データベースシステム　122－43
(2000. 7. 27)

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

図子 泰三[†]　吉田 尚史[††]　清木 康[†††]　北川 高嗣[††††]

†慶應義塾大学 政策・メディア研究科　††筑波大学大学院 工学研究科
†††慶應義塾大学 環境情報学部　††††筑波大学 電子・情報工学系

〒252-0816 神奈川県藤沢市遠藤 5322 慶応義塾大学 i501
TEL: 0466-47-5000(53251)　E-mail: tz@mdbl.sfc.keio.ac.jp

あらまし

本稿では，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を用いた意味的知識発見方式について示す．本方式の特徴は，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，本稿では，意味的にクラスタリングされたドキュメントデータのメタデータを対象としたサブクラスタ生成方式を導入する．ドキュメントデータ群を用いた実験結果を示し，本方式の実現可能性および有効性を確認する．

キーワード　ドキュメントマイニング，意味的連想処理，クラスタリング，知識発見

A Semantic Knowledge Discovery Method Using Context Dependent Dynamic Clustering for Document Data

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

†Graduate School of Media and Governance, KEIO University
††Doctroal Program in Engneering, University of Tsukuba
†††Faculty of Environmental Information, KEIO University
††††Institute of Information Sciences and Electronics, University of Tsukuba

5322 Endo, Fujisawa, Kanagawa, Japan, 252-0816
Keio University i501
TEL: 0466-47-5000(53251)　E-mail: tz@mdbl.sfc.keio.ac.jp

Abstract

In this paper we present a semantic knowledge discovery method using context dependent dynamic clustering for document data. The main feature of the method is to make clustering for document 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. We introduce a clustering method that further classifies each target cluster. This method is applied to the metadata attached to the document data formed in the semantic clustering. We clarify feasibility and effectiveness of the method by showing several experimental results using actual document data.

key words　*Document Mining, Semantic Associative Processing, Clustering, Knowledge Discovery*

1. はじめに

近年，コンピュータネットワークの急速な普及により，多種多様なドキュメントデータが検索対象となっている．World Wide Web の普及に伴い，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いる．ドキュメントデータを対象とした的確な情報獲得の方式の実現およびデータマイニングの方式の実現が重要な課題となっている．

データマイニングに関する研究[4]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[7]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[11),12)]．

本稿では，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を用いた意味的知識発見方式を示す．本方式の特徴は，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を対象として，ドキュメントデータを構成するメタデータを用いた詳細なサブクラスタを生成可能な点にある．これにより，文脈依存動的クラスタリングにより得られた意味的なクラスタを対象として，詳細な分析が可能となる．

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),8)]．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，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．さらに，メタデータを用いたサブクラスタの生成により，洗練された対話的なクラスタ分析を可能とする．

2. 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概要

本節では，本方式の概要を示す．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[5),6),10)]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確信度を用いて，サブクラスタを生成する．

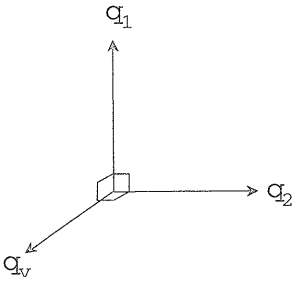

図1 Step-1: 正規直交空間の生成 (q_1 ～ q_v: 正規直交軸)

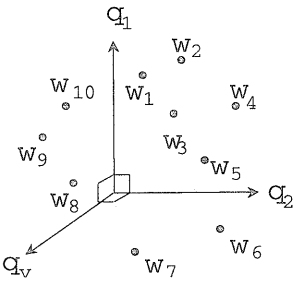

図2 Step-2: 分析対象アイテム群の正規直交空間へのマッピング (w_1 ～ w_{10}: 分析対象アイテム)

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．

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 5 ステップにより実現される．Step-1 から Step-5 の定式化および詳細については，文献[12)]に示している．

