ドキュメントデータ群を対象とした文脈依存動的クラスタリングおよび意味的データマイニング方式

吉田尚史[†] 図子泰三[††]
清木康[††] 北川高嗣[†††]

本論文では，ドキュメントデータ群を対象とした文脈依存動的クラスタリングおよび意味的データマイニング方式を提案する．本方式の特徴は，ドキュメントデータの意味を考慮し，文脈に応じて動的にドキュメントデータ群のクラスタリングを行い，さらにクラスタ群からの知識発掘を可能とする点にある．本方式により，分析対象であるドキュメントデータ群を対象として，文脈や視点に応じた意味的分析結果を動的に得ることが可能となる．さらに，本方式により，多数のドキュメントデータ群を対象として効率的なブラウジングを可能とする．実際のドキュメントデータ群を用いた実験結果を示し，提案方式の実現可能性および有効性を確認する．

A Context Dependent Dynamic Clustering and Semantic Data Mining Method for Document Data

Naofumi Yoshida,[†] Taizo Zushi,[††] Yasushi Kiyoki[††] and Takashi Kitagawa[†††]

In this paper we propose a context dependent dynamic clustering and semantic data mining method for document data. The main feature of the method is to make clustering for raw data semantically according to a given context. By using this method, we can dynamically obtain a set of semantic clusters of documents from a set of raw data according to a given context. This method also enables efficient browsing for a large volume of document data. We clarify the feasibility and effectiveness of the method by showing several experimental results using real document data.

1. はじめに

近年，コンピュータネットワーク上の多種多様なドキュメントデータ(文書データ)が検索対象となっている．特に電子新聞や World Wide Web の普及に伴い，検索対象となるドキュメントデータの数は爆発的に増えつつある．ドキュメントデータの検索方式として，全文検索エンジンなどが広く用いられている．しかし，それらの検索結果は一般に膨大な数のドキュメントデータの集合となり，多数のドキュメントデータ群を対象とした的確なデータ獲得の実現が課題となっている．

† 筑波大学大学院 工学研究科
Doctroal Program in Engneering, University of Tsukuba
†† 慶應義塾大学 環境情報学部
Faculty of Environmental Information, Keio University
††† 筑波大学 電子・情報工学系
Institute of Information Sciences and Electronics, University of Tsukuba

このような課題について，ドキュメントデータ群を対象とした自動分類に関する研究[7]，ドキュメントデータ群の構造の把握を支援する研究[4]，ドキュメントデータ群を対象とした問い合わせ言語や問い合わせ方式の研究[11]が行われてきた．また，データマイニングに関する研究[5),8)]を応用し，ドキュメントデータ群から静的な知識やルールを発見するドキュメントマイニングの研究[12]も活発である．それらの研究では，ドキュメントデータ内やドキュメントデータ群について，主としてドキュメントの静的な性質を対象とした知識獲得または知識発掘の方式を示している．

通常，ドキュメントデータは多くの事象を内包しており，その重要となる部分は分析時や検索時の視点に依存する．本論文の目的は，そのようなドキュメントデータの意味的な多面性を考慮して，多数のドキュメントデータ群を対象とした文脈や視点に応じた動的な集約化または集合化を実現することにより，文脈や視点に応じた知識獲得を実現することにある．

本論文では，ドキュメントデータ群を対象とした文脈依存動的クラスタリングおよび意味的データマイニング方式を提案する．提案方式の特徴は，ドキュメントデータの意味を考慮し文脈に応じて動的にそれらのクラスタリング分析を行う点，および，クラスタリング分析によって集約化されたドキュメントデータ群を対象とした知識獲得を可能とする点にある．本方式により，分析対象であるドキュメントデータ群を対象として，文脈に応じて多数の意味的分析結果を得ることが可能となる．さらに，本方式により，多数のドキュメントデータ群を対象として効率的なブラウジングが可能となる．

クラスタリングについては，多変量解析の分野やデータベースの分野において多くの方式が提案されている[3),15)]．従来の方式との比較において，本方式の特徴は，文脈に応じて動的に分析結果を獲得することができる点にある．すなわち，本方式は，一つの分析対象について静的に分析結果を得るのではなく，文脈や状況に応じて動的に分析結果を得ることを可能とする．

文脈依存動的クラスタリング方式におけるデータの文脈に応じた動的な意味的解釈については，意味的連想処理機構[9),10),16)]を用いて実現している．意味的連想処理機構では，直交空間の部分空間選択を行う演算を定義し，その演算によりデータの意味を文脈に応じて動的に解釈する機構を実現している．文脈依存動的クラスタリング方式は，この部分空間選択の機構を用いて，文脈を反映した部分空間上に（ドキュメント）データ群のマッピングを行った後に，それらのマッピングされたデータ群を対象としたクラスタリングを行うことにより，文脈に応じた動的なクラスタリングを実現する．この方式では，この部分空間選択の後に，クラスタリングのアルゴリズムを適用する[17)]．分析対象に応じて，自由にクラスタリングのアルゴリズムを選択可能である．

本方式は，意味的連想処理機構を用いて実現される．この機構はデータ間の意味的な関係を文脈に応じて動的に計算する体系を与えている．本方式は，意味的連想処理機構の文脈理解機能を応用し，文脈に応じた意味的なクラスタリングを行うことを可能とする．データの意味を計算する方式についての従来研究として，多変量解析による空間生成を用いた情報検索方式，検索者の印象によるメディアデータ抽出に関する研究，および，曖昧検索に関する研究がある．動的なクラスタリング方式を実現している研究として，文献6)が挙げられる．検索者や分析者の多様な観点に漸進的に適応し動的なクラスタリングを可能とする方式を示している．これらの研究との比較において，本方式は次の点を特徴とする．本機構では，直交空間における部分空間の選択を行う演算を定義し，その演算によりデータ（ドキュメントデータ）の意味を文脈に応じて動的に解釈する方式を実現している．この機構により，データ間の意味的な関係を，与えられた文脈に応じて動的に計算することを可能としている．

提案方式では，各ドキュメントデータにメタデータとして複数の単語群が付与されていることを前提とする．このメタデータは，自動的または半自動的に各ドキュメントデータごとに付与される．このメタデータから自動生成された特徴つきベクトルを対象として，ドキュメントデータを対象とした意味的連想処理を実現している．さらに，このメタデータを用いることにより，意味的データマイニングを可能としている．

2. 文脈依存動的クラスタリングおよび意味的データマイニング方式の概要

本節では，本方式（文脈依存動的クラスタリングおよび意味的データマイニング方式）の概要を示す．本方式は，多数のドキュメントデータ群を対象とした分析者の文脈に応じた動的なクラスタリング分析を行う段階と，抽出されたクラスタを対象として各クラスタ内のドキュメントデータ群に共通する性質を知識として抽出する段階により構成する．前者を Phase-1，後者を Phase-2 とし，以下でその概要について示す．

Phase-1 ：文脈依存動的クラスタリング

多数のドキュメントデータ群を対象とした分析者の文脈に応じた動的なクラスタリング分析を行う．文脈に応じたデータの動的な意味的解釈については意味的連想処理機構[9),10),16)]を応用し，ドキュメントデータ間の意味的相関量を計算することにより文脈依存動的クラスタリングを実現する．

Phase-2 ：意味的データマイニング

Phase-1 により抽出されたクラスタを対象として各クラスタ内のドキュメントデータ群のメタデータに着目し，ドキュメント群を構成するメタデータを対象としてデータマイニングのアルゴリズムを適用し，共通する性質を知識として抽出する．各ドキュメントデータに付与されたメタデータは，分析対象となるドキュメントデータ群において表現形式について正規化されていることを前提とする．

2.1 Phase-1 の概要

Phase-1(文脈依存動的クラスタリング)は，次の4ステップにより実現される．

Step-1 ：正規直交空間の生成

分析対象アイテム群を特徴づける特徴量群を抽出し，正規直交空間を生成する．

Step-2 ：分析対象アイテム群の正規直交空間へのマッ

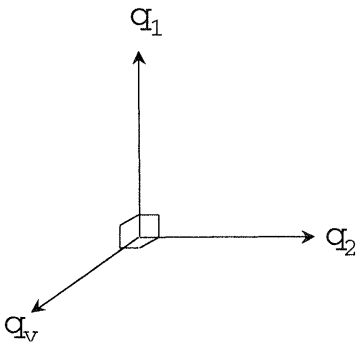


図1 Step-1: 正規直交空間の生成 (q_1 ~ q_v: 正規直交軸)
Fig. 1 Step-1: Creation of the orthogonal space (q_1 ~ q_v: orthogonal axes).

