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の再帰的適用による意味的知識発見方式

図子泰三[†] 吉田尚史[†] 清木康[††]

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の再帰的適用による意味的知識発見方式を提案する．本方式の特徴は，次の 2 点にまとめられる．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見を実現する点，および，共通の性質を有するより多くのドキュメントが含まれるクラスタの抽出を可能とする点である．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．応用分野として，医療ドキュメントデータ群を用いたシステム構築，および，実験結果を示し，本方式を適用したマイニングシステムの実現可能性および有効性を明らかにする．

A Semantic Knowledge Discovery Method by Recursively Applying Context Dependent Dynamic Clustering to Document Data

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

In this paper we present a semantic knowledge discovery method which recursively applies context dependent dynamic clustering to document data. The features of the method can be summerized in following two points. One is to perform knowledge discovery from document clusters which are dynamically partitioned according to a given context, and the other is to extract clusters including more documents with common features. By using this method, we can dynamically obtain a set of semantic clusters of documents according to a given context. We have implemented a system for a medical domument set as an application, and clarify feasibility and effectiveness of the mining system based on the method by showing several experimental results.

1. はじめに

近年，コンピュータネットワーク上に存在する多種多様なドキュメントデータを検索対象とするシステムの実現が行われてきている．それにともない，それらのドキュメントデータを対象とした情報獲得の機会は増大しており，ドキュメントデータ群を対象とした的確な情報獲得方式の実現が重要な課題となっている．それらの多様なドキュメント群を対象とした知識発見，データマイニングの実現が行われれば，単純な検索に比べ，より高度な情報獲得を行うことが可能となる．

データマイニングに関する研究[3),6)] 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[9)] が活発である．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．

我々は，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と考える．そして，多数のドキュメントデータを対象として，文脈や視点に応じたデータマイニングを実現する方式として，文脈依存動的クラスタリングを提案している[15),16),18)]．

本論文では，ドキュメントデータ群を対象として，文献 15)，18) において示した文脈依存動的クラスタリングを再帰的に適用する機能を加えた意味的知識発見方式を示す．また，その機能を加えた文脈依存動的クラスタリング方式を実際にシステムとして実現し，

† 慶應義塾大学政策・メディア研究科
Graduate School of Media and Governance, Keio University
†† 慶應義塾大学環境情報学部
Faculty of Environmental Information, Keio University

そのシステムの応用分野として医療ドキュメントデータベースを対象とした実現方式を示す．本論文は，再帰的クラスタリング機能を加えることにより文脈依存動的クラスタリング方式を拡張し，さらに，実践的な対象を扱う方法を示す論文として位置付けることができる．意味的知識発見方式は，文脈依存動的クラスタリングにより形成されたクラスタに含まれるドキュメントを構成するメタデータを対象として，データマイニングアルゴリズムとクラスタリングアルゴリズムを繰り返し再帰的に適用し，より詳細なサブクラスタを生成する方式である．ドキュメントを対象としたクラスタリングにおいては，共通の性質を有するより多くのドキュメントが含まれるクラスタを最適クラスタとする．その最適クラスタを発見するために，本方式では，再帰的に得られるクラスタの中間結果を含めて，最終クラスタ（分析者に出力として与えるクラスタ）の候補とし，より多くのドキュメントを含みかつ共通性（確信度によって表される）の高い性質を持つクラスタを抽出する．

再帰的にアルゴリズムを適用する方式については，データベースや情報検索などの分野において研究されている（たとえば文献 8)，12)）．それらとの比較において本方式の特徴は，再帰的に生成されるクラスタのすべての中間結果を対象として，より多くのドキュメントが含まれ，かつ，共通の性質を有するクラスタを抽出可能な点にある．この特徴を持つ本方式により，より多くドキュメントデータを含みかつ性質が集約されたクラスタ群を，最適なクラスタとして分析者に与えることが可能となる．本論文では，医療ドキュメントデータを対象としたシステム構築および実験により，本方式の実現可能性および有効性を示す．

本方式では，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している．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付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，意味的解釈をともなった知識発見を可能としている．

Dublin Core[4] や Resource DescriptionFramework（RDF）[10] といったデジタルドキュメントデータを対象としたメタデータの標準形式が提案されている．今後は，これらのメタデータが付与されたデジタルドキュメントデータが急増すると予想され，本方式の適用範囲も拡大されると考えられる．このような環境においては，クラスタリングの精度を評価する際も，メタデータの類似性を評価することが有効であると考えられる．

2. 関 連 研 究

本方式における文脈依存動的クラスタリング方式の文脈に応じた動的な意味的解釈については，意味の数学モデル[7),13)] における意味的連想処理機構を用いて実現している．この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．この方式は，部分空間上でのクラスタリングアルゴリズムには依存していないので，分析対象に応じたクラスタリングアルゴリズムを選択可能である．

クラスタリングをともなったデータ分析は，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[5),6)]．それらは，いずれも対象オブジェクト群の全体的な傾向を調べるための方式として位置付けることができる．しかし，データマイニングでは，データのおおまかな傾向よりも，詳細な，部分的に成立するルールやパターンに着目することが多い．本論文で提案する意味的知識発見方式は，対象オブジェクト空間に対して，意味的連想処理機構を適用し，分析者の問合せに応じた部分空間上でクラスタリングを行う方式である．この方式により，分析者が与える文脈に応じた意味的解釈をともなったデータ分析が可能となる．

3. 文脈依存動的クラスタリングの再帰的適用による意味的知識発見方式の概要

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される．前者を Phase-1，後者を Phase-2 とし，それらの概要について示す．さらに，Phase-1 は 5 つのステップ（Step-1〜Step-5）からなる．Phase-1（Step-1〜Step-4），Phase-2 に関しては，文献 18) において提案した．本論文では，新たに Phase-1 の Step-5 を追加する．この Step を追

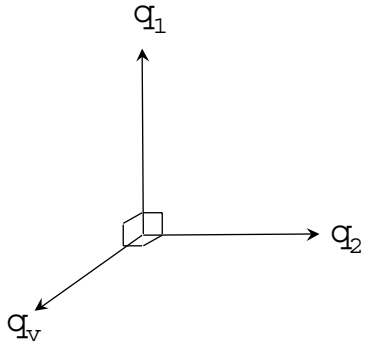

図 1 Step-1：正規直交空間の生成（q_1 ～ q_v：正規直交軸）
Fig. 1 Step-1: Creation of the orthogonal space (q_1 ～ q_v: orthogonal axes).

