データベースシステム 118−12
情 報 学 基 礎 54−12
（1999. 5. 17）

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

吉田 尚史† 図子 泰三†† 清木 康†† 北川 高嗣†††

† 筑波大学 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

要旨

本論文では，ドキュメントデータ群を対象としたデータマイニングを実現するための基礎となる動的クラスタリ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う点にある．本方式により，同一解析対象であるドキュメントデータを対象として文脈に応じて多数の意味的解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に対象として効率的なブラウジングを可能とする．ここでは，実験結果を示し，提案方式の有効性を確認する．

キーワード: 動的クラスタリング，意味的連想処理，データマイニング

Dynamic Clustering Method with Semantic Associative Processing for Document Data

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

† Doctroal Program in Engneering, University of Tsukuba

†† Faculty of Environmental Information, Keio University

††† Institute of Information Sciences and Electronics, University of Tsukuba

abstract

In this paper we propose a dynamic clustering method as the basis of data mining 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 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 of the method by showing several experimental results.

keywods: dynamic clustering, semantic associative processing, data mining

1 はじめに

近年，データベースにおける知識獲得 (KDD: Knowledge Discovery in Databases) の研究が注目されている [3, 9, 11]. KDD の一プロセスであり，その重要な役割を担うデータマイニングに関する技術が注目されている [5]. データマイニングとは，データベースとして格納されている大量データの中から，そこに潜在する知識や事実を発見する手法である．

データマイニングの本質は，分析対象である大量の生データ群に内在する知識を抽出することにある．そのためには，多量のデータを対象として意味的に近いデータ群をグループ化し，クラスタ分析することが有効であると考えられる．

本稿では，ドキュメント (文書) データ群を対象として，データの意味的解釈を伴うデータマイニングを実現するための基礎となる動的クラスタリ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を行う点にある．本方式により，解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解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に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[12]において多くの方式が提案されている．従来の統計的解析法との比較において，本方式の特徴は，文脈に応じて動的に解析結果を求め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解析結果を得るだけでなく，文脈や状況に応じて動的に解析結果を得ることを可能とする．

本方式は，意味的連想処理機構[6, 7, 13]を用いて実現される．意味的連想検索機構では，データ間の意味的な同一性，差異性は，静的な関係によって決定されるのではなく，文脈や状況に応じて動的に変化するものと考える．このモデル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，意味的連想検索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．

意味的連想検索機構では，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．意味的動的クラスタリング方式は，この部分空間の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，部分空間の選択後に，クラスタリングのアルゴリズムを適用する．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．

2 動的クラスタリング方式の概要

本節では，動的クラスタリング方式の概要を示す[14].

2.1 概要

本方式は，次の５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．

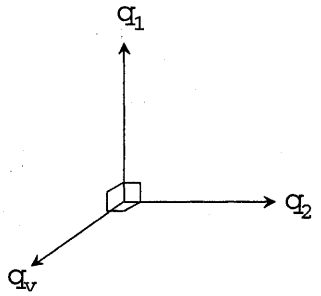

図 1: Step-1: 正規直交空間の生成 (q_1 ～ q_v: 正規直交軸)

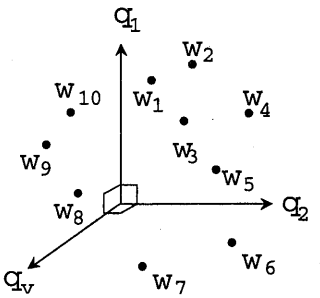

図 2: Step-2: 分析対象アイテム群の意味空間へのマッピング (w_1 ～ w_{10}: 分析対象アイテム)

Step-2 ：分析対象アイテム群の意味空間へのマッピング
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
意味的連想検索方式の応用により，問合せ（文脈語列）に応じて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．