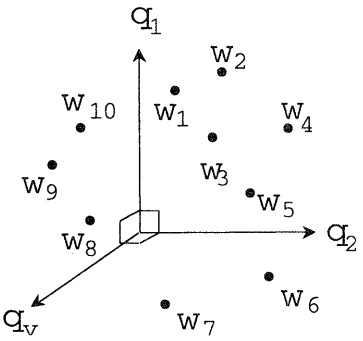


図2 Step-2: 分析対象アイテム群の正規直交空間へのマッピング (w_1 ~ w_{10}: 分析対象アイテム)
Fig. 2 Step-2: Mapping target items to the orthogonal space (w_1 ~ w_{10}: target items).

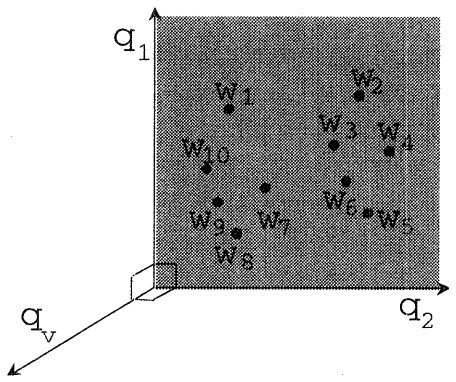


図3 Step-3: 問合せに応じた部分空間選択
Fig. 3 Step-3: Selection of subspace according to the given query.

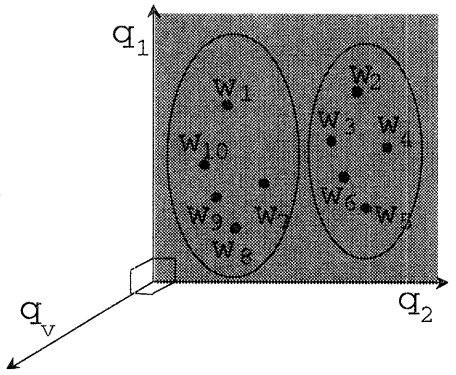


図4 Step-4: 部分空間上での分析対象アイテム群のクラスタリング
Fig. 4 Step-4: Clustering target items on the subspace.

ピング

分析対象アイテム群を抽出した特徴量群で特徴づけ，**Step-1** で生成した正規直交空間にマッピングする．

Step-3 ：問合せに応じた部分空間選択

意味的連想処理方式の応用により，分析者により与えられた問合せ（文脈語列）に応じて正規直交空間の部分空間選択を行う．

Step-4 ：部分空間上での分析対象アイテム群のクラスタリング

Step-3 で選択された正規直交空間の部分空間上において，距離が近いアイテム群を意味的に近いアイテム群として分析対象アイテム群をクラスタリングする．

2.1.1 Step-1: 正規直交空間の生成

まず，全ての分析対象アイテム群を特徴づけることができる特徴量群を抽出する．それを用いて，相関量を計算する場となる正規直交空間を生成する（図1）．

2.1.2 Step-2: 分析対象アイテム群の正規直交空間へのマッピング

全ての分析対象アイテム群を，前項で抽出した特徴量群で特徴づける．それを用いて，生成した正規直交空間に分析対象アイテム群をマッピングする（図2）．

2.1.3 Step-3: 問合せに応じた部分空間選択

意味的連想処理機構[9),10),16)]の特徴である部分空間選択の方式を用いて，分析者より文脈あるいは視点として与えられた問合せに応じて，生成した正規直交空間の部分空間を動的に選択する（図3）．全ての分析対象アイテム群は，選択された部分空間にマッピングされる．

2.1.4 Step-4: 部分空間上での分析対象アイテム群のクラスタリング

前項で選択された正規直交空間の部分空間上において，分析対象アイテム群をクラスタリングする（図4）．すなわち，文脈に応じた意味的解釈を伴う動的なクラスタリングを行う．この手続きにより，分析者の多様な視点または文脈に動的に対応することが可能である．

2.2 Phase-2 の概要

本方式における Phase-2(意味的データマイニング) の概要は，次の通りである．Phase-1 により得られたクラスタ群を対象に分析を行い，ドキュメントデータを対象としたデータマイニングを実現する．すなわち，生成された各クラスタを分析し，知識発見を自動的または半自動的に行う．具体的には，生成された各クラスタごとにおいて，各ドキュメントデータに付与されたメタデータを対象としてデータマイニングのアルゴリズムを適用し，クラスタを構成する分析対象アイテム群（ドキュメントデータ群）に共通する性質を知識として獲得する．

3. 文脈依存動的クラスタリング方式および意味的データマイニング方式の定式化

本節では，提案方式（文脈依存動的クラスタリング方式および意味的データマイニング方式）の定式化について述べる．

3.1 定 式 化

本節では，ユーザに与えられた文脈（具体的には，任意に与えられた ℓ 個の単語列により構成される文脈）に応じた動的なクラスタリングおよび意味的データマイニングの数学的定式化を示す．

各ドキュメントは複数個のメタデータで構成されているものとする．ここでメタデータは，単語であることを前提とする．

全メタデータの集合を $\mathcal{M}$ とし，その要素を md_i で表すものとする．メタデータの集合 $\mathcal{M}$ の全要素の数を m 個とする．すなわち $\mathcal{M}$ は次の通りである．

$$\mathcal{M} = \{md_1, md_2, ..., md_m\},\quad \#(\mathcal{M}) = m$$

ここで $\#(A)$ は，集合 A の要素数を表すものとする．

全ドキュメントの集合を $\mathcal{D}$ とし，その要素を doc_i で表すものとする．全ドキュメントの数を n とすれば，全ドキュメントの集合 $\mathcal{D}$ および各ドキュメント doc_i は下の通りである．

$$\mathcal{D} = \{doc_1, doc_2, ..., doc_n\}, \#(\mathcal{D}) = n$$
$$doc_i = \{md_{i_1}, md_{i_2}, ..., md_{i_k}\},$$
$$md_{i_j} \in \mathcal{M}, (j = 1, 2, ...k)$$

doc_i の要素数 k はドキュメントごとに（すなわち i に依存して）異なる．

1つのクラスタ Cl_i は，1つまたは複数個のドキュメントからなる．クラスタリングにより生成されるクラスタの組を $\mathcal{C}$ とする．また，各クラスタ Cl_i はドキュメントの集合のべき集合 $2^{\mathcal{D}}$，すなわち $\mathcal{D}$ の全ての部分集合の集合の要素となる．つまり下の通りである．

$$\mathcal{C} = \{Cl_1, Cl_2, ..., Cl_p\},\quad Cl_i \in 2^{\mathcal{D}}$$
$$Cl_i = \{doc_{i_1}, doc_{i_2}, ..., doc_{i_q}\},\quad doc_{i_j} \in \mathcal{D}$$