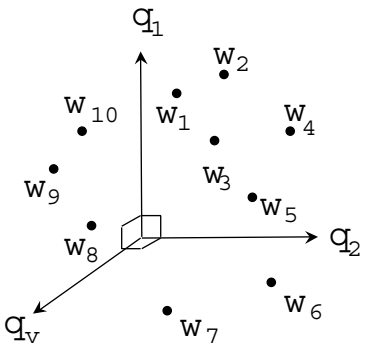

図 2 Step-2：分析対象アイテム群の正規直交空間へのマッピング（w_1 ～ w_{10}：分析対象アイテム）
Fig. 2 Step-2: Mapping target items to the orthogonal space (w_1 ～ w_{10}: target items).

加拡張することによって，文献 18) において提案した手法と比較して，より精度の高いクラスタ群の獲得が可能となる．

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[7),13)] 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．さらに，形成された各クラスタを対象として，ドキュメントデータ群を構成するメタデータの確信度を計算し，確信度の高いクラスタ群が形成されるまで，文脈依存動的クラスタリングを再帰的に適用する．また，生成されたクラスタ群のうち，多くのドキュメントを含み，かつ確信度の高いメタデータを有するクラスタを抽出する．

Phase-2： 意味的データマイニング方式
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．

3.1 Phase-1 の概要

Phase-1（文脈依存動的クラスタリング）は，次の 5 ステップにより実現される．

3.1.1 Step-1：正規直交空間の生成

まず，すべ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**）．

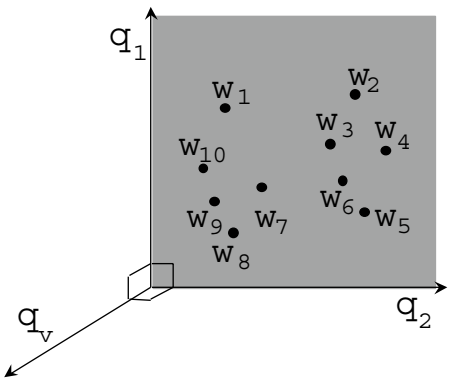

図 3 Step-3：問合せに応じた部分空間選択
Fig. 3 Step-3: Selection of subspace according to the given query.

3.1.2 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

すべ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**）．

3.1.3 Step-3：問合せに応じた部分空間選択

意味的連想処理機構[7),13)] 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**）．すべ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．ここでの文脈とは，具体的には，分析者が分析対象について，分析する事象を単語（群）として表現する．

3.1.4 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

3.1.3 項で選択された正規直交空間の部分空間上に

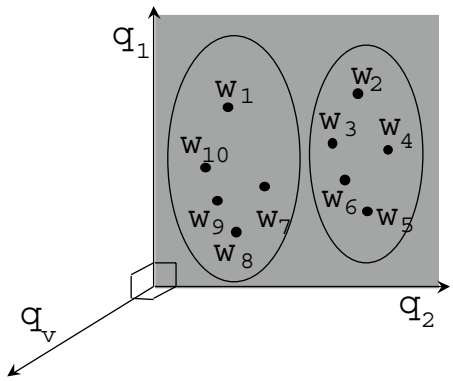

図 4 Step-4：部分空間上での分析対象アイテム群のクラスタリング
Fig. 4 Step-4: Clustering target items on the subspace.

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ともな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．

3.1.5 Step-5：メタデータを用いた再帰的クラスタリング

Step-5 は方式 A と方式 B の 2 方式によって構成される．方式 A は各クラスタ内で再帰的にクラスタリングを行うことによって，より共通的な性質を持ったドキュメント群を獲得する方式である．方式 B は，方式 A の再帰的にクラスタリングを行う過程において，クラスタ内のドキュメント数が多く，共通的な性質を有するクラスタを抽出する方式である．方式 A を適用した後に方式 B を適用することによって，ドキュメントの共通性と数のバランスのとれたクラスタ群の獲得が可能となる．方式 A のみを適用すると，確かに共通性の高いドキュメントが集まったクラスタが形成されるが，極端にドキュメントの数が少ないクラスタが形成される可能性がある．本研究では，ドキュメントの数が極端に少ないクラスタよりも，ある程度，ドキュメントの数がまとまったクラスタを有効なクラスタであると考える．

方式 A： 3.1.4 項で形成された各クラスタについて，各クラスタに含まれるそれぞれのメタデータの出現頻度に関する確信度☆を計算し，その中で最大の値を持つ最大確信度☆☆を求める．この最大確信度の値が，あらかじめ設定した閾値を超えない場合は，そのクラスタ内においてドキュメントデータを構成するメタデータを対象として，さらにクラスタリングを行う．この操作により，各サブクラスタ内のメタデータの出現に関する確信度が閾値を超えるまで，クラスタ内で再帰的にクラスタリングを行うことによって，元のクラスタよりも，より共通的な性質（共通性の高いメタデータ）を持ったドキュメント群から構成されるクラスタを形成することができる．ここでは，クラスタの確信度が高くなるほど，クラスタ内のドキュメントの共通性が高くなるということを示している．すなわち，閾値を変化させることによって，形成されるクラスタ内のドキュメントの共通性をコントロールすることができる．

方式 B： 方式 A によって最終的に生成されるクラスタは，含まれるドキュメント群が共通の性質を有しているが，含まれるドキュメント群の数が非常に少数になっている可能性がある．方式 B は，共通の性質を有し，少数のドキュメント群からなるクラスタよりも，共通性は低くなるが，ドキュメント数がより多いクラスタを獲得することを重視する方式である．具体的には，クラスタの最大確信度が閾値を超えるまで行った再帰的クラスタリング（方式 A）の中間結果を含むクラスタ群を対象として，クラスタの最大確信度とクラスタ内のドキュメント数から算出される評価値を求める．そして，この評価値が高い順にランキングすることによって，クラスタ内のドキュメント数が多く，かつ共通性の高い順にクラスタを獲得することができる．

3.2 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，次のとおりである．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．

4. 提案方式のアルゴリズム

本章では，本論文で提案する再帰的意味的クラスタ

☆ クラスタ内の各ドキュメントに付与されたメタデータの出現回数をクラスタ内の全ドキュメントの数で割った値である．
☆☆ ほかにも確信度の平均などが考えられるが，最も直観的な基準として最大確信度を採用している．一般的な確信度についての考察は，文献 11) に示されている．

リング方式（3 章における Phase-1 の Step-5）について示す．また，本方式（文脈依存動的クラスタリングを用いた意味的知識発見方式）の基礎となる 2 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）のアルゴリズムを示す．このアルゴリズムの数学的定式化の詳細は，文献 18) に示している．

4.1 再帰的意味的クラスタリング方式

本節では，本論文で提案する再帰的意味的クラスタリング方式のアルゴリズムについて示す．再帰的意味的クラスタリングは 3 章における Phase-1 の Step-5 に対応する部分であり，方式 A と方式 B の 2 方式によって構成される．

