多次元データマイニングを用いたWeb空間の構造解析

山下　由展† 　大森　匡† 　星　守†

† 電気通信大学大学院情報システム学研究科　〒 182–8585 東京都調布市調布ヶ丘 1–5–1
E-mail: †{yamashita,omori}@hol.is.uec.ac.jp

あらまし　著者らは, 従来から, 多次元的制約の下でデータマイニングを OLAP 的に行なうデータキューブ機構「アイテムセットキューブ」を試作しており, 多様なデータマイニング応用の中核として使うことを目指している. 本稿は, このデータマイニング機構の応用として, 多次元的制約下で Web 空間の構造計算を行なった結果を報告する. Web 構造マイニングの研究では, 従来から, 完全２部グラフ (コア) 計算によるコミュニティの検出やコア間の関連計算などがある. 本稿では, 電気通信大学ドメイン uec.ac.jp の Web 空間を対象に, どの集団から見て解析するか, どの集団にとって重要か, などの多次元的制約の下でコア間の関連性を表すグラフ作成とランク計算をアイテムセットキューブ上で行なうシステム例を報告する.

キーワード　データマイニング, OLAP, Web とインターネット, アイテムセットキューブ

Mining Web Structures Using Multi-dimensional Data Mining Model

Yoshinobu YAMASHITA†, Tadashi OHMORI†, and Mamoru HOSHI†

† The University of Electro-Communications, Graduate School of Information Systems　Chofugaoka 1–5–1, Chofu, Tokyo, 182–8585 Japan
E-mail: †{yamashita,omori}@hol.is.uec.ac.jp

Abstract　We make data cube mechanism ”Item set cube” that does data mining in OLAP style under a multi-dimensional restriction so far. Our objective is to use it as a kernel of various data mining applications. In this paper we report results of calculating the structure of an intra-net web space under a multi-dimensional restriction as an application of this data mining mechanism. Researches of web structural mining include complete bipartite-graph (core) detection and relationship calculation between cores. In this paper, we report a system of making such a graph that shows both relationship between cores and their ranks under multi-dimensional restrictions.

Key words　Data mining, OLAP, Web and Internet, Itemset cube

1. はじめに

Web ページ間を結ぶハイパーリンクによって構成されるグラフ構造を解析し, 隠れた有益な情報を抽出する研究は Web 構造マイニングと呼ばれる. これら Web 構造マイニングの研究では, 完全２部グラフ (コア) 計算による Web 上のコミュニティの検出 [5] や, コア間の関連計算 [3] 等により Web 空間の解析を行っている.

従来の Web 構造マイニングの研究では Web 空間全体を分析対象としている [3] [4] [5] [6]. そのため, 分析対象を 1 組織に限定するなど対象範囲が限られてくる場合, 従来の手法を用いると分析の粒度が粗くなることが考えられる.

一方, 著者らは従来から, 指定した着眼点についてすぐにデータマイニングを行うことができるデータキューブ機構「アイテムセットキューブ」を試作している [1] [2]. アイテムセットキューブは, 分割を与えた属性を次元としてもつ多次元構造をしており, 属性に囲まれたセルには, そのセルの条件を満たす高頻度アイテムセットを格納している. このアイテムセットキューブは多次元データマイニングモデルを採用しており, 多次元制約下で各属性別に着目した解析を行うことができる.

本論文では, 高頻度アイテムセット計算をコア計算と対応づけて考えることにより, アイテムセットキューブが支援する多次元データマイニングを用いて, 電気通信大学リンク構造データを対象に, どの集団から見て解析するか, どの集団にとって重要か, などの多次元的制約の下でコア間の関連性を表すグラフ作成とランク計算を行う. また, 本論文では, 多次元データマイニングを Web 構造分析に適用した時のコミュニティ検出能力の有効性について評価を行う. さらに, そのときのアイテムセットキューブシステムの有効性を評価する.

本論文の構成は以下である. 2 章で準備として関連研究やアイ

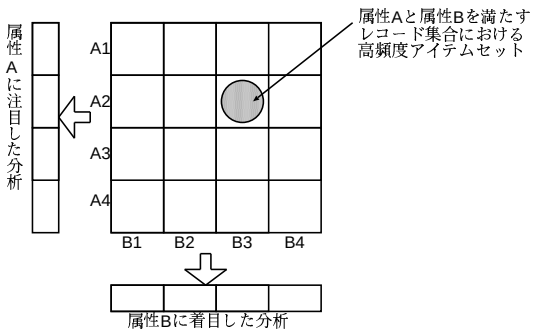

図 1　アイテムセットキューブ

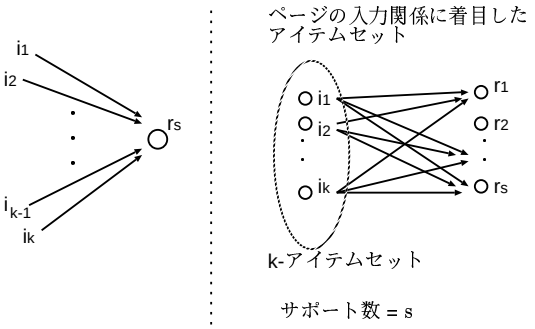

図 2　アイテムセットを用いたコア計算

テムセットキューブ, アイテムセット計算等について触れ, 3 章では, 本論文で提案する多次元データマイニングを用いた Web 構造解析手法の導入部について述べる. 4 章では提案する手法の各手順について細かく説明し, 5 章で実際にクローラーを用いて収集した Web リンク構造データへ本論文の手法を適用し, 考察する. 6 章では多次元制約を変化させた場合のコアコミュニティの関連性について考え, 7 章ではアイテムセットキューブ作成手順による効率化について述べる. そして, 最後に 8 章でまとめを述べる.