それぞれの集合の要素数 p, q は，下記のクラスタリング関数によって決定される．ただし，クラスタ数 p は，アルゴリズムによっては，分析者より与えられる場合もある．

クラスタリング関数 $f_c(\mathcal{D}; s_\ell)$ は，ドキュメントの全集合を，分析者から与えられた文脈 s_ℓ，即ち ℓ 個の文脈を規定する単語列に応じて，動的にいくつかのクラスタの組 $\mathcal{C}$ に分割する．クラスタリング関数は，$\mathcal{D}$ と S_ℓ の直積集合から，クラスタの組への写像である．つまり次のように定義される．

$$f_c : \mathcal{D} \otimes S_\ell \mapsto \mathcal{C}$$

一般にデータマイニングの場合，c_i, c_j を，ドキュメントに関する条件式とするとき，コンフィデンス (確信度) 関数 $confidence(c_i, c_j)$ は，以下の式で与えられる．

$$confidence(c_i, c_j) = \frac{\#\{doc \in \mathcal{D} \mid doc\ satisfies\ c_i \wedge doc\ satisfies\ c_j\}}{\#\{doc \in \mathcal{D} \mid doc\ satisfies\ c_i\}}$$

また通常のデータマイニングにおいては，上式の分母 $\#\{doc \in \mathcal{D} \mid doc\ satisfies\ c_i\}$ を全ドキュメント数 n でわったものは，$Support(c_i)$ と書かれ c_i のサポート（支持率）と呼ばれる．つまりサポート $Support(c_i)$ とは，全ドキュメントの内条件 c_i を満たすものの割合を表す．これに対応し，動的クラスタリングの場合，Cl_i に含まれるドキュメントの中で，そのドキュメント群のうちメタデータ md_j を含むものの割合を，そのクラスタとメタデータの関連性の確信度を計量するためのコンフィデンス関数 $conf(Cl_i, md_j)$ として用い，次の式により定義する．

$$conf(Cl_i, md_j) = \frac{\#\{doc \in \mathcal{D} \mid doc \in Cl_i \wedge md_j \in doc\}}{\#\{doc \in \mathcal{D} \mid doc \in Cl_i\}}$$

上式は，条件 c_i を $doc \in Cl_i$，条件 c_j を $md_j \in doc$ とした場合の $confidence(c_i, c_j)$ の式に対応する．

3.2 文脈依存動的クラスタリングの定式化

ここでは，意味的連想処理機構[9),10)] による文脈依存動的クラスタリングの定式化について述べる．すなわち，2節における Phase-1 についての詳細な定式化について述べる．

本方式では，次の3種類の特徴付ベクトル群が与えられていることを前提とする．第1は，イメージ空間を生成するための特徴付ベクトル群である．第2は，文脈語列（問合わせ）ための特徴付ベクトル群である．第3は，分析対象アイテム群に対応する特徴付ベクトル群である．これらのベクトル群は，各メタデータ (イメージ空間生成用メタデータ，文脈語列のためのメタデータ，

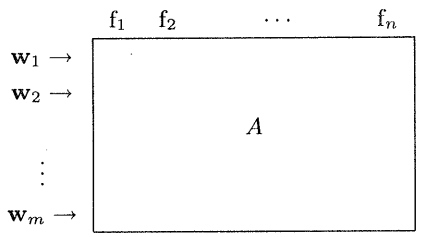


図5 データ行列 A の構成
Fig. 5 Data Matrix A.

そして，分析対象アイテム群のためのメタデータ) から自動生成されることを前提とする．

3.2.1 イメージ空間 $\mathcal{I}$ の設定

ここでは，m 個の単語について各々 n 個の特徴 $(f_1, f_2, \cdots, f_n)$ を列挙した各単語に対する特徴付ベクトル $\mathbf{w}_i (i = 1, \cdots, m)$ が与えられているものとし，そのベクトルを並べた m 行 n 列のデータ行列を A とする (図5).

(1) データ行列 A の相関行列 A^TA を作る．

(2) A^TA を固有値分解する．

$$A^TA = Q \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_\nu & & \\ & & & 0 & \\ & & & & \ddots 0 \end{pmatrix} Q^T,$$

$0 \leq \nu \leq n.$

ここで行列 Q は，

$$Q = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n)^T$$

である．この $\mathbf{q}_i$ は，相関行列の固有ベクトル，つまり意味素である．

(3) このとき，イメージ空間 $\mathcal{I}$ を以下のように定義する．

$$\mathcal{I} := span(\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_\nu).$$

$(\mathbf{q}_1, \cdots, \mathbf{q}_\nu)$ は $\mathcal{I}$ の正規直交基底である．

3.2.2 意味射影集合 Π_ν の設定

P_{λ_i} を次の様に定義する．

$P_{\lambda_i} \overset{\mathrm{d}}{\Longleftrightarrow} \lambda_i$ に対応する固有空間への射影，

i.e. $P_{\lambda_i} : \mathcal{I} \to span(\mathbf{q}_i)$.

意味射影の集合 Π_ν を次のように定義する．

$$\Pi_\nu := \{\, 0, P_{\lambda_1}, P_{\lambda_2}, \cdots, P_{\lambda_\nu}, P_{\lambda_1} + P_{\lambda_2}, P_{\lambda_1} + P_{\lambda_3}, \cdots, P_{\lambda_{\nu-1}} + P_{\lambda_\nu}, \vdots \; P_{\lambda_1} + P_{\lambda_2} + \cdots + P_{\lambda_\nu} \}.$$

Π_ν の要素の個数は 2^ν 個であり，これは 2^ν 通りの意味の様相表現ができることを示している．

3.2.3 意味解釈オペレータ S_p の構成（意味空間の選択）

文脈ベクトル

$$s_\ell = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell)$$

と，しきい値 $\varepsilon_s (0 \leq \varepsilon_s < 1)$ が与えられたとき，意味解釈オペレータ S_p は，その文脈ベクトル s_ℓ に応じて，意味射影 $P_{\varepsilon_s}(s_\ell)$ を決定する．すなわち，$s_\ell \in T_\ell$ （ここで T_ℓ は ℓ 語によって構成される語群シーケンスのすべての集合である．），$\Pi_\nu \ni P_{\varepsilon_s}(s_\ell)$ とすると，意味解釈オペレータ S_p は，T_ℓ から Π_ν への作用素として定義される．また，$\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell\}$ は，特徴付ベクトルであり，データ行列 A の特徴と同一の特徴を用いている．

オペレータ S_p は次のように定義される．

(1) $\mathbf{u}_i (i = 1, 2, \cdots, \ell)$ をフーリエ展開する．

$\mathbf{u}_i$ と $\mathbf{q}_j$ の内積を u_{ij} とする．

$$u_{ij} := (\mathbf{u}_i, \mathbf{q}_j), \; j = 1, 2, \cdots, \nu.$$

ベクトル $\widehat{\mathbf{u}}_i \in \mathcal{I}$ を次のように定める．

$$\widehat{\mathbf{u}}_i := (u_{i1}, u_{i2}, \cdots, u_{i\nu}).$$

これは，単語 $\mathbf{u}_i$ をイメージ空間 $\mathcal{I}$ に写像したものである．

(2) 文脈ベクトル s_ℓ の意味重心 $\mathbf{G}^+(s_\ell)$ を求める．

$$\mathbf{G}^+(s_\ell) := \frac{\left(\sum_{i=1}^{\ell} u_{i1}, \sum_{i=1}^{\ell} u_{i2}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right)}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \sum_{i=1}^{\ell} u_{i2}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty}$$