方式 A： 文脈依存動的クラスタリング方式と意味的データマイニング方式を交互に再帰的に適用し，意味的に近いメタデータを有するクラスタを自動生成する．方式 A のアルゴリズムを以下に示す．前提として，再帰の停止条件となる最小確信度（閾値）ε_c が与えられているとする．

手順 1 クラスタリング関数により文脈依存動的クラスタリングを適用し，クラスタ群を生成する．

手順 2 生成された各クラスタについて，確信度関数により意味的データマイニング方式を適用し，最大確信度を求める．

手順 3 クラスタの最大確信度が閾値 ε_c を超えない場合，そのクラスタを対象として，このアルゴリズム（手順 1〜3）を再帰的に適用する．

方式 B： 方式 A によって得られた結果（中間結果を含む）の中から，まとまった数のドキュメント群を有するクラスタを抽出する（少数のドキュメントを含むクラスタを排除する）．方式 B のアルゴリズムを以下に示す．

手順 1 方式 A によって生成された全クラスタに対して，クラスタに含まれるドキュメントの数を求める．

手順 2 各クラスタに対して，クラスタの最大確信度とドキュメントの数を使用し，評価関数によって評価値を算出する．

手順 3 クラスタ群を評価値の高い順にランキングする．

ここでの評価関数は，クラスタの最大確信度とドキュメント数のバランスを評価するために使用される評価値を算出する関数である．

4.2 文脈依存動的クラスタリング

本節では，文脈依存動的クラスタリング，すなわち，3 章における Phase-1 のアルゴリズムを示す．

4.2.1 意味的連想処理機構

本項では，意味の数学モデルによるドキュメントデータの意味的連想処理方式について概説する．この方式の詳細は，文献 7)，13) に示されている．ここでは，3 項における Phase-1 の Step-1 から Step-3 に対応する，文脈に応じた部分空間選択について述べる．

(1) **イメージ空間 $\mathcal{I}$ の設定：**
検索対象となるメディアデータをベクトルで表現したデータをマッピングするための正規直交空間（以下，メタデータ空間 $\mathcal{I}$）を設定する．

(2) **分析対象アイテムのマッピング：**
設定されたイメージ空間 $\mathcal{I}$ へ分析対象となるメディアデータのメタデータをベクトル化しマッピングする．これにより，同じ空間に検索対象データのメタデータがイメージ空間上に配置されることになり，検索対象データ間の意味的な関係を空間上での距離として計算することが可能となる．
メディアデータには，メタデータとして複数の単語が付与されていることを前提としている．各単語は，ベクトル表現されたデータを持っている．各メディアデータは，メタデータとして付与されている複数の単語が合成されベクトル表現された後，イメージ空間 $\mathcal{I}$ へマッピングされる．

(3) **$\mathcal{I}$ の部分空間（意味空間）の選択：**
分析者は与える文脈を複数の単語を用いて表現する．分析者が与える単語の集合を文脈語列と呼ぶ．この文脈語列を用いてイメージ空間 $\mathcal{I}$ に各文脈語に対応するベクトルをマッピングする．これらのベクトルは，イメージ空間 $\mathcal{I}$ において合成され，意味重心を表すベクトル生成される．意味重心から各軸への射影値を相関とし，閾値を超えた相関値（以下，重み）を持つ軸からなる部分空間（以下，意味空間）が選択される．

4.2.2 部分空間上での分析対象アイテム群のクラスタリング方式

本項では，部分空間上での分析対象アイテム群のクラスタリング方式，すなわち，3 章における Phase-1 の Step-4 に対応する部分の詳細について述べる．

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．

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．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．

4.3 意味的データマイニング方式

本節では，意味的データマイニング，すなわち，3章における Phase-2 についてのアルゴリズムを示す．

本方式の Phase-1 において得られた各クラスタについて，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．これによって，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関数を示す．

関数： 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 1 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ために付与されたすべてのメタデータが含まれている．この集合から，任意の組合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．

アプリオリアルゴリズムとは，データマイニングの一手法である相関ルール抽出の高速化の一方式であり，アイテム集合に対して，アイテム間の共起出現情報を抽出するのに効果的な方式である[2)]．アプリオリアルゴリズムを採用した理由は，ドキュメントに付与されているメタデータをアイテム集合としてアプリオリアルゴリズムを適用すれば，ドキュメントデータ間の共通性や，頻出するメタデータを調べることが可能であると判断したからである．

5. 実　　験

本章では，ドキュメントデータを対象とした実験により，本論文で示している文脈依存動的クラスタリングの再帰的適用による意味的知識発見方式の実現可能性および有効性について検証する．

5.1 実 験 環 境

ここではサンプルドキュメントとして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，本章に示す方法によって生成した．

5.1.1 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，4.2.1 項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” として生成する．ここでは，“0”，“1”，“-1” といった離散値によって特徴づけを行っているが，連続値による特徴づけも可能である．しかし，医療分野に関して，言葉と言葉の関係を連続値によって適切に表している指標が存在しないため，本実験では，“0”（関係なし），“1”（関係あり），“-1”（否定的な意味で関係あり）といった離散的な値による特徴づけを行った．

5.1.2 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．つまり，各ドキュメントデータに対し，5.1.1 項における 1,048 個のメタデータの単語群（meta words）を用いてメタデータを付与した（各ドキュメントのメタデータは，1,048 個の単語群のサブセットである）．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すべてについて，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

```
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 5 実験に使用したメタデータの例
Fig. 5 Examples of metadata for experiments.

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 5** 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．

5.1.3 文脈語列（問合せ）のメタデータの生成

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 words1,048 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．

5.1.4 実験システム

4 章で述べた方式により実験システムを構築した．4.2.2 項におけるクラスタリングアルゴリズムに関して，本実験では，融合法[14] を採用した．その理由としては，融合法はクラスタリングアルゴリズムの中で一般的なものであるので，クラスタリングアルゴリズムの影響に左右されず，本方式の精度を確認できるからである．再帰的意味的クラスタリングと意味的データマイニングのプログラムに関しては，Perl を用いて実装した．それ以外のプログラムに関しては，C 言語を用いて実装した．使用した計算機は Sun Enterprise 150，OS は SunOS 5.5.1 である．

