 f_k が出現する文書の数を $N_x(f_k)$ として、曖昧性解消に対する文書 d_i の有効性 $EFF(d_i)$ を以下のように定義する。

$$EFF(d_i)=\frac{\sum_{f_k\in F_b} w_i(f_k)N_x(f_k)}{\sqrt{\sum_{f_k\in F_b}\{w_i(f_k)\}^2}}$$

この値が高い文書は、C_x に属する文書の主題を表した特徴要素を多く含むと考えられ、例示に用いることで C_x と他のクラスタの結合に対する曖昧性を解消することができる。

6. 評価実験

6.1 実験に用いた文書集合

インターネット関連のニュースを扱ったサイト[9]に掲載された 840 の記事、電子情報通信学会のサイト[10]で公開されている 1997 年から 2000 年の論文誌に掲載された 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、特徴要素として不適切な単語(名詞の中でも “接尾辞”、“接尾”、“非自立”、“代名詞”、“接頭詞”、“数”、“形容動詞語幹”、“副詞可能”と判定されたもの)を除いた。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て採用した。その結果、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、3520、3135 となった。

6.2 評価方法

データベース作成に用いた文書群をあらかじめ分類しておき(分類結果を表 1、表 2 に示す)、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判断し、以下の値を用いて評価を行った。

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

本研究で提案した特徴要素選出手法と他の特徴要素選出手法の比較評価を行った。また、提案した例示文書の選出法の有効性を評価した。

(1) 他の特徴要素選出手法との比較評価

比較対象として、文献[7]の比較実験で用いられている DF 法、IG 法、χ^2 法を取り上げ、用意した 2 種類のデータベースに対して実験を行った。

あらかじめ分類した各グループから無作為に 3 つ、5 つの文書を取り出し、それを分類例示とした。それぞれに対し、比較対象の 3 手法と本手法を用いて特徴要素の選出順序を決め、選出個数を 10, 20, 30, ･･･と変化させて、precision を求めた。ただし、precision は 10 回の分類例示より得られた値の平均として求めた。

(2) 例示文書の選出法の評価

クラスタリング処理過程に、本手法に基づいて曖昧性の解消に適した文書を選出し、それを分類例に追加していった場合と、無作為に選出された特定数の文書を用いて分類例示した場合に対して、例示文書数を 10, 20, 30, ･･･と変化させて、precision を求めた。ただし、本手法において、最初の例示文書は以下に述べる手法に基づいて 10 個(少なくとも各グループから 1 つ以上)選出した。

文書を分類する場合、特定の文書のみに出現する特徴要素よりも、多くの文書に出現する特徴要

素が重要である[7]。そこで、多くの文書に出現する特徴要素に対する重みが高い文書を選出するために、以下に示す評価値 $EFI(d_i)$ の高い文書から順に選出した。

$$EFI(d_i) = \frac{\sum_{f_k \in F_t} w_i(f_k) N_0(f_k)}{\sqrt{\sum_{f_k \in F_t} \{w_i(f_k)\}^2}}$$

ここで、F_t は特徴要素の全体集合から既に選出されている文書に含まれる特徴要素を除いた部分集合、$N_0(f_k)$ は全文書の中で特徴要素 f_k が出現する文書の数である。

表1　分類結果(ニュース記事)

トピック	文書数
電子商取引	87
コンテンツ配信	81
携帯端末	54
セキュリティ	84
ネットワーク	58
金融	99
検索サービス	97
パソコン	100
業界動向	99
プロバイダ	81
	計 840

表2　分類結果(論文)

トピック	文書数
音声	78
通信網	81
ﾊﾞｲｵｻｲﾊﾞﾈﾃｨｯｸｽ	157
アンテナ	131
非線形問題	82
ﾃﾞｨｼﾞﾀﾙ信号処理	95
電磁環境	67
画像処理	145
	計 836

6.3　実験結果

(1) 他の特徴要素選出手法との比較評価

実験結果を図3～6に示す。比較対象に取り上げた DF 法は、多くの文書に出現する特徴要素から順に選出するというものであり、分類例示によらない手法である。本手法と DF 法を比較することで、特に特徴要素の選出個数が少ない部分では、分類例示によって大きな効果が得られていることが分かる。IG 法、χ^2 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、本手法がわずかながら良い結果を示している。

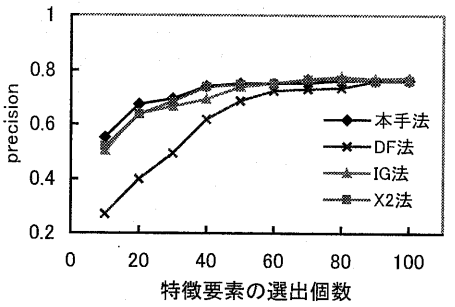

図3　他の特徴要素選出手法との比較
(ニュース記事、例示文書 30)