この $\| \cdot \|_\infty$ は，無限大ノルムを示す．

(3) 意味射影 $P_{\varepsilon_s}(s_\ell)$ を決定し，イメージ空間 $\mathcal{I}$ の部分空間（以下，意味空間とよぶ）を選択する．

$$P_{\varepsilon_s}(s_\ell) := \sum_{i \in \Lambda_{\varepsilon_s}} P_{\lambda_i} \quad \in \Pi_\nu.$$

但し $\Lambda_{\varepsilon_s} := \{\, i \mid \left(\mathbf{G}^+(s_\ell)\right)_i > \varepsilon_s \}$ とする．

3.2.4 意味空間における距離の定義

文脈語ベクトル s_ℓ が与えられたとする．また，分析対象アイテム x と分析対象アイテム y の特徴つきベクトルを，イメージ空間に写像したベクトルを $\mathbf{x} \in \mathcal{I}$, $\mathbf{y} \in \mathcal{I}$ とする．このデータ間の距離 $\rho(\mathbf{x}, \mathbf{y}; s_\ell)$ を次のように定める．

$$\rho(\mathbf{x}, \mathbf{y}; s_\ell) = \sqrt{\sum_{j \in \Lambda_{\varepsilon_s}} \{c_j(s_\ell)(x_j - y_j)\}^2},$$

ここで，$c_j(s_\ell)$ は，文脈ベクトル s_ℓ に依存して決まる重みであり，次のように定義する．

$$c_j(s_\ell) := \frac{\sum_{i=1}^{\ell} u_{ij}}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty}, \quad j \in \Lambda_{\varepsilon_s}.$$

このように，距離計算において，イメージ空間を構成する各意味素（固有ベクトル）に重みづけ $(c_j(s_\ell))$ を行うことにより，ε_s の値が小さい場合，すなわち，3.2.3節 (3) において意味空間を構成するために選択される固有ベクトルの数が多くなる場合においても，文脈の認識に関する ε_s の値の影響を小さくしている．

3.2.5 意味空間におけるクラスタリング方式

本方式は，文脈を反映した意味空間 (イメージ空間 $\mathcal{I}$ の部分空間) 上に分析対象アイテム群のマッピングを行った後に，それらのマッピングされたアイテム群を対象としたクラスタリングを行うことにより，文脈に応じた動的なクラスタリングを実現する方式である．

本方式の特徴の一つは，文脈に対応した意味空間内のクラスタリングのアルゴリズムを自由に選択できる点である．ここでは，得られた意味空間上で動的にクラスタリングを行う具体的方法として，以下の3関数を示す．これらは，3.1節のクラスタリング関数 f_c として適用される関数である．

関数 A_1: 意味空間における分析対象データ間の距離によるクラスタリング方式

関数 A_2: 意味空間における特定の軸上の値によるクラスタリング方式

関数 A_3: 意味空間における分析対象データの原点からの距離によるクラスタリング方式

関数 A_1: 意味空間における分析対象アイテム間の距離によるクラスタリング方法

関数 A_1 は，与えられた文脈に対応する意味空間 (イメージ空間 $\mathcal{I}$ の部分空間) において，分析対象アイテム間の意味的な距離によりクラスタリングを行う関数である．具体的には，すべての分析対象アイテム間の距離を求め，それによりクラスタ群を生成する．

ここでは，クラスタリング・アルゴリズムとして融合法[15] を例として採用する．融合法は，以下のように記述される．

(1) k 分析対象アイテムについて，全ての分析対象アイテムから全ての分析対象アイテムへの距離を求める．すなわち，3.2.4 節で定義した距離計算を $k(k-1)/2$ 回行う．

(2) k 分析対象アイテムを，k 個のクラスタとみなす．各々のクラスタは，意味空間上の座標として，各々のクラスタを構成する 1 つの分析対象アイテムの意味空間上の座標を持つ．

(3) 最小距離を持つ一組の分析対象アイテムを一つのクラスタとする．生成されたクラスタを意味空間上の 1 点で代表させる．

(4) (3) の操作を，分析対象アイテム群が指定された個数のクラスタになるまで繰り返す．

ここで，(3) における個々のクラスタを意味空間上の1点に代表させる方法については，次の5方法がある．

(a) クラスタを構成するある分析対象アイテムの座標を用いる．

(b) クラスタの重心の座標を用いる．

(c) クラスタ間の最小距離を求めるごとに，クラスタを構成する分析対象アイテムのうち，距離が最小となる分析対象アイテムの座標を用いる．

(d) クラスタ間の最小距離を求めるごとに，クラスタを構成する分析対象アイテムのうち，距離が最大となる分析対象アイテムの座標を用いる．

(e) クラスタ間の最小距離を求めるごとに，クラスタを構成する各分析対象アイテムのうち平均の距離となる座標を用いる．

関数 A_2: 意味空間における特定の軸上の値によるクラスタリング方式

関数 A_2 は，与えられた文脈に対応する意味空間 (イメージ空間 $\mathcal{I}$ の部分空間) において，各分析対象アイテムの特定の軸上の値よりクラスタリングを行う関数である．次の3ステップにより実現する．

Step-1: 意味空間における特定の軸を選択する．

Step-2: すべての分析対象アイテム群を選択された軸に射影する．

Step-3: すべての分析対象アイテム群間について選択された軸上における距離を求め，クラスタ群を生成する．

関数 A_3: 意味空間における分析対象データ間の原点からの距離によるクラスタリング方式

関数 A_3 は，与えられた文脈に対応する意味空間 (イメージ空間 $\mathcal{I}$ の部分空間) において，各分析対象アイテム群について原点からの距離を求め，その距離によりクラスタリングを行う関数である．クラスタ群の生成については，関数 A_1，関数 A_2 と同様である．

3.2.6 意味的データマイニングの定式化

本節では，意味的データマイニングの定式化，すなわち，2節における Phase-2 についての詳細な定式化について述べる．

本方式の Phase-1 において得られたクラスタを対象として，ドキュメントデータ群を説明するメタデータを対象としてデータマイニングの手法を適用することによって，文脈を反映した各クラスタを構成しているドキュメントデータ群に共通する意味を知識として抽出する．

知識の抽出の具体的方法として，分析対象アイテム群のメタデータを用いる以下の2種類の関数を示す．

関数 B_1: クラスタのIDと分析対象アイテム群のメタデータを対象に相関ルールアルゴリズム[1),2)]を適用する．具体的には，クラスタIDおよびメタデータの2属性を持つデータベースを対象に相関ルールアルゴリズムを適用する．これにより，クラスタIDとクラスタに含まれるドキュメントデータのメタデータとの相関ルールが抽出できる．すなわち，各クラスタごとに，含まれるドキュメントのメタデータのうち出現頻度の高いメタデータが知識として抽出できる．

関数 B_2: 各クラスタについて，分析対象アイテム群のメタデータを対象にアプリオリアルゴリズム[1),2)]を適用する．具体的には，各クラスタごとに，次の手続きを行う．クラスタに含まれるドキュメントのメタデータ群をひとつの集合と考える．このとき，この集合には，注目しているクラスタに含まれるドキュメントを説明する全てのメタデータが含まれている．この集合から，任意の組み合わせのメタデータについて出現頻度を求める．求めたメタデータの組と出現頻度のうち，出現頻度の高いメタデータを知識として採用する．これにより，各クラスタごとの意味的な概要を検索者や分析者に与えることが可能となる．

4. 実　　験

本節では，ドキュメントデータを対象とした実験により，提案方式である文脈依存動的クラスタリングおよび意味的データマイニング方式の実現可能性および有効性について検証する．