```
cluster 0:
doc101 doc108 doc811 doc111 doc812 docA06
doc205 docA12 doc208 doc302 doc911 doc202
docA05 doc309 doc409 docB02 docA03 doc110
doc103 doc506 doc304 doc201 doc104

cluster 1:
doc105 doc204 doc910 doc512 doc904 docA01
doc401 doc508 doc306 doc107 doc402 doc303
doc305 doc310 doc403 doc901 doc810 doc307
doc308 doc209 doc406 docA11 doc504 doc509
doc507 doc510 doc902 docB01 doc407 doc905
docA09 doc206 doc502 doc908 doc903 doc511
doc505 docA07 doc601 docB09 doc501

cluster 2:
doc701 docB06 docB03 docB08 docB05 docB10
doc704 doc708 doc707 doc709 doc909 doc705
docB07 docB04 doc907
```

図 6 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. 6 Result-1 of context dependent dynamic clustering.

5.2 実 験 A

［方法・結果］

実験 A は，クラスタリングにおいて，再帰的適用を行わない文脈依存動的クラスタリング方式[18] に関する実験である．実験 A の目的は，文脈に応じてクラスタリングの結果が変化することを確認することである．具体的には，同一分析対象ドキュメントデータセットを対象として，2 種類の異なる文脈語列を与え，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，生成されたクラスタから知識を抽出し，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．

文脈語列には，次の 2 種類を与えた．“ストレス，不安”，“疲労，五十肩” である．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．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超える組をルールとして採用する方式である．本実験では，ドキュメントデータ群のメタデータを分析の対象とし，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．

文脈語列 “ストレス，不安” に対応するクラスタリング結果，分析結果は，それぞれ**図 6**，**図 7** である．同様に，文脈語列 “疲労，五十肩” に対応するクラスタリング結果，分析結果は，それぞれ**図 8**，**図 9** である．図 7 および図 9 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ている．

```
================================
cluster 0:
  13  がん
  7  糖尿病
  6  肺がん
  5  胃
  5  胃がん
================================
cluster 1:
  13  がん
  11  心臓
  10  心臓病
================================
cluster 2:
  7  疲れ
  4  過労
  4  ストレス
  3  自律神経
  3  過労死
================================
```

図 7 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. 7 Result-1 of semantic data mining.

```
cluster 0:
doc101 doc908 doc909 doc911 docA01 docA02
docA05 docA04 doc906 doc910 doc102 doc110
doc907 docA03 doc207 docB07 doc108 doc202
doc109 docB06 doc104 doc204 doc502 docB04
doc106 doc107 doc203 doc210 doc205 docB05
doc209 doc103 doc501 doc504 doc206 doc201
doc111 doc503 doc105

cluster 1:
doc208 doc601 doc509

cluster 2:
doc301 docB01 docA08 doc310 doc903 docB02
doc401 doc409 doc307 doc101 doc404 doc306
docA07 docA12 docB09 doc302
```

図 8 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. 8 Result-2 of context dependent dynamic clustering.

［考　　察］

クラスタリング結果において，互いに意味的に類似しているドキュメント群が同一のクラスタに含まれていることが確認できる．5.1.2 項で示したとおり，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．以上から，図 6，図 8 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．

また，与えた 2 種類の文脈によるクラスタリング結果および意味的データマイニング結果により，文脈に

```
====================
cluster 0:
  36  がん
  17  肺がん
  13  肺
  13  早期がん
  12  胃がん
  11  胃
====================
cluster 1:
  3  生活習慣病
  2  冠動脈疾患
  2  心臓病
  2  心臓
====================
cluster 2:
  5  疲れ
  4  頭痛
  2  発熱
  2  微熱
  2  慢性疲労
  2  うつ症状
  2  筋肉
  2  脳しゅよう
  2  自律神経
  2  過労
  2  不眠
====================
```

図 9 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. 9 Result-2 of semantic data mining.

応じてクラスタ構成の様子が変化していることが確認できる．図 7，図 9 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．cluster-0 にクラスタに依存しないクラスタが生成され,「がん」に関するドキュメントデータによって形成されている．その他のクラスタについては，文脈に依存して変化するクラスタが生成されている．特に，cluster-3 には，それぞれ文脈と相関の強いクラスタが生成されていることが分かる．

5.3 実　験　B

［方法・結果］

実験 B は，本論文で新たに提案する文脈依存動的クラスタリングの再帰的適用（4.1 節の方式 A）に関する実験である．実験 B の目的は，実験 A で得られたクラスタ群と比較して，さらに洗練されたクラスタ群を生成することである．具体的には，文脈に応じたクラスタリングによって生成される各クラスタを対象として，さらに，再帰的にサブクラスタを生成する．また，生成されたサブクラスタから提案方式によって知識を抽出し，より詳しいクラスタの特徴が発見できることを示す．

実験 A の文脈語列 “ストレス，不安” に対するクラ

スタリング結果（図 6）の各クラスタについて，クラスタの確信度が 4.1 節における最小確信度 ε_c（閾値）を超えないものに関して，さらに 2 つのサブクラスタに分割する．このサブクラスタに分割する操作を，すべてのクラスタが最小確信度 ε_c（閾値）を超えるようになるまで再帰的に行った．

ここでの確信度は，クラスタ内のすべてのドキュメントデータに付与されているメタデータの集合の中で，最も出現頻度の高いメタデータの出現回数を，クラスタ内のドキュメントの数で割った値である．本実験では，最小確信度 ε_c（閾値）を 40%とした．

実験 A と同様に，生成された各クラスタを対象としてアプリオリアルゴリズムを適用し，分析を行った．最小支持度は 30%である．

得られたクラスタリング結果，分析結果を，それぞれ**図 10**，**図 11** に示す．

［考　　察］

クラスタリング結果において，互いに意味的に類似しているドキュメント群が同一のサブクラスタに含まれていることが確認できる．図 10 は，クラスタごとに，クラスタに含まれる文書群と，最大の出現頻度を持つメタデータ（括弧内はその確信度の値）を示している．

実験 B の結果が，実験 A における図 6 と異なる点は，最大の出現頻度を持つメタデータの確信度の値が，あらかじめ設定した最小確信度（40%=0.4）を満たないクラスタ，つまり cluster 1 について，さらにサブクラスタ（cluster 1-0，cluster 1-1）に分割していることである．また，cluster 1-0 においても，同様の条件を満たさないため，さらに小さなサブクラスタ（cluster 1-0-0，cluster 1-0-1）に分割している．つまり，最終的には，最小確信度を満たしている 5 つのクラスタ（cluster 0，cluster 1-0-0，cluster 1-0-1，cluster 1-1，cluster 2）が結果として得られたことになる．このことから，クラスタリングが再帰的に機能していることを確認できる．

さらに，意味的データマイニング結果により，サブクラスタごとに異なった概念の単語情報が記述されていることが確認できる．具体的には，図 11 より，cluster 1-0-0，cluster 1-0-1 はそれぞれ，「がん」，「心臓」に関するサブクラスタであることが分かる．このことは，元のクラスタは「がん」と「心臓」という 2 つの概念が混ざったクラスタであったのが，サブクラスタに分けることによって 2 つの概念ごとのクラスタに分割されたということを示している．つまり，サブクラスタに分ける前の状態では得ることのできなかった，明確な情報を知識として自動的に獲得できることを示している．