Step-5 ：クラスタ群の分析
Step-4 で得られたクラスタ群を分析し，各クラスタ内の分析対象アイテム群に共通して現れる性質を知識として抽出する．

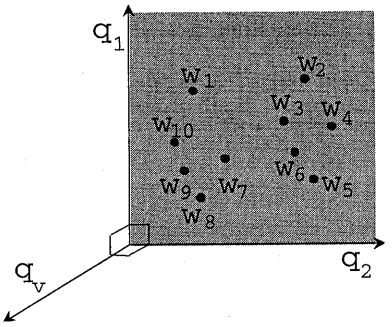

図 3: Step-3: 問合せに応じた部分空間選択

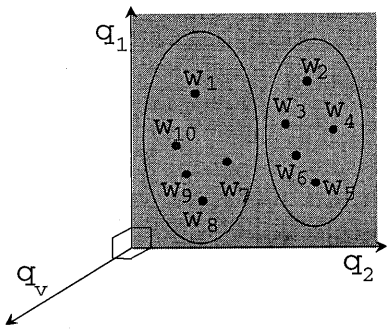

図 4: Step-4,5: 部分空間上での分析対象アイテム群のクラスタリング，および，クラスタ群の分析

2.2 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）．

2.3 Step-2: 分析対象アイテム群の意味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）．

2.4 Step-3: 問合せに応じた部分空間選択

意味的連想検索機構 [6, 7, 13] の特徴である部分空間選択の方式を用いて，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．

2.5 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に動的に対応することが可能である．

2.6 Step-5: クラスタ群の分析

前項で得られたクラスタ群を対象に分析を行う（図 4）．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．

3 動的クラスタリング方式の定式化

ここでは，意味的連想処理機構 [6, 7] による動的クラスタリングの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，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．

3.1 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 $A^T A$ を作る．

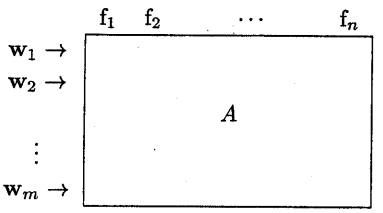

図 5: データ行列 A の構成

2. A^TA を固有値分解する.

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

$$0 \leq \nu \leq n.$$

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 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.

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

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.

3.3 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.

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

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

3.4 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，3.3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．

3.5 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．

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方法として，以下の3方法を提案する．

方法 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

方法 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

方法 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

3.5.1 方法 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

方法 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方法である．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 [12] 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，3.4 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 4 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毎に，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毎に，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．

3.5.2 方法 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

方法 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方法である．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．

3.5.3 方法 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

方法 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方法である．クラスタ群の生成については，方法 A_1，方法 A_2 と同様である．