2. 準　　備

ここでは準備として, まず始めに Web 構造マイニングの用語と関連研究について述べ, 次にアイテムセットキューブに関する説明を行う. そして最後にアイテムセット計算とコア計算との関係性について述べる.

2.1　Web 構造マイニングの用語と関連研究について

Web 空間上には互いに興味を持つページ集合が存在すると考えられており, そのような互いに興味を持ち合うページ集合はコミュニティと呼ばれる. また, コミュニティの中核をなすページ集合はコアと呼ばれる. Web 空間上でページをノード, ハイパーリンクを辺と見たときにできるグラフ構造において, このグラフに含まれる完全二部グラフは, 互いに興味を持つページを表すコミュニティのコアであると IBM [5] は定義している (ここで完全二部グラフが $|F|=i$, $|C|=j$ であれば, それは (i, j) コアと呼ばれる. また F はハブ, C はオーソリティと呼ばれる).

Web 構造マイニングの研究では, コアとなりえない辺を除去することで, 効率的にコミュニティのコアを列挙する研究 [5] や, コミュニティに含まれるページのページタイトルの中から頻出する 10 単語を取りだし, そのコミュニティを表すインデックスとすることで, コミュニティ検索を行う研究 [4], コミュニティ間の関連性を考えることで, コミュニティを単位とした地図をつくるウェブコミュニティチャート [3] と呼ばれる研究などがある.

2.2　アイテムセットキューブ [2]

データマイニングにおいて, 数値による分析は視覚的に分かりやすく, 着眼点を発見しやすい手法であるが, その要因を知ることができないという短所がある. そこで, 高頻度に出現するアイテム集合 (以下高頻度アイテムセットと呼ぶ) による分析で, 発見した着眼点で何が起きているのかを知る方法がある (与えられたレコード集合 D においてアイテムセット I が成立するレコード数を一般にサポート数と呼ぶ。ここで, アイテムセット I が高頻度であるとは, ある閾値 θ を与えたとき, レコード集合 D におけるアイテムセット I の出現頻度が (サポート数 s_I)/$|D|$ $> \theta$ となる時をいう). これは, 着眼点における条件を満たすレコード群から, 高頻度アイテムセットを算出して, 分析する方法である.

この高頻度アイテムセット計算をもとに, 指定した着眼点においてすぐにデータマイニングを行うことができるデータマイニング機構として「アイテムセットキューブ」がある.

アイテムセットキューブは, 属性を次元軸としてもった多次元構造をしており, そのセルには条件を満たすレコード群から算出された高頻度アイテムセットが格納されている (図 1). アイテムセットキューブを実体化しておくことにより, 各属性別に着目した分析を行うことができ, また, アドホックな問い合わせに対して瞬時に高頻度アイテムセットをかえすことができるため, 効率の良いアイテムセット分析を行うことが可能である.

2.3　アイテムセット計算とコア計算との関係性

アイテムセット計算を Web リンク構造データに用いることで (i, j) コアを求めることができる.

アイテムセット計算により (i, j) コアを Web リンク構造データから抽出するためには, Web リンク構造データに対し高頻度アイテムセットを求めるための $Apriori$ アルゴリズムを適用する. 閾値 (%) を定めたとき, アイテムセットの部分集合の一つでも閾値を下回っているものがあれば, そのアイテムセットはもはや高頻度ではないという考え方を $Apriori$ アルゴリズムでは採用しており, 長さが 1 短い高頻度アイテムセットから候補となるアイテムセットを生成し, その後データベーススキャンを行い, 高頻度アイテムセットを求めていく.

Web リンク構造データはリンクの入力関係に注目するとアイテムセット表現で表すことができる.

以下はリンクの入力関係に着目した時のアイテムセット表現である.

定義. ページの入力関係に着目したアイテムセット

ページの入力関係に着目したアイテムセットのレコード書式は以下で表される.

入力関係レコード書式:

レコード長 対象ページ 入辺の始点ページ 1

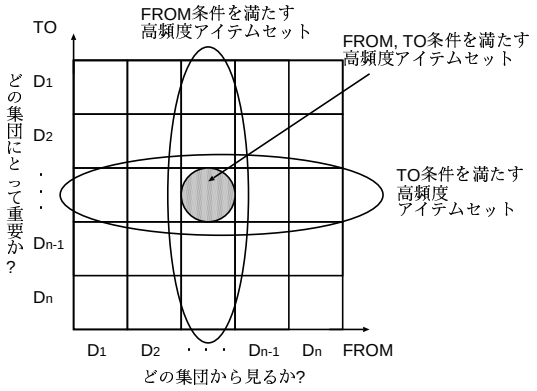

図 3　多次元データマイニングモデル

入辺の始点ページ 2 … 入辺の始点ページ m

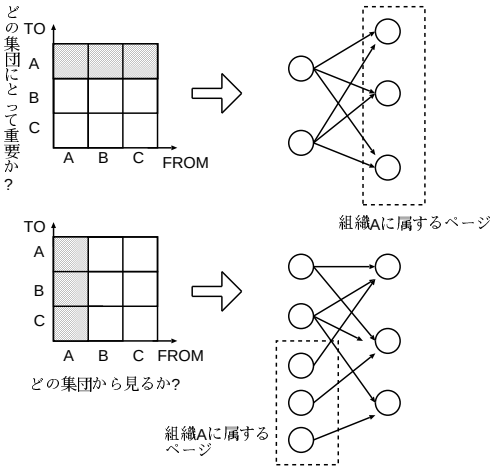

図 4　多次元制約がコアに与える影響

表 1　クローリング情報

クローラが収集したページ/サイト数	
リンク情報を収集した際に把握した uec.ac.jp 下のサイト数	362
全サイトのページ総数	108631

ページ i_1,i_2,...,i_k を入辺の始点ページ集合として含むページを考える. この時, i_1,i_2,...,i_k を k-アイテムセットと考えて, この i_1,i_2,...,i_k を入辺の始点ページに含むページ数をサポート数と考える (図 2)。

Apriori アルゴリズムを Web リンク構造データに適用する際, ページの入力関係に着目したアイテムセットを考えれば、ハブとしての役割が強い (i, j) コアを求めることができる.

なお, *Apriori* アルゴリズムによって求められるコアは以下で表されるものである.

- **ページの入力関係に着目した時にできるコア**

… ある閾値 θ を与えたとき, レコード集合 D において高頻度となる入辺の始点ページの組み合わせ (アイテムセット) を I とする. また, I を含む全てのレコードの対象ページ集合を A とすると, 求められるコアは (I, A) となる.

3. 多次元データマイニングによる Web 構造解析手法の導入

3.1 研究の動機

本論文では電気通信大学 uec.ac.jp 下のリンク構造 (これ以後電気通信大学リンク構造と呼ぶ) の分析を行う.

ここではまず始めに, 従来の Web 構造マイニングの研究手法を用いて分析を行った場合を考える. 従来の Web 構造マイニングの研究では, Web 空間全体を分析対象としている. Web 空全体 (10 億ページ以上の Web 空間) を対象とするような従来の研究では, サイト間リンクのみを分析対象としており, また分析対象空間も大きい. そのため, これらの手法を電気通信大学リンク構造のような対象 Web 空間の規模が限られてくるイントラネット組織の分析に用いると, 分析の粒度が粗くなることが考えられる.

そこで, 本論文では, サイト間リンクとサイト内リンクの両方を分析対象として, どの組織にとって重要か (TO), どの組織から見るか (FROM) の 2 つの属性を持つ多次元データマイニングモデルに, 高頻度アイテムセット計算を用いた (i, j) コア計算を組み合わせて, どの集団にとって重要か, どの集団から見るか, などの多次元制約下のコア計算をもとに, 電気通信大学リンク構造データを細かく分析する.

3.2 多次元制約とコア計算

今回考える多次元データマイニングモデルは 対象ページの所属と入辺の始点ページの所属関係に着目した場合にできるものである (図 3). この多次元データマイニングモデルでは, どの集団にとって重要か (TO), どの集団から見るか (FROM) の 2 つの属性を持っており, 属性 TO と属性 FROM はそれぞれ D_1, D_2, ..., D_n の n 個の組織で分割されている. 対象ページの所属ドメインが D_{k2}, 入辺の始点ページのなかで所属ドメインが D_{k1} であるものが含まれていれば, そのレコードはセル (D_{k1},D_{k2}) に所属するものとする. セル (D_{k1},D_{k2}) には, 上の条件を満たすレコード集合上の高頻度アイテムセットを含んでいる (図 3).

TO 側の集団を制約することで, コアのオーソリティ側には制約した TO 側の集団しか含まれないようにすることができ, また, FROM 側の集団を制約することで, 制約した FROM 側の集団からリンク参照されている集団 (重要と見なされている集団) がオーソリティ側に来るようになる.

例えば, TO を組織 A に制約することで, コアのオーソリティ側には組織 A の集団しか含まれないようになり, また, FROM を組織 A に制約することで, 組織 A からリンク参照されている (重要と見なされている) 集団がオーソリティ側に来るようになる (図 4)

3.3 対象とする Web 空間の基本的性質

ここでは, 今回分析対象とする電気通信大学リンク構造データの基本的性質について述べる. 今回分析対象とするのは, 2005 年 1 月にクローラーで収集した電気通信大学リンク構造データ (クローラーが集めた情報の内訳については表 1 参照のこと) である. 電気通信大学組織は uec.ac.