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

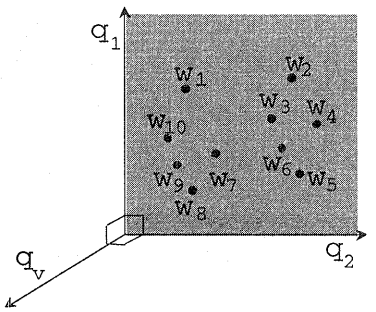

図3 Step-3: 問合せに応じた部分空間選択

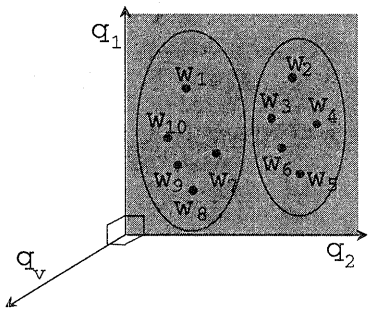

図4 Step-4: 部分空間上での分析対象アイテム群のクラスタリング

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[5),6),10)]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．

2.1.5 Step-5: メタデータを用いたサブクラスタの生成

前項で形成されたクラスタ群のうち，分析者より指定されたあるクラスタを対象として，ドキュメントデータを構成するメタデータの確信度をもちいてさらに分割し，サブクラスタを生成する．これにより，分析者と対話的な洗練された分析が可能となる．

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. 定式化およびアルゴリズム

本節では，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）の定式化およびアルゴリズムについて述べる[12)]．

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を用いた意味的知識発見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 意味的連想処理機構

本節では，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[5),6)] に示されている．

(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．

3.3 アルゴリズム

3.3.1 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．

ここでは，クラスタリング・アルゴリズムとして融合法[9)] 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．

(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．

3.3.2 メタデータを用いたサブクラスタの生成方式

分析者によって指定されたサブクラスタの数に基づいて，その数分のサブクラスタの候補を生成する．具体的には，Step-1 で形成された，ある特定のクラスタについて，そのクラスタ内に含まれるアイテム群の任意の組合せを生成し，サブクラスタの候補とする．さらに，各候補サブクラスタ内に含まれるメタデータの確信度（confidence）を計算し，最も確信度の高いものを選択する．ここでの確信度とは，サブクラスタ内でのメタデータの出現回数をサブクラスタ内に含まれるアイテム数で割った値である．そして，各候補サブクラスタ群ごとに，評価関数を適用し，最も高い評価値を得たものを結果サブクラスタの組とする．ここでの評価関数 f_e を以下に示す．

$$f_e = \frac{1}{n} \sum_{i=1}^{n} N_i \cdot confidence(md_i)$$

N_i は i 番目のサブクラスタに含まれるアイテムの

数，md_i は i 番目のサブクラスタの中で最も確信度の高いメタデータ，n は生成するサブクラスタの数をそれぞれ示している．

3.3.3 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のアルゴリズム，すなわち，2節における Phase-2 についての詳細なアルゴリズムについて述べる．

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. データ構造と基本演算子

ここでは、文脈依存動的クラスタリング方式と、その結果を対象とした対話的サブクラスタ生成方式の両方式を有するドキュメントデータ分析システムを実現するためのデータ構造と基本演算子について述べる。

4.1 文脈依存動的クラスタリングにおけるデータ構造と基本演算子

文脈依存動的クラスタリングにおけるデータ構造と基本演算子を示す。

文脈依存動的クラスタリングにおいては、1ドキュメントデータが、空間上の1ドキュメントデータ・ベクトルに対応する。基本演算子は、意味的連想処理によって導出される意味空間上でのクラスタリングである。パラメータとしては、文脈語列、検索対象ドキュメントデータ・ベクトル集合、生成するクラスタ数を指定する。

この演算子を、次のように定義する。

- `(clustering_by_context [context] [target] [clusnum])`

ここで、`context` は文脈語列、`target` は検索対象ドキュメントデータ・ベクトル集合、`clusnum` は生成するクラスタ数を表すパラメータである。