4.1 実験環境

医療分野のドキュメントデータを対象に実験を行った．本方式における，意味空間上での距離計算に用いられる各メタデータについては，本節に示す方法によって生成した．

4.1.1 イメージ空間生成用のメタデータの生成

医療分野を説明するに十分な単語である316単語を特徴語群（feature words）として用意した．医療分野において部位，症状，病名を表す1,048単語を，空間生成用メタデータの単語群（meta words）として用意した．

次の操作を行うことにより，3.2.1節におけるイメージ空間の作成に使用するデータ行列 A を生成した．空間生成用メタデータの単語（meta words 1,048語）について，各単語の説明語としてfeature wordsを用いて説明し，1,048行316列の行列 A を作成した．その単語群（meta words）を説明するfeature wordsが肯定の意味に用いられていた場合“1”，否定の場合“-1”，使用されていない場合“0”とし，見出し語自身が特徴である場合その特徴の要素を“1”として自動生成する．その操作後に，列ごとに2ノルムで正規化する．

```
doc101: がん 肺がん 肺 リンパ節
doc102: がん 肺がん 肺 腰椎 しびれ ぎっくり腰
doc103: がん 胃がん 早期がん 胃 吐血 下血
doc201: 胃 胃がん がん 食道 痛み 異物感 ポリープ
        早期がん 消化器
doc202: 胃 胃がん がん 胃かいよう 吐血 ポリープ
        粘膜 消化器
doc203: 胃 胃がん がん 胃かいよう 粘膜 胃壁 早期がん
doc501: 心臓病 心臓 不整脈 発作 疲れ ストレス
        意識不明 心臓疾患 心室
doc502: 心臓病 心臓 心筋梗塞 虚血性心疾患
        高脂血症 糖尿病 高血圧 高尿酸血症
        動脈硬化
doc503: 心臓病 心臓 心筋梗塞 血栓
        虚血性心疾患 動脈硬化 狭心症 ストレス
        血小板
```

図6 実験に使用したメタデータの例
Fig. 6 Examples of metadata for experiments.

3.2.1節における固有値分解の際の固有値の数，すなわちイメージ空間の次元数は，270であった．

4.1.2 分析対象アイテム群のメタデータの生成

イメージ空間へ写像する分析対象アイテム群のメタデータ生成については，医療分野の95ドキュメントデータを用いた．このドキュメントデータ群は，新聞記事の連載記事群である．95の各ドキュメントデータに対し，メタデータとして複数のmeta wordsを半自動的に付与した．具体的には，次の手順によりメタデータを付与した．まず，各ドキュメントデータから形態素解析などの技術を用い各ドキュメントに含まれる単語群を自動抽出した．次に，95のドキュメントデータ全てについて，各ドキュメントデータに対応する単語群から不要な単語や全ドキュメント群中で一貫していない単語を排除した．さらに，各ドキュメントデータについて，自動抽出されずかつ重要と思われる単語もメタデータとして加えた．以上の手順で，各ドキュメントデータに複数の単語(meta words)を付与した．

このメタデータの一部を図6に示す．ここで，ドキュメントデータのIDを「docXYY」という形式とした．Xは新聞記事の連載の種類を示し，YYは連載内のシリアルナンバである．連載の種類Xは，互いに番号が近いほど近い内容の連載であることを表している．ただし，XがAのとき連載の番号は10，XがBのとき連載の番号は11であることを示している．

4.1.3 文脈語列（問合わせ）メタデータの生成

イメージ空間へ写像する文脈語列のメタデータを，次

のように生成した．医療分野において部位，症状，病名を表す1,048単語を，空間生成用メタデータの単語（meta words）として用意した．meta wordsについて，feature wordsにより，空間生成用メタデータと同様に特徴づけを行った．

具体的には，空間生成用メタデータの単語（meta words 1,048語）を文脈語列メタデータとして用いた．各単語の説明語としてfeature wordsを用いて説明した．

4.1.4 クラスタリングのアルゴリズム

意味空間上でのクラスタリングのアルゴリズムは，3.2.5節で述べた．意味空間におけるクラスタリング方式については，関数A_1からA_3は共通する性質を持っていると考えられるので，関数A_1を採用して実験を行った．

3.2.6節で述べた意味的ドキュメントマイニングにおける分析方式の関数については，意味的には関数B_2における第1段階が関数B_1に相当し，関数B_2は関数B_1を含んでいると考えられるので，関数B_2を採用した．

ここで，個々のクラスタを意味空間上の1点に代表させる方法については，そのクラスタの重心の座標を用いる方法を採用した．この方法については，3.2.5節の(b)として述べた．(c)(d)および(e)と比較して，(b)は計算量が比較的少なく，さらに，(a)と比較して，(b)はそのクラスタを代表させるために最も客観的な方法であると考えられる．

4.1.5 実験システム

3節で述べた提案方式により実験システムを構築した．

4.2 実験方法

分析対象アイテム群に様々な文脈語列（問合わせ）を与え，分析結果を得る．分析対象アイテム群から文脈に応じた解析結果が得られることを確認する．さらに，本方式による分析結果から，有効な知識を抽出する．実験Aおよび実験Bの2種類の実験を行う．

実験Aの目的は，本方式によるクラスタリングの文脈による変化を検証することである．具体的には，同一分析対象セットを対象として3種類の異なる文脈を与え，文脈に依存してクラスタリングの結果が変化する様子を確認する．さらに，抽出されたクラスタから提案方式によって知識を抽出し，文脈によって変化するクラスタの特徴が発見できることを示す．

実験Bの目的は，単純なクラスタリング方式と提案方式によるクラスタリング方式の比較により提案方式の有効性を示すことである．比較的単純なベクトル空間モデルによるクラスタリング方式を適用した場合と提案方式

```
cluster 0:
doc101 doc102 doc103 doc104 doc105 doc106
doc107 doc108 doc109 doc110 doc111 doc201
doc202 doc203 doc204 doc205 doc206 doc207
doc208 doc209 doc210 doc301 doc302 doc303
doc304 doc305 doc306 doc307 doc308 doc309
doc310 doc311 doc401 doc402 doc403 doc404
doc505 doc509 doc906 doc909 doc910
cluster 1:
doc406 doc407 doc409 doc501 doc502 doc503
doc504 doc506 doc507 doc508 doc510 doc511
doc512 doc701 doc703 doc705 doc707 doc708
doc709 doc810 doc903 doc904 doc908 docA09
docA11 docB02 docB05 docB06 docB07 docB08
cluster 2:
doc601 doc704 doc811 doc901 doc907 docA01
docA02 docA04 docA05 docA06 docA07 docA12
docB03 docB04 docB09
```

図7 文脈“ストレス，不安”の文脈依存動的クラスタリングの結果（部分）

Fig. 7 Result-1 of context dependent dynamic clustering.

```
cluster 0:
L1:
        37      がん
        17      肺がん
        13      肺
        13      早期がん
        13      胃がん
L2:
        17      がん，肺がん
        13      がん，肺
        13      肺，肺がん
        13      がん，早期がん
========================================
cluster 1:
L1:
        13      心臓
        12      心臓病
L2:
        11      心臓，心臓病
========================================
cluster 2:
L1:
        7       糖尿病
L2:
```

図8 文脈“ストレス，不安”の意味的データマイニングの結果（部分）

Fig. 8 Result-1 of semantic data mining.