```
========================================
cluster 0:
doc101 doc108 doc811 doc111 doc812 docA06
doc205 docA12 doc208 doc302 doc911 doc202
docA05 doc309 doc409 docB02 docA03 doc110
doc103 doc506 doc304 doc201 doc104
最大確信度をもつメタデータ：がん（0.61）
========================================
cluster 1:
doc105 doc204 doc910 doc512 doc904 docA01
doc401 doc508 doc306 doc107 doc402 doc303
doc305 doc310 doc403 doc901 doc810 doc307
doc308 doc209 doc406 docA11 doc504 doc509
doc507 doc510 doc902 docB01 doc407 doc905
docA09 doc206 doc502 doc908 doc903 doc511
doc505 docA07 doc601 docB09 doc501
最大確信度をもつメタデータ：がん（0.32）
    ========================================
    cluster 1-0:
    doc105 doc204 doc403 doc406 doc504 doc502
    doc401 doc307 doc303 doc308 doc107 doc810
    doc910 doc903 doc601 docA11 doc902 docB01
    doc508 doc501 docB09 doc905 docA01 doc512
    doc511 doc507 docA07 docA09 doc510 doc509
    doc908 doc901 doc407 doc904 doc206 doc305
    doc402 doc505 doc310 doc209
    最大確信度をもつメタデータ：心臓（0.3）
        ========================================
        cluster 1-0-0:
        doc105 doc508 doc601 doc810 doc905 doc901
        doc904 doc910 doc908 doc407 doc204 doc401
        doc107 doc303 doc307 doc308 doc310 docB01
        doc209 docA09 docB09 docA07 doc305 doc402
        doc403 docA01 doc206
        最大確信度をもつメタデータ：がん（0.44）
        ========================================
        cluster 1-0-1:
        doc406 doc507 doc501 doc502 doc510 doc902
        doc505 docA11 doc511 doc512 doc504 doc903
        doc509
        最大確信度をもつメタデータ：心臓（0.85）
    ========================================
    cluster 1-1:
    doc306
    最大確信度をもつメタデータ：肺（1.0）
========================================
cluster 2:
doc701 docB06 docB03 docB08 docB05 docB10
doc704 doc708 doc707 doc709 doc909 doc705
docB07 docB04 doc907
最大確信度をもつメタデータ：疲れ（0.47）
========================================
```

図 10　再帰的意味的クラスタリング（方式 A）の結果（部分）
Fig. 10　Result-3 of recursive semantic clustering (Method-A).

5.4 実　験　C

［方法・結果］

実験 C は文脈依存動的クラスタリングの再帰的適用に関する解析的な実験である．実験 C の目的は，意味的連想処理およびクラスタリングに関連した，分析者の与えるパラメータ群によって，最終的に得られるクラスタ数をコントロール可能であり，また，クラスタ数の初期設定値に対して依存度の低いクラスタ群を形成することが可能であることを確認することである．具体的には，実験 A，実験 B で使用したものと同一