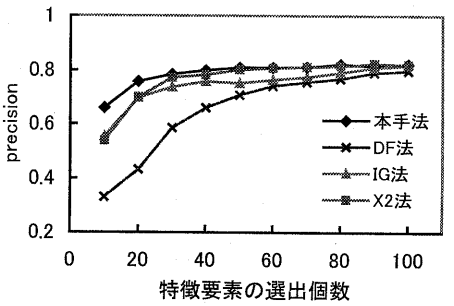

図4　他の特徴要素選出手法との比較
(ニュース記事、例示文書 50)

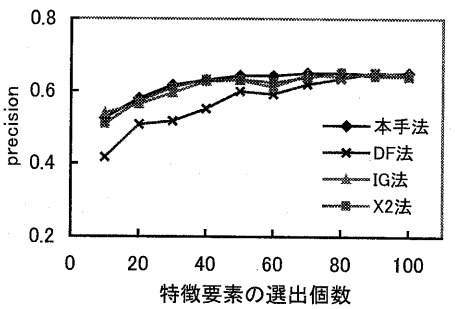

図5　他の特徴要素選出手法との比較
(論文、例示文書 24)

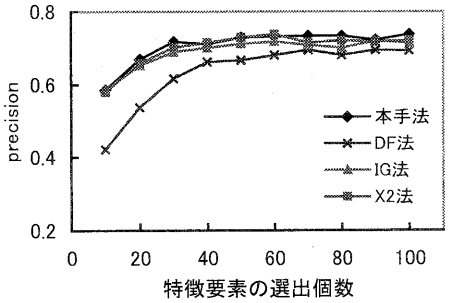

図6　他の特徴要素選出手法との比較
(論文、例示文書40)

(2) 例示文書の選出法の評価

実験結果を図7、8に示す。本手法により例示文書を選出することでprecisionが向上し、本手法の有効性を確かめられた。また、例示に用いた文書の数が多くなると、無作為に例示文書を選出した場合との差がなくなる傾向が見られた。

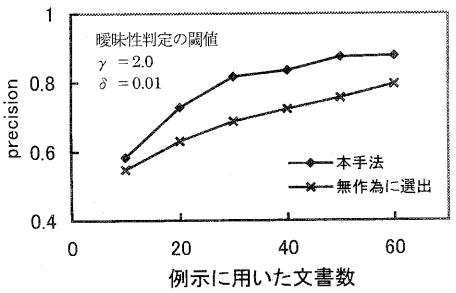

図7　例示文書選出法の評価実験(ニュース記事)

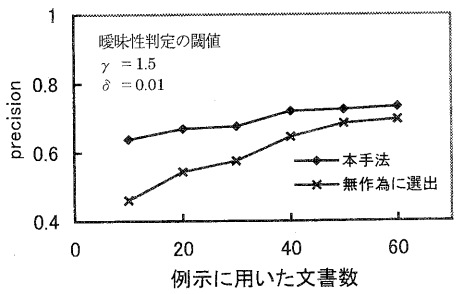

図8　例示文書選出法の評価実験(論文)

7. まとめ

本研究では、利用者の分類に対する意図を反映した文書クラスタリングを行うために、利用者の分類例示に基づいて特徴要素を選出する手法を提案した。評価実験により、①少数の特徴要素で高い分類精度が得られる、②少数の例示文書で適切な特徴要素を選出できる、ということを示した。これにより、クラスタリング処理時間の短縮、例示の際に利用者にかかる負担を軽減できると考えられる。クラスタリングの精度をさらに上げるために、同義語、多義語を考慮して意味的に独立した特徴要素を特徴ベクトルの基底とすることが、今後の課題として挙げられる。

参考文献

[1] S. Chakrabarti, B. Dom, R. Agrawal and P. Raghavan, "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies," The VLDB Journal 7(3), pp. 163-178, 1998.

[2] M. Dash and H. Liu, "Feature Selection for Clustering," In Proceedings of PAKDD 2000, pp. 110-121, 2000.

[3] D. Koller and M. Sahami, "Toward Optimal Feature Selection," In Proceedings of ICML '96, pp. 284-292, 1996.

[4] D. Mladenić, "Feature subset selection in text-learning," In Proceedings of ECML '98, pp. 95-100, 1998.

[5] H. Schütze, D. Hull and J. Pedersen, "A Comparison of Classifiers and Document Representations for the Routing Problem," In Proceedings of SIGIR '95, pp. 229-237, 1995.

[6] Y. Yang, "Noise Reduction in a Statistical Approach to Text Categorization," In Proceedings of SIGIR '95, pp. 256-265, 1995.

[7] Y. Yang and J. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proceedings of ICML '97, pp. 412-420, 1997.

[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://search.impress.co.jp/

[10] http://search.ieice.org/