を適用した場合を比較し，提案方式が有効なクラスタリング方式であることを示す．

4.2.1 実験A

文脈語列（問合わせ）には，次の3種類を与えた．“ストレス，不安”，“疲れ，疲労”そして“疲労，五十肩”である．このうち，“ストレス，不安”と“疲れ，疲労”という文脈は互いに意味的に大きく異なるが，“疲れ，疲労”と“疲労，五十肩”とは互いに意味的に類

```
cluster 0:
doc101 doc102 doc103 doc104 doc105 doc106
doc107 doc108 doc109 doc110 doc111 doc202
doc203 doc204 doc205 doc206 doc207 doc209
doc210 doc301 doc302 doc303 doc304 doc305
doc306 doc307 doc308 doc309 doc310 doc311
doc401 doc402 doc403 doc404 doc909 doc910
cluster 1:
doc201 doc208 doc505 doc509
cluster 2:
doc406 doc407 doc409 doc501 doc502 doc503
doc504 doc506 doc507 doc508 doc510 doc511
doc512 doc601 doc701 doc703 doc704 doc705
doc707 doc708 doc709 doc810 doc812 doc903
doc904 doc907 doc908 docA01 docA02 docA04
docA05 docA06 docA07 docA09 docA10 docA11
docA12 docB01 docB02 docB03 docB05 docB07
docB08 docB09
```

図 9 文脈“疲れ，疲労”の文脈依存動的クラスタリングの結果（部分）
Fig. 9 Result-2 of context dependent dynamic clustering.

似している文脈である．

また，クラスタリングの際のパラメータとして，クラスタ数を 10 に設定した．これは，分析対象アイテム群の数の約 1/10 の数である．すなわち，1 クラスタを構成する分析対象アイテムの数は平均約 10 である．

本方式の分析に用いているアプリオリアルゴリズムは，多数のデータを対象としてデータの組（セット）の出現頻度が最小支持度を越える組をルールとして採用する方式である．本実験では，各クラスタに含まれるドキュメントデータ群のメタデータを分析の対象とし，組を構成する要素の数は 2 つまでとしてアプリオリアルゴリズムを適用した．最小支持度は 30 %とした．

4.2.2 実 験 B

単純なベクトル空間モデルによるクラスタリング方式と提案方式によるクラスタリング方式を比較する．前者の実験環境として，提案方式の実験システムから文脈理解機能を取り除いた実験環境を構築した．これは，単純なベクトル空間モデルによるクラスタリング方式の実験システムに相当する．

この実験システムを対象として，3 種類の文脈“ストレス，不安”，“疲れ，疲労”そして“疲労，五十肩”を与え，他の条件も実験 A の際の条件と同一にし，文脈依存動的クラスタリングの結果と意味的データマイニングの結果を実験 A の結果と比較する．

4.3 実験結果

4.3.1 実験 A の結果

実験結果は次の通りである．文脈語列“ストレス，不安”に対応する実験結果は，図 7および図 8である．図 7はクラスタリングの結果であり，図 8はそれを分析した結果である．すなわち図 7は“ストレス，不安”という文脈において文脈依存動的クラスタリング方式 (Phase-1) を適用した結果であり，図 8は意味的ドキュメントマイニング方式を適用した結果である．

```
cluster 0:
L1:
        34      がん
        17      肺がん
        13      肺
        12      早期がん
L2:
        17      がん，肺がん
        13      がん，肺
        13      肺，肺がん
        12      がん，早期がん
==========================================
cluster 1:
L1:
        3       生活習慣病
        2       心臓病
        2       がん
        2       胃
        2       冠動脈疾患
        2       胃がん
        2       心臓

L2:
        2       胃，胃がん
        2       がん，胃
        2       心臓，心臓病
        2       心臓，生活習慣病
        2       がん，胃がん
        2       心臓病，生活習慣病
==========================================
cluster 2:
L1:
        14      糖尿病
L2:
```

図 10 文脈“疲れ，疲労”の意味的データマイニングの結果（部分）
Fig. 10 Result-2 of semantic data mining.

同様に，文脈語列“疲れ，疲労”に対応する実験結果は，図 9および図 10である．図 9は文脈依存動的クラスタリングの結果であり，図 10は意味的データマイニングの結果である．

文脈語列“疲労，五十肩”に対応する実験結果は，図 11および図 12である．図 11は文脈依存動的クラスタリングの結果であり，図 12は意味的データマイニングの結果である．

図 8，図 10，および図 12において，L1 ではメタデータの出現頻度（出現回数）とそのメタデータを示し，L2 では 2 つのメタデータを組として算出した出現頻度（出現回数）とそのメタデータの組を示している．

4.3.2 実験 B の結果

実験 B の結果は文脈に依存しないクラスタリングであ

```
cluster 0:
doc101 doc102 doc103 doc104 doc105 doc106
doc107 doc108 doc109 doc110 doc111 doc202
doc203 doc204 doc205 doc206 doc207 doc209
doc210 doc301 doc302 doc303 doc304 doc305
doc306 doc307 doc308 doc309 doc310 doc311
doc401 doc402 doc403 doc404 doc909 doc910
cluster 1:
doc201 doc208 doc505 doc509
cluster 2:
doc406 doc407 doc409 doc501 doc502 doc503
doc504 doc506 doc507 doc508 doc510 doc511
doc512 doc601 doc701 doc703 doc704 doc705
doc707 doc708 doc709 doc810 doc903 doc904
doc907 doc908 docA01 docA02 docA04 docA05
docA06 docA07 docA09 docA10 docA11 docA12
docB02 docB03 docB05 docB07 docB08 docB09
```

図11 文脈“疲労，五十肩”の文脈依存動的クラスタリングの結果（部分）
Fig. 11 Result-3 of context dependent dynamic clustering.

り，同一のクラスタが結果として得られる．3種類の異なる文脈を与えた場合について，そのクラスタの構成を図13および図14に示す．

4.4 考　　察

実験結果Aより，互いに意味的に相関の強い単語が文脈に依存して同一クラスタを構成していることが確認できる．意味的に類似しているドキュメント群（メタデータが類似しているドキュメント群）が同一クラスタに含まれている．4.1.2節で示した通り，ドキュメントデータのIDの形式「docXYY」のうち，Xは新聞記事の連載の種類を示し，YYは連載内のシリアルナンバを示している．さらに，連載の種類Xは，互いに番号が近いほど近い内容の連載であることを示している．メタデータの例は，図6に示している．以上から，図6，7，9，11より，IDが互いに近いドキュメントデータが同一のクラスタを形成していることが分かる．

本方式のPhase-2(意味的データマイニング)の適用により，各クラスタの意味が知識として抽出されている．図8において「cluster-0」についてL1で示される知識により，このクラスタは「がん」のクラスタであることを抽出できる．また，L2で示される知識により，このクラスタ(cluster-0)は，「がん」および「肺がん」に相関の強いクラスタであることが抽出できる．これにより，分析対象のドキュメントデータ群を概観することができる．得られた知識により，膨大な数のドキュメントデータ群を対象として，利用者や分析者にブラウジングを行う順序やドキュメントデータ群の概要を指針として与えることが可能となる．

与えた文脈（文脈語列）により，クラスタ構成の様子

```
cluster 0:
L1:
        34      がん
        17      肺がん
        13      肺
        12      早期がん
L2:
        17      がん，肺がん
        13      がん，肺
        13      肺，肺がん
        12      がん，早期がん
========================================
cluster 1:
L1:
        3       生活習慣病
        2       心臓病
        2       がん
        2       胃
        2       冠動脈疾患
        2       胃がん
        2       心臓
L2:
        2       胃，胃がん
        2       がん，胃
        2       心臓，心臓病
        2       心臓，生活習慣病
        2       がん，胃がん
        2       心臓病，生活習慣病
========================================
cluster 2:
L1:
        13      糖尿病
L2:
```

図12 文脈“疲労，五十肩”の意味的データマイニングの結果（部分）
Fig. 12 Result-3 of semantic data mining.