```
============================
cluster 0:
  13  がん
  7  糖尿病
  6  肺がん
  5  胃
  5  胃がん
============================
cluster 1:
  13  がん
  11  心臓
  10  心臓病
  7  肺がん
  6  早期がん
  6  心筋梗塞
  6  肺
    ============================
    cluster 1-0:
      12  心臓
      11  心臓病
      11  がん
      6  肺がん
      6  心筋梗塞
        ============================
        cluster 1-0-0:
          12  がん
          6  肺がん
          5  肺
          5  早期がん
          4  扁平上皮がん
        ============================
        cluster 1-0-1:
          10  心臓
          9  心臓病
          4  心筋梗塞
          4  動脈硬化
    ============================
    cluster 1-1:
      1  肺
      1  肺がん
      1  がん
============================
cluster 2:
  7  疲れ
  4  過労
  4  ストレス
  3  自律神経
  3  過労死
============================
```

図 11 全クラスタを対象とした意味的データマイニングの結果（部分）
Fig. 11 Result-3 of semantic data mining.

のドキュメントデータを対象として，分析者によって与えられる初期設定クラスタ数と，そのパラメータによって最終的に生成されるクラスタ数の関係を調べる．同様に，分析者によって与えられる最小確信度（閾値）と，そのパラメータによって最終的に生成されるクラスタ数の関係について調べる．

以下のような 4 種類のパラメータ設定を与え，再帰的意味的クラスタリング方式を適用し，それによって生成されるクラスタ数を調べた．

(1) パラメータ設定 1
最小確信度（ε_c） 0.4
初期設定クラスタ数 2〜40

(2) パラメータ設定 2
最小確信度（ε_c） 0.5
初期設定クラスタ数 2〜40

(3) パラメータ設定 3
初期設定クラスタ数 10
最小確信度（ε_c） 0.1〜1.0

(4) パラメータ設定 4
初期設定クラスタ数 15
最小確信度（ε_c） 0.1〜1.0

さらに，それぞれに関して，文脈語として,「ストレス・不安」,「かぜ・鼻づまり」,「がん・悪性しゅよう」,「アルツハイマー病」,「文脈なし」の 5 種類を与えた．「文脈なし」の場合は，4.2.1 項で述べた意味的連想処理機構における部分空間選択を行わない状態，すなわち，イメージ空間 $\mathcal{I}$ 上でクラスタリングを行った．パラメータ設定 1〜4 に対応する結果は，それぞれ図 12〜15 である．それぞれの図中には，比較のために，再帰を行わないクラスタリングによる結果もグラフとして表示した．

［考　察］

初期設定クラスタ数と生成されたクラスタ数に関する実験結果（**図 12**，**13**）において，初期設定クラスタ数が小さいときは，そのクラスタ数に対して生成されたクラスタ数の比が大きく，逆に，初期設定クラスタ数が多いときは，その比が小さくなっている（再帰なしの場合にはその比は変化しない）．再帰なしの場合との比較において，本方式では初期設定クラスタ数の設定値に対する依存度の小さいクラスタ群を形成することが可能であることが確認できる．

最小確信度（閾値）と生成されたクラスタ数に関する実験結果（**図 14**，**15**）において，最小確信度の値が大きくなるにつれて，生成されたクラスタ数が多くなっている．このことから，分析者の与える最小確信度の値によって，生成されるクラスタ数をコントロールすることが可能であることが確認できる．同時に，最小確信度の値を大きく設定することにより，より共通の性質を有するクラスタ群の獲得が可能であることを示している．

実験 C において得られた最大の知見は，本方式が閾値 ε_c を与えられることによって分析結果をコントロール可能な点である．図 12 および図 13 より，初

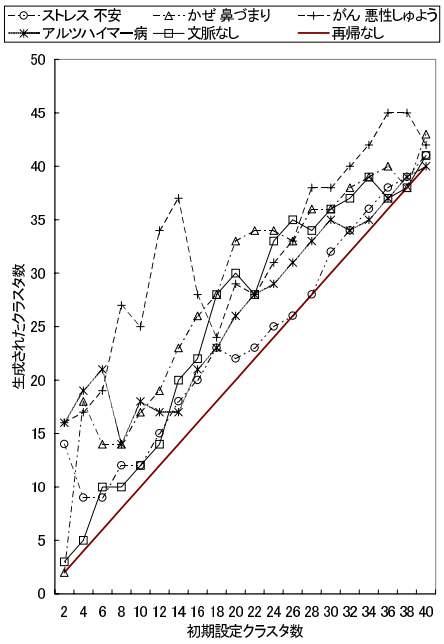

図 12 パラメータ設定 1 に対する結果
Fig. 12 Result for parameter setting 1.

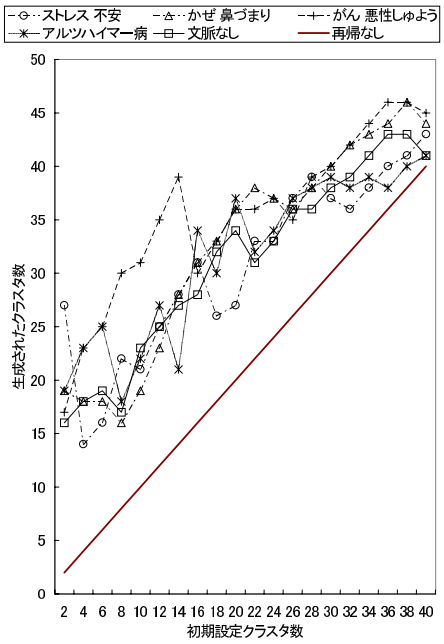

図 13 パラメータ設定 2 に対する結果
Fig. 13 Result for parameter setting 2.

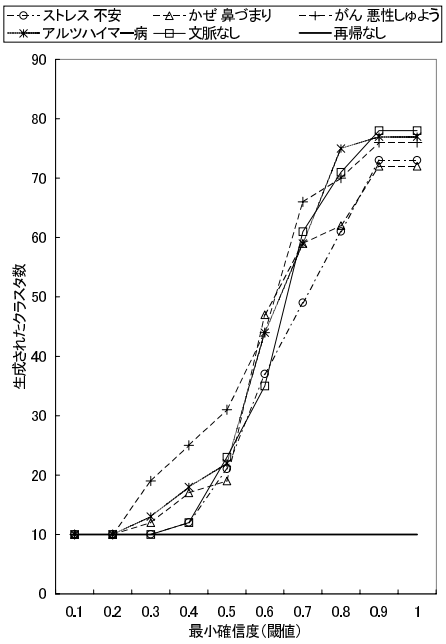

図 14 パラメータ設定 3 に対する結果
Fig. 14 Result for parameter setting 3.

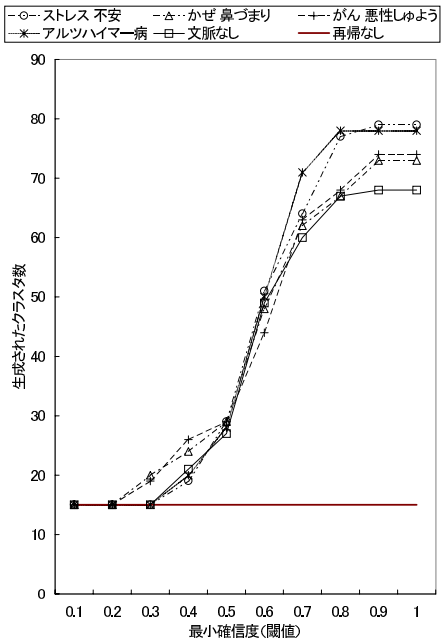

図 15 パラメータ設定 4 に対する結果
Fig. 15 Result for parameter setting 4.

期設定クラスタ数を比較的小さい値に設定しても有効な数のクラスタが生成可能であることが分かる．さらに，図 14 および図 15 より，本方式は，閾値として最小確信度 ε_c によりクラスタ数をコントロール可能であることが示された．以上 2 項目より，本方式は，初期設定クラスタ数が低く設定され，最小確信度 ε_c のみを与えられることにより有効な数のクラスタを得られる方式であることが分かる．