4.2 サブクラスタ生成クラスタリングにおけるデータ構造と基本演算子

サブクラスタ生成クラスタリングにおけるデータ構造と基本演算子を示す。

サブクラスタ生成クラスタリングにおけるデータ構造は、ドキュメントデータの ID とそれに対応するメタデータである。基本演算子は、メタデータの確信度によるクラスタリングである。パラメータとしては、入力としてのクラスタリング結果、サブクラスタに分割するクラスタ ID、分割するサブクラスタの数を指定する。

この演算子を、次のように定義する。

- `(clustering_by_confidence [target] [clusid] [clusnum])`

ここで、`target` は入力としてのクラスタリング結果、`clusid` はサブクラスタに分割するクラスタ ID、`clusnum` は分割するサブクラスタの数を表すパラメータである。

5. 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を用いた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

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．

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，3.2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．

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．具体的には，次の手順によりメタデータを

```
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

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．

5.1.4 実験システム

3節で述べた方式により実験システムを構築した．

5.2 実験方法

分析対象ドキュメントデータ群に対して様々な文脈語列（問い合わせ）を与え，分析結果を得る．実験 A および実験 B の 2 種類の実験を行なう

実験 A の目的は，文脈に応じたクラスタリングの変化をを検証する．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．

```
(clustering_by_context '(ストレス 不安) target 10)
```

図6 実験 A に用いた問い合わせ（文脈：ストレス，不安）

```
cluster 0:
doc101 docA04 doc906 doc910 docA05 docA02
docA01 doc911 doc909 doc908 docA03 doc110
doc102 doc907 doc207 doc208 doc407 doc106
doc107 doc502 docB06 doc206 doc202 doc104
doc108 doc109 docB04 doc204 doc203 doc205
doc210 docB05 doc209 doc103 docB07 doc105
doc504 doc201 doc111 doc503

cluster 1:
doc302 doc707 doc811 doc709 doc812 doc903
doc505 docA11 docA10 doc507 doc509 doc601
doc901 doc403 doc512 doc511 doc404 doc305
doc506 doc510 doc703 doc704 doc508 docB08
doc701 doc309 doc308 doc905 doc303

cluster 2:
doc306 docA08 docB03 docB02 docA09 docB01
docA12 docA07 docA06 doc810 doc904 doc902
doc708
```

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）

実験 B の目的は，分析者との対話的な分析の有効性を示すことである．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クラスタから，本方式によって複数のサブクラスタが生成される様子を確認する．さらに，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，分析者の要求に応じたクラスタの特徴が発見できることを示す．

5.2.1 実験 A

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．

文脈語列“ストレス，不安”に対応する問い合わせ，クラスタリング結果，分析結果は，それぞれ図 6，図 7，図 8である．

同様に，文脈語列“疲労，五十肩”に対応する問い合わせ，クラスタリング結果，分析結果は，それぞれ図 9，図 10，図 11である．

図 8および図 11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

cluster 0			
L1:		L2:	
38	がん	17	がん, 肺がん
17	肺がん	13	がん, 早期がん
13	胃がん	13	がん, 肺
13	肺	13	肺, 肺がん
13	早期がん	12	がん, 胃がん
12	胃	12	胃, 胃がん
9	扁平上皮がん	11	がん, 胃
		9	肺がん, 扁平上皮がん
		9	がん, 扁平上皮がん

cluster 1			
L1:		L2:	
15	心臓	13	心臓, 心臓病
14	心臓病	9	心筋梗塞, 心臓病
10	心筋梗塞	8	心筋梗塞, 心臓
10	糖尿病	6	虚血性心疾患, 心臓病
7	狭心症		
7	ストレス		
6	虚血性心疾患		

cluster 2			
L1:		L2:	
8	疲れ	4	過労, 疲れ
4	糖尿病	3	高血糖, 糖尿病
4	過労	3	ストレス, 疲れ
3	高血糖	3	うつ症状, 疲れ
3	慢性疲労	3	疲れ, 慢性疲労
3	うつ症状		
3	ストレス		

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）