が変化しているのが確認できる．実験Aにおける文脈“ストレス，不安”に対応する結果と文脈“疲れ，疲労”に対応する結果を比較すると，文脈に依存して変化するクラスタと変化しないクラスタが存在することが分かる．図7と図9との比較において，cluster-0は文脈に依存しないクラスタであることが分かる．これは，分析対象である95ドキュメントデータのうち「がん」についてのドキュメントデータが多く，それらがcluster-0を形成していると考えられる．さらに，cluster-1とcluster-2を比較すると，図7ではそれぞれ心臓病と糖尿病のクラスタであるが，図9においてはそれぞれ生活習慣病と糖尿病のクラスタであることが分かる．具体的には，文脈“ストレス，不安”を与えた場合のcluster-2の意味は，L1として抽出されている知識より「糖尿病」であると理解できる．一方，文脈“疲れ，疲労”を与えた場合のcluster-1の意味は同様に「生活習慣病」であり，cluster-2の意味は同様に「糖尿病」であると分かる．つまり，“ストレス，不安”という文脈では糖

```
cluster 0:
doc101 doc102 doc103 doc104 doc105 doc106
doc107 doc108 doc109 doc110 doc111 doc201
doc202 doc203 doc204 doc205 doc206 doc207
doc208 doc209 doc210 doc301 doc302 doc303
doc304 doc305 doc306 doc307 doc308 doc309
doc310 doc311 doc401 doc402 doc403 doc404
doc505 doc509 doc909 doc910 doc911
cluster 1:
doc406 doc407 doc409 doc501 doc502 doc503
doc504 doc506 doc507 doc508 doc510 doc511
doc512 doc601 doc701 doc703 doc704 doc705
doc707 doc708 doc709 doc810 doc812 doc901
doc902 doc903 doc905 doc908 docA06 docA07
docA12 docB01 docB02 docB03 docB05 docB06
docB07 docB08
cluster 2:
doc811 doc907
```

図 13 文脈“ストレス，不安”，“疲れ，疲労”，“疲労，五十肩”のクラスタリングの結果（部分）

Fig. 13 The result of clustering for the context-1,2 and 3.

尿病のドキュメントデータが集約されるが，“疲れ，疲労”という視点に立った場合にはより一般的な生活習慣病に関するドキュメントデータと，糖尿病のドキュメントデータが別のクラスタとして形成されることが発見できる．文脈が“疲れ，疲労”の場合は文脈が“ストレス，不安”の時よりも多くの糖尿病のドキュメントデータが集約されていることも発見できる．

さらに，実験Aからは，異なる2つの文脈に対応する結果は異なる結果であるが類似する2つの文脈に対応する結果は酷似していることが観測できる．

すなわち，実験Aにおける文脈“疲れ，疲労”に対応する結果と，文脈“疲労，五十肩”に対応する結果を比較すると，文脈依存動的クラスタリングの結果も意味的データマイニングの結果も酷似していることが分かる．図9と図11とを比較すると，cluster-0 および cluster-1 は全く同一であり，cluster-2 についても後者の結果において2ドキュメントデータが別のクラスタに属してしまったことが確認できる．図10と図12も同様に結果は酷似している．

提案方式によるクラスタリングでは，分析対象であるドキュメントデータの特徴のうち，特に利用者の文脈と相関が強い部分に着目してクラスタリング処理を行っている．これにより文脈に応じたクラスタリング分析を可能としている．さらに，以上のように，クラスタ群の文脈に依存して変化する特徴を獲得できる．

実験Bより，提案方式がクラスタリング方式として有効であることが示せる．図7，図9，図11および図13を比較すると，提案方式によるクラスタリングの方が類似したドキュメントデータが同一クラスタに多く含まれる

```
cluster 0:
L1:
        37      がん
        17      肺がん
        13      肺
        13      早期がん
        13      胃がん
L2:
        17      がん, 肺がん
        13      がん, 肺
        13      肺, 肺がん
        13      がん, 早期がん
==========================================
cluster 1:
L1:
        13      心臓
        12      心臓病
L2:
==========================================
cluster 2:
L1:
        2       頭痛
        2       発熱
        1       高熱
        1       膠原病
        1       鼻炎
        1       微熱
        1       くしゃみ
        1       発しん
        1       ウイルス感染症
        1       肺炎
        1       腹痛
        1       鼻水
        1       かぜ
L2:
        2       頭痛, 発熱
        1       かぜ, 鼻炎
        1       発しん, 発熱
        1       肺炎, 腹痛
        1       発熱, 鼻水
        1       発しん, 膠原病
        1       かぜ, 発熱
        1       頭痛, 肺炎
        1       頭痛, 腹痛
        1       高熱, 発熱
        1       かぜ, 鼻水
        1       発熱, 膠原病
        1       発しん, 微熱
        1       高熱, 膠原病
        1       発熱, 微熱
        1       ウイルス感染症, 肺炎
        1       ウイルス感染症, 腹痛
        1       高熱, 微熱
        1       頭痛, 鼻炎
        1       肺炎, 発熱
        1       微熱, 膠原病
        1       ウイルス感染症, 頭痛
        1       肺炎, 膠原病
        1       腹痛, 膠原病
        1       頭痛, 鼻水
        1       肺炎, 微熱
        1       頭痛, 膠原病
        1       鼻炎, 鼻水
        1       ウイルス感染症, 発熱
        1       頭痛, 微熱
        1       高熱, 発しん
        1       くしゃみ, 頭痛
        1       ウイルス感染症, 高熱
        1       ウイルス感染症, 膠原病
        1       くしゃみ, 鼻炎
        1       ウイルス感染症, 微熱
        1       肺炎, 発しん
        1       くしゃみ, 発熱
        1       発しん, 腹痛
        1       くしゃみ, 鼻水
        1       かぜ, くしゃみ
        1       発熱, 腹痛
        1       頭痛, 発しん
        1       高熱, 肺炎
        1       高熱, 腹痛
        1       かぜ, 頭痛
        1       発熱, 鼻炎
        1       高熱, 頭痛
        1       微熱, 腹痛
        1       ウイルス感染症, 発しん
```

図 14 文脈“ストレス，不安”，“疲れ，疲労”，“疲労，五十肩”の意味的データマイニングの結果（部分）

Fig. 14 The result of semantic data mining for the context-1,2 and 3.