```
グループ 1
doc10: アレルギー アレルギー疾患 アレルギー症状
doc11: アレルギー アレルギー疾患 アレルギー性疾患
doc12: アレルギー アレルギー疾患 アレルギー性鼻炎
          :
グループ 2
doc20: 呼吸まひ 呼吸困難 呼吸障害
doc21: 呼吸まひ 呼吸困難 呼吸停止
doc22: 呼吸まひ 呼吸困難 呼吸不全
          :
グループ 3
doc30: 脳血管 脳血管疾患 脳血管障害
doc31: 脳血管 脳血管疾患 脳血管性痴呆
doc32: 脳血管 脳血管疾患 脳血栓
          :
          :
グループ 8
doc80: 胸 胸やけ 胸筋
doc81: 胸 胸やけ 胸痛
doc82: 胸 胸やけ 胸部
          :
```

図 16 実験に使用したテストデータ（一部）
Fig. 16 Test data for experiment.

5.5 実 験 D

[方法・結果]

実験 D は，実験 C と同様に，文脈依存動的クラスタリングの再帰的適用に関する解析的な実験である．実験 D の目的は，本論文で提案するクラスタリング方式の精度を確認することである．具体的には，テストドキュメントデータを対象として，分析者によって与えられる最小確信度と，そのパラメータによって最終的に生成されるクラスタ群の精度との対応を調べる．ここでは，再帰的クラスタリングと再帰しない場合との比較において，本方式の精度を示す．

以下のようにしてテストデータを作成した．

(1) 8 種類の単語（メタデータ）の組を作成した．それぞれの組には関連のある単語が 5 個ずつ含まれている．

(2) 各組の 5 個の単語のうち，3 個の組合せをメタデータとして持つ 8 個のグループからなるドキュメントデータを作成した．つまり，各グループ 10 のドキュメントデータ，全体で 80 のドキュメントデータを用意した．これらの 8 個のドキュメントの集合を,「クラスタ」という用語と区別して，今後「グループ」と呼ぶ．

作成したテストデータの一部を**図 16** に示す．

80 のテストドキュメントデータを対象に，初期クラスタ数を 4 として，再帰的意味的クラスタリングを適用した．この際，文脈として「がん・かぜ・呼吸まひ・胸・老化」を与えた場合と，文脈を与えない場合の 2 種類を行った．それぞれのクラスタリング結果に関して，クラスタの精度を評価した．ここでの精度とは，クラスタ内に含まれる全ドキュメントデータのうち，同一のグループに含まれるドキュメントデータが含まれる割合を，全クラスタにおいて平均した値である．実験結果を**図 17** に示す．

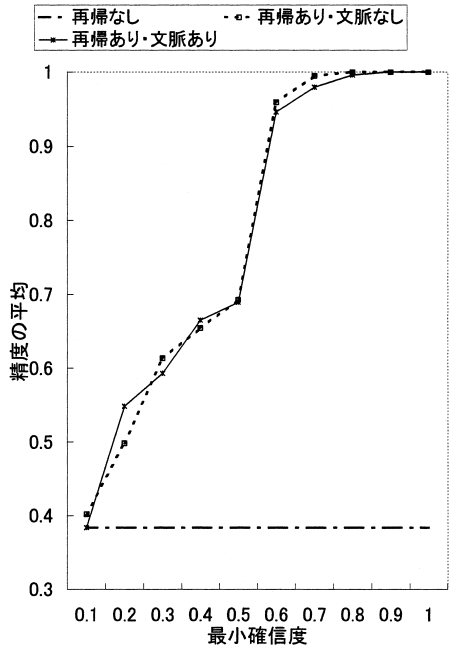

図 17 クラスタリングの精度に関する実験結果
Fig. 17 Result for precision of clustering.

[考 察]

最小確信度と生成されたクラスタの精度に関する実験結果（図 17）において，再帰あり（本方式）の場合では，最小確信度の値を大きくすると，生成されるクラスタの精度が上昇する傾向を確認できる．このことから，分析者が与える最小確信度の値を大きくすることによって，より共通の性質を有するクラスタ群の獲得が可能となることを確認できる．さらに，本方式では，文脈を与えた場合と，文脈を与えない（部分空間選択を行わない）場合で，ほぼ同じような精度の変化が得られている．このことから，メタデータの出現頻度を利用した再帰的なクラスタリング方式が，一般的なベクトル空間において，汎用的に動作するということを確認できる．

5.6 実 験 E

[方法・結果]

実験 E は，再帰的意味的クラスタリングの結果抽出（4.1 節の方式 B）に関する実験である．実験 E の目的は，実験 B で得られたクラスタ群に対して，クラスタ内に含まれるドキュメントの数が少ないものを

表 1 各クラスタの評価値
Table 1 Evaluated values for each cluster.

クラスタ ID	最大確信度	ドキュメント数	評価値
Cluster 0	0.61	23	14.03
Cluster 1	0.32	41	13.12
Cluster 1-0	0.3	40	12.0
Cluster 1-0-0	0.44	27	11.88
Cluster 1-0-1	0.85	13	11.05
Cluster 1-1	1.0	1	1.0
Cluster 2	0.47	15	7.05

排除し，ドキュメント数のより多いクラスタを抽出することである．これによって，ドキュメントの共通性と数のバランスのとれたクラスタ群の獲得が可能となることを示す．

実験 B の文脈語列 “ストレス，不安” に対する再帰的クラスタリング結果（図 10）の全クラスタについて，4.1 節における方式 B の評価関数を適用し，各クラスタの評価値を算出する．この実験における評価関数 f_e を以下の式で示す．

$$f_e(mc_i, dn_i) = mc_i \cdot dn_i$$

評価関数の引数は，クラスタ Cl_i について，4.1 節における方式 A の手順 2 によって求められた最大確信度 mc_i と，含まれるドキュメント数 dn_i をそれぞれ表している．この評価関数は，クラスタの最大確信度とドキュメント数を単純に掛け合わせたものである．この関数は，クラスタ内のドキュメントの共通性と数の値を反映した評価値を計算するものであり，両方の性質を評価するための一関数として採用している．

各クラスタを対象として，評価関数によって算出された評価値を**表 1** に示す．表 1 の評価値の高い順にクラスタをランキングして表示したものを**図 18** に示す．

［考　　察］

クラスタリングの表示結果（図 18）について，上位にランキングされているクラスタ（評価値が高いクラスタ）は，分析者が参照する場合において，十分な数のドキュメントを有していることが確認できる．このことは，本方式が分析者に対して，知識を発掘するのに十分なドキュメントのまとまりを提示することが可能であることを示している．逆に，下位にランキングされているクラスタ（評価値が低いクラスタ）は，クラスタ内に十分ドキュメントを有していない．特に，Cluster 1-1 は高い確信度を示しているが，1 つのドキュメントしか有していない．これらは，本方式が多くのドキュメントを有し高い確信度を有するクラスタを最適クラスタとして抽出可能であり，それ以外のクラスタを最適でないクラスタとして排除可能であることを示している．