```
(clustering_by_context '(疲労 五十肩) target 10)
```

図9 実験Aに用いた問い合わせ（文脈：疲労 五十肩）

```
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10 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）

数）とそのメタデータの組を示している．

5.2.2 実 験 B

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7）におけるcluster2について，クラスタリングを施し，サブクラスタに分割した．クラスタリングの際のパラメータとして，クラスタ数2に設定した．

実験Aと同様に，アプリオリアルゴリズムを適用し，分析を行なった．最小支持度は30%である．

cluster 0			
L1:		L2:	
36	がん	17	がん, 肺がん
17	肺がん	13	肺, 肺がん
13	肺	13	がん, 肺
13	早期がん	13	がん, 早期がん
12	胃がん	11	胃, 胃がん
11	胃	11	がん, 胃がん
		10	がん, 胃

cluster 1			
3	生活習慣病	2	心臓, 心臓病
2	冠動脈疾患	2	心臓, 生活習慣病
2	心臓病	2	心臓病, 生活習慣病
2	心臓		

cluster 2			
L1:		L2:	
5	疲れ	2	疲れ, 不眠
4	頭痛	2	疲れ, 慢性疲労
2	発熱	2	自律神経, 疲れ
2	微熱	2	うつ症状, 不眠
2	慢性疲労	2	過労, 慢性疲労
2	うつ症状	2	うつ症状, 疲れ
2	筋肉	2	過労, 疲れ
2	脳しゅよう	2	頭痛, 発熱
2	自律神経	2	筋肉, 疲れ
2	過労		
2	不眠		

図11 文脈“疲労，五十肩”の意味的データマイニングの結果（部分）

```
(clustering_by_confidence target cluster2 2)
```

図12 実験Bに用いた問い合わせ

```
subcluster 2-0:
docA08 docB03 docB02 docA09 docB01 docA12
docA07 docA06

subcluster 2-1:
doc306 doc810 doc904 doc902 doc708
```

図13 クラスタリングによって得られたサブクラスタ

発行した問い合わせ，クラスタリング結果，分析結果を，それぞれ図12，図13，図14に示す．

5.3 考 察

実験Aの結果より，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．以上から，図7，図10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8，図11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

subcluster 2-0

L1:		L2:	
7	疲れ	3	うつ症状, 疲れ
3	うつ症状	3	過労, 疲れ
3	過労	3	疲れ, 慢性疲労
3	慢性疲労		

subcluster 2-1

L1:

3	インスリン非依存型糖尿病
3	糖尿病
2	高血糖

L2:

2	高血糖, 糖尿病
3	インスリン非依存型糖尿病, 糖尿病
2	インスリン非依存型糖尿病, 高血糖

図14 サブクラスタの意味的データマイニングの結果

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．

実験 B の結果より，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 13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．つまり，メタデータを利用した確信度によるクラスタリングが効果的に機能していることを示している．

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．図 14より，subcluster-2-0, subcluster-2-1 はそれぞれ，「疲れ」，「糖尿病」に関するサブクラスタであることがわかる．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獲得できることを示している．

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

本稿では，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を用いた意味的知識発見方式について示した．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．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，クラスタリング結果をもとに，分析者の要求に応じて，特定のクラスタを対話的に詳しく分析す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．

今後は，本方式の高速化，ドキュメントデータ群の重み付きメタデータの取り扱い，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．

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) 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.
5) 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.
6) 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.
7) 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.
8) Jain, A.K., Murty, M.N. and Flynn, P.J.: "Data Clustering: A Review," ACM Comuting Surveys, Vol. 31, No. 3, 1999.
9) 塩谷實:"多変量解析概論," 朝倉書店, 1990.
10) 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.
11) 酒井大, 清木康, 吉田尚史: "ドキュメントデータ群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," 第 11 回データ工学ワークショップ (DEWS'00) 論文集, 電子情報通信学会,2000.
12) 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.