ことが分かる．図8，図10，図12および図14を比較すると，提案方式によるクラスタリングの方が有用な知識が抽出されていることが分かる．図13および図14が示している結果は，単純なベクトル計算モデルによるクラスタリング方式，またはパターンマッチングを用いたク

ラスタリング方式に相当すると考えられる．これらの方式と比較して本方式は，クラスタリング方式として有効であると考えられる．

処理時間については，以下のように考察できる．

提案方式のプロセスは，合計5ステップにより構成される(2節における Phase-1 の Step-1 から Step-4，および Phase-2)．このうち，分析時に動的に処理する必要があるのは Phase-1 の Step-3，Step-4，および Phase-2 の3ステップである．Phase-1 の Step-1 は，意味的解釈の事前に一度だけ実行すればよい．Phase-1 の Step-2 については，分析対象アイテムの更新時に一度だけ実行すればよい．

このとき，Phase-1 の Step-3，Step-4 および Step-5 は，分析者から文脈が与えらるごとに動的に処理する．Phase-1 の Step-4 では，処理時間は，意味的解釈のための数値計算の処理時間とクラスタリング処理の処理時間の合計である．Phase-1 の Step-5 では，データマイニングアルゴリズムを適用する処理時間が必要となる．

Phase-1 の Step-3 は，部分空間（意味空間）選択は分析対象アイテムの数 n とは無関係の計算量である．Phase-1 の Step-4 においては，この実験に用いた方法では，クラスタリング処理のための距離計算とクラスタリングアルゴリズムの処理時間それぞれについて $O(n^2)$ の計算量が必要となる．現在の実験では，Phase-1 の Step-3，Step-4 および Phase-2 の処理時間の合計は，$n = 95$ について実時間で 1.47 [sec.] である(5回の実験の平均値)．

以上より，これらの実験結果は，文脈に応じたドキュメントデータ群の動的なクラスタリングおよびデータマイニングが可能な本方式の実現可能性および有効性を示している．

5. 結　論

本論文では，データの意味的な解釈を伴うドキュメントマイニングを行うための文脈依存動的クラスタリングおよび意味的データマイニング方式を提案した．本方式は，文脈に依存した意味的な相関に応じた動的なクラスタリングを実現する点が特徴である．さらに，既存のクラスタリングのアルゴリズムを自由に組み合わせることが可能である点も特徴である．本方式により，分析対象のデータに対して，文脈に応じて動的に意味的分析結果を得ることが可能となる．ドキュメントデータ群を対象とした実験により，本方式の実現可能性および有効性を確認した．

本論文における主要な提案は文脈依存動的クラスタリングおよび意味的データマイニング方式であり，具体的なクラスタリング・アルゴリズムとして用いている方法は比較的単純な方法である．高速化アルゴリズムの採用については今後検討を行う．分析対象アイテム群の距離計算の計算方式，およびクラスタリング・アルゴリズムについては，意味的検索の高速化アルゴリズムの適用[14]，クラスタリング・アルゴリズムとしてよく用いられる k-means 方式[15] などにより高速化が可能である．Phase-2 についても同様に，8) などに示される高速化アルゴリズムにより高速化が実現可能であると考えられる．

今後は，本方式の高速化，メタデータ空間（正規直交空間）の生成方式についての検討[13]，分析対象アイテム群の特徴量抽出方式の確立，および，本方式の各種マルチメディアデータへの適用を行う予定である．

謝　辞

本論文で用いた実験データは，田畑裕秋氏をはじめとする読売新聞社のデータベース分野の方々との共同作成によるものです．ここに記して感謝致します．

参考文献

1) Agrawal, R., Imielinski, T., Swami, A.: "Mining Association Rules between Sets of Items in Large Databases," Proc. of ACM SIGMOD, pp.207-216, 1993.
2) Agrawal, R., and Srikant, R.: "Fast Algorithms for Mining Association Rules," Proc. of the 20th International Conference on Very Large Data Bases, pp.487-489, 1994.
3) Ankerst, M., Breunig, M., Kriegel, H.P., and Sander, J.: "OPTICS: Ordering Points To Identify the Clustering Structure," Proc. of the ACM SIGMOD Conf. on Management of Data, ACM, 1999.
4) Botafogo, A. R. and Shineiderman, B.: "Identifying Aggregates in Hypertext Structures," Proc. of the 3rd ACM Conference on Hypertext, ACM, pp.63-73, 1991.
5) Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. and Simoudis, E.: "Mining Business Databases," Communications of the ACM, Vol.39, No.11, pp. 41-48, Nov. 1996.
6) 江口浩之, 伊藤秀隆, 隈元昭, 金田彌吉: "漸次的に拡張されたクエリを用いた適応的文書クラスタリング法", 電子情報通信学会論文誌 (D-I), Vol. J82-D-I, No.1, pp.140-149 (1999).
7) 波多野賢治, 佐野綾一, 段一為, 田中克己: "自己組織化マップと検索エンジンを用いた Web 文書の分類ビュー機構," 情報処理学会論文誌: データベー

ス,Vol. 40, No. SIG3(TOD1), pp. 47-59, 1999.
8) 喜連川優: "データマイニングにおける相関ルール抽出技法," 人工知能学会誌, Vol. 12, No. 4, pp. 513-520, Jul. 1997.
9) 清木康, 金子昌史, 北川高嗣: "意味の数学モデルによる画像データベース探索方式とその学習機構," 電子情報通信学会論文誌,D-II,Vol.J79-D-II,No. 4,pp. 509-519, 1996.
10) Kiyoki, Y., Kitagawa, T. and Hayama, T.: "A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning," Multimedia Data Management – using metadata to integrate and apply digital media –, McGrawHill, A. Sheth and W. Klas(editors), Chapter 7, 1998.
11) Konopnicki, D. and Shmueli, O.: "Information Gathering in the World Wide Web: The W3QL Query Language and the W3QS System," ACM Transactions on Database Systems, Vol. 23, No. 4, pp. 369-410, Dec. 1998.
12) Lent, B., Agrawal, R., Srikant, R.: "Discovering Trends in Text Databases," Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), pp. 227-230, 1997.
13) 宮川祥子, 清木康: "特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式," 情報処理学会論文誌: データベース,Vol. 40, No. SIG5(TOD2), pp. 15-27, 1999.
14) 宮原隆行, 清木康, 北川高嗣: "意味の数学モデルによる意味的連想検索の高速化アルゴリズムとその実現方式," 情報処理学会論文誌, Vol.38, No.7, pp.1399-1411, 1997.
15) 塩谷實:"多変量解析概論," 朝倉書店, 1990.
16) 吉田尚史, 清木康, 北川高嗣: "意味的連想検索機能を持つメディア情報検索システムの実現方式," 情報処理学会論文誌, Vol. 39, No. 4, pp. 911-922, 1998.
17) 吉田尚史, 図子泰三, 清木康, 北川高嗣: "ドキュメントデータを対象とした意味的連想処理機構による動的クラスタリング方式," 情報処理学会研究報告, 99-DBS-118, pp.89-96, 1999.

(平成1999年 9 月22日受付)
(平成1999年12月21日採録)

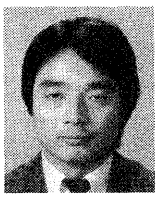
(担当編集委員 安達 淳)

吉田 尚史 (学生会員)

1972 年生. 1996 年筑波大学第三学群情報学類卒業. 1998 年同大学大学院理工学研究科修了. 現在, 同大学院工学研究科在学中. データベースシステム, マルチメディアシステムに関する研究に従事. ACM 会員.

図子 泰三

1976 年生. 1999 年慶應義塾大学環境情報学部卒業. データベースシステム, ドキュメントデータベースシステム, データマイニングシステムに関する研究に興味を持つ.



清木 康 (正会員)

1978 年 慶應義塾大学工学部電気工学科卒業. 1983 年 同大学院工学研究科博士課程了. 工学博士. 同年, 日本電信電話公社武蔵野電気通信研究所入所. 1984 年～ 1995 年筑波大学電子・情報工学系講師, 助教授を経て, 1996 年, 慶應義塾大学環境情報学部助教授, 1998 年同学部教授. データベースシステム, 知識ベースシステム, マルチメディアシステムの研究に従事. ACM, IEEE, 電子情報通信学会, 日本ソフトウェア科学会各会員.

北川 高嗣

1978 年名古屋大学工学部卒業. 1983 年同大学院工学研究科博士過程修了. 工学博士. スタンフォード大学計算機科学科客員研究員, 愛媛大学理学部数学科講師を経て 1990 年より筑波大学電子・情報工学系に勤務. 現在同学系助教授. 数値解析, 逆問題, マルチメディア情報システムの研究に従事. 日本応用数理学会会員.