```
========================================
cluster 0: (評価値: 14.03)
doc101 doc108 doc811 doc111 doc812 docA06
doc205 docA12 doc208 doc302 doc911 doc202
docA05 doc309 doc409 docB02 docA03 doc110
doc103 doc506 doc304 doc201 doc104
========================================
cluster 1: (評価値: 13.12)
doc105 doc204 doc910 doc512 doc904 docA01
doc401 doc508 doc306 doc107 doc402 doc303
doc305 doc310 doc403 doc901 doc810 doc307
doc308 doc209 doc406 docA11 doc504 doc509
doc507 doc510 doc902 docB01 doc407 doc905
docA09 doc206 doc502 doc908 doc903 doc511
doc505 docA07 doc601 docB09 doc501
========================================
cluster 1-0: (評価値: 12.0)
doc105 doc204 doc403 doc406 doc504 doc502
doc401 doc307 doc303 doc308 doc107 doc810
doc910 doc903 doc601 docA11 doc902 docB01
doc508 doc501 docB09 doc905 docA01 doc512
doc511 doc507 docA07 docA09 doc510 doc509
doc908 doc901 doc407 doc904 doc206 doc305
doc402 doc505 doc310 doc209
========================================
cluster 1-0-0: (評価値: 11.88)
doc105 doc508 doc601 doc810 doc905 doc901
doc904 doc910 doc908 doc407 doc204 doc401
doc107 doc303 doc307 doc308 doc310 docB01
doc209 docA09 docB09 docA07 doc305 doc402
doc403 docA01 doc206
========================================
cluster 1-0-1: (評価値: 11.05)
doc406 doc507 doc501 doc502 doc510 doc902
doc505 docA11 doc511 doc512 doc504 doc903
doc509
========================================
cluster 2: (評価値: 7.05)
doc701 docB06 docB03 docB08 docB05 docB10
doc704 doc708 doc707 doc709 doc909 doc705
docB07 docB04 doc907
========================================
cluster 1-1: (評価値: 1.0)
doc306
========================================
```

図 18 再帰的意味的クラスタリング（方式 B）の結果
Fig. 18 Result for recursive semantic clustering (Method-B).

実験 E の結果（図 18）は，実験 A の結果（図 6 および図 8）との比較において，本方式が，cluster 1, cluster 1-0, cluster 1-0-0 などのすべての中間結果を評価値によりランキングをともなって獲得可能なことを示している．これは，本論文で提案している中間結果を含めて評価値によりランキングする方式（3.1.5 項の方式 B）の実現可能性および妥当性を示している．

本実験の結果は，3.1.5 項における Phase-1 の Step-5 の方式 A を適用した後に方式 B を適用することにより，ドキュメントの共通性とクラスタに含まれるド

キュメント数のバランスのとれたクラスタ群の獲得が可能であることを示している．

5.7 実験全体の考察

本方式で与える閾値については，次のように考察できる．本方式は，任意の対象データについて，分析者が閾値を設定することなく有効な意味的クラスタ群を抽出可能である．分析者が与える閾値は，ε_c および初期設定クラスタ数のみである．初期設定クラスタ数は，実験 C より，比較的小さい値に設定しておけばよいという事実が得られた．図 17（実験 D に対応する結果）において 0.5 以降の急激な上昇は，少数のドキュメントによるクラスタが多数生成されたことにより確信度が上昇したものと考えられるため，本実験に用いたデータを対象とした場合は 0.5 程度が最適と考えられる．しかし，確信度とドキュメント数に依存した評価値によるランキングにより，ε_c を高く設定しておけば，全自動的に有効なクラスタ群を抽出可能である．以上より，本方式は任意の対象データについて，分析者が閾値を設定することなく有効な意味的クラスタ群を全自動的に抽出可能であると考えられる．

処理時間に関しては，次のように考察できる．提案方式である再帰的意味的クラスタリング方式は，方式 A，方式 B より構成されている．方式 A に関して，95 検索対象についての処理時間は，実時間で 6.02 [sec.] である（5 回の実験の平均値）．方式 B に関して，同一検索対象についての処理時間は，実時間で 0.17 [sec.] である（5 回の実験の平均値）．すなわち，提案方式の処理時間は，これらの 2 つの平均値を足した値である 6.19 [sec.] ということになる．この結果は，提案方式が現実的な時間内で処理可能であることを示している．しかし，対象ドキュメント数が多くなるに従い，再帰的意味的クラスタリングの処理量が増大するので，本方式におけるクラスタリングアルゴリズムの高速化が必用となる．これについては，今後の研究課題とする．

ここで，本実験において生成されるクラスタ数や精度については，イメージ空間 $\mathcal{I}$ の生成方式に依存する．イメージ空間 $\mathcal{I}$ の生成方式については，文献 17) に示している．

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の再帰的適用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．

6. 結　　論

本論文では，データの意味的な解釈をともな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の再帰的適用による意味的知識発見方式について述べた．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．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．さらに，文脈依存動的クラスタリングを再帰的に適用することによって，共通の性質を有するより多くのドキュメントが含まれる最適クラスタの発見が可能となった．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．

今後は，ドキュメントデータ群の重みつきメタデータの自動抽出方式，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 本研究にあたって貴重なご助言をいただいた筑波大学電子・情報工学系北川高嗣先生に感謝の意を表します．

参 考 文 献

1) Agrawal, R., Imielinski, T. and Swami, A.: Mining Association Rules between Sets of Items in Large Databases, *Proc. ACM SIGMOD*, pp.207–216 (1993).
2) Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, *Proc. 20th International Conference on Very Large Data Bases*, pp.487–489 (1994).
3) Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. and Simoudis, E.: Mining Business Databases, *Comm. ACM*, Vol.39, No.11, pp.41–48 (1996).
4) Dublin Core Metadata Initiative: http://purl.org/dc/
5) Jain, A.K., Murty, M.N. and Flynn, P.J.: Data Clustering: A Review, *ACM Computing Surveys*, Vol.31, No.3 (1999).
6) Han, J. and Kamber, M.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers (Aug. 2000).
7) 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*, Sheth, A. and Klas, W.(Eds.), Chapter 7, McGrawHill (1998).
8) Kobayashi, M. and Takeda, K.: Information retrieval on the web, *ACM Comput. Surv.*, Vol.32, No.2, pp.144–173 (2000).
9) Lent, B., Agrawal, R. and Srikant, R.: Discovering Trends in Text Databases, *Proc. 3rd In-*

ternational Conference on Knowledge Discovery and Data Mining (*KDD-97*), pp.227–230 (1997).
10) Resource Description Framework (RDF): http://www.w3.org/RDF/
11) Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T.: Mining Optimized Association Rules for Numeric Attributes, *PODS'96*, pp.182–191 (1996).
12) Wan, S.J., Wong, S.K.M. and Prusinkiewicz, P.: An algorithm for multidimensional data clustering, *ACM Trans. Math. Softw.*, Vol.14, No.2, pp.153–162 (1988).
13) 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).
14) 塩谷　實：多変量解析概論，朝倉書店 (1990).
15) 図子泰三，吉田尚史，清木　康，北川高嗣：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を用いた意味的知識発見方式，情報処理学会研究報告，情報処理学会データベースシステム研究会，2000-DBS-122, pp.331–338 (2000).
16) 図子泰三，吉田尚史，清木　康，北川高嗣：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の再帰的適用による意味的知識発見方式，データベースと Web 情報システムに関する合同シンポジウム（DBWeb2000），pp.221–228 (2000).
17) 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.SIG 5 (TOD2), pp.15–27 (1999).
18) 吉田尚史，図子泰三，清木　康，北川高嗣：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，情報処理学会論文誌：データベース，Vol.41, No.SIG 1 (TOD5), pp.127–139 (2000).

(平成 13 年 9 月 25 日受付)
(平成 13 年 12 月 28 日採録)

（担当編集委員　田中 克己）

図子　泰三（学生会員）

1976 年生．1999 年慶應義塾大学環境情報学部卒業．現在，同大学院政策・メディア研究科修士課程に在学中．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．

吉田　尚史（正会員）

1972 年生．1996 年筑波大学第三学群情報学類卒業．1998 年同大学院修士課程理工学研究科修了．2001 年同大学院博士課程工学研究科修了．博士（工学）．2001 年より慶應義塾大学大学院政策・メディア研究科講師．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．

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修了．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，電子情報通信学会各会員．