3.6 クラスタ群の分析方式

得られたクラスタを分析する方式として，分析対象アイテム群のメタデータを用いる 2 種類の方法を提案する．

方法 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．

方法 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．

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，提案方式である動的クラスタリング方式の実現可能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，3.1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．

```
doc001: がん 肺がん 肺 リンパ節
doc002: がん 肺がん 肺 腰椎 しびれ ぎっくり腰
doc003: がん 胃がん 早期がん 胃 吐血 下血
doc601: 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
doc602: 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
doc603: 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: 実験に使用した分析対象アイテム群の例

3.1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 100 ドキュメントデータを用いた．100 の各ドキュメントデータに対し，メタデータとして数語の meta words を付与した．このメタデータの一部を図 6に示す．

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．

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，3.5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，方法 A_1 を採用した．

クラスタ群の分析方式については，方法 B_1 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，3.5.1節の (b) として述べた．これは，(c) および (d) 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

ClusterID	meta words	Support	Confidence
cluster1	がん	0.287671	0.440476
cluster1	肺がん	0.287671	0.202381
cluster1	肺	0.287671	0.154762
cluster1	早期がん	0.287671	0.154762
cluster1	胃がん	0.287671	0.154762
...			
cluster2	糖尿病	0.363014	0.113208
cluster2	心臓	0.363014	0.0660377
cluster2	心臓病	0.363014	0.0660377
cluster2	心筋梗塞	0.363014	0.0566038
cluster2	肥満	0.363014	0.0566038
...			
cluster3	ホルモン	0.00684932	0.5
cluster3	糖尿病	0.00684932	0.5
cluster3	じん臓病	0.00684932	0.5
...			

図 7: 実験結果 1 (文脈: ストレス, 不安) (部分)

ClusterID	meta words	Support	Confidence
cluster1	がん	0.243151	0.478873
cluster1	肺がん	0.243151	0.239437
cluster1	肺	0.243151	0.183099
cluster1	早期がん	0.243151	0.169014
cluster1	胃がん	0.243151	0.15493
...			
cluster2	生活習慣病	0.0650685	0.157895
cluster2	冠動脈疾患	0.0650685	0.105263
cluster2	胃	0.0650685	0.105263
cluster2	心臓病	0.0650685	0.105263
cluster2	心臓	0.0650685	0.105263
...			
cluster3	疲れ	0.236301	0.0724638
cluster3	頭痛	0.236301	0.0724638
cluster3	糖尿病	0.236301	0.0434783
cluster3	筋肉	0.236301	0.0434783
cluster3	不眠	0.236301	0.0289855
...			

図 8: 実験結果 2 (文脈: 疲れ, 疲労) (部分)

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, 解析結果を得る. 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.

文脈語列 (問合わせ) には, 次の 2 種類を与えた. "ストレス, 不安", "疲れ, 疲労" 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.

4.3 実験結果

実験結果を次のように示す. 2 文脈語列 "ストレス, 不安", "疲れ, 疲労" に対応する実験結果は, それぞれ, 図 7, 図 8である.

4.4 考察

実験結果より, 互いに意味的に相関の強い単語が同一クラスタを構成していることが確認できる. さらに, 与えた文脈語列により, クラスタ構成の様子が変化しているのが確認できる.

さらに, 相関ルールアルゴリズムの適用により, 各クラスタの意味が抽出でき, 分析対象のドキュメントデータを概観することができる.

また, 実験結果 1 と実験結果 2 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.

この実験結果は, 文脈に応じた動的なクラスタリングが可能な本方式の実現可能性を示している.

5 結論

本稿では, データの意味的な解釈を伴うデータマイニングを行うための基礎となる動的クラスタリ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て動的にクラスタリングができる点が特徴である. 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. 本方式により, 解析対象のデータに対して, 文脈に応じて動的に意味的解析結果を得ることが可能となる. また, ドキュメントデータを対象とした実験により, 本方式の実現可能性を確認した.

今後は, 本方式の高速化, 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

チメディアへの適用を行う予定である.

謝辞

本稿で用いた実験データは，読売新聞社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] 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.

[4] Kitagawa, T. and Kiyoki, Y.: The mathematical model of meaning and its application to multidatabase systems, Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp. 130-135, April 1993.

[5] 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.

[6] 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.

[7] Kiyoki, Y., Kitagawa, T. and Hayama, T.: A metadatabase system for semantic image search by a mathematical model of meaning, ACM SIGMOD Record, vol. 23, no. 4, pp. 34-41, 1994.

[8] Kiyoki, Y.,Kitagawa, T. and Hitomi, Y.: A fundamental framework for realizing semantic interoperability in a multidatabase environment, Journal of Integrated Computer-Aided Engineering, Vol.2, No.1, pp. 3-20, John Wiley & Sons, Jan. 1995.

[9] 河野浩之,"データベースからの知識発見の現状と動向," 人工知能学会誌, Vol.12, No.4, pp. 497-504, Jul. 1997.

[10] Longman Dictionary of Contemporary English, Longman, 1987.

[11] 西尾章治郎,"大規模データベースにおける知識獲得," 情報処理, Vol. 34, No. 3, pp. 343-350, 1993.

[12] 塩谷實,"多変量解析概論," 朝倉書店, 1990.

[13] 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.

[14] 吉田尚史, 清木康, 北川高嗣, "意味的連想処理機構を用いた大量データ分析のための動的クラスタリング方式," 情報処理学会研究報告, 98-DBS-116(1), pp.143-150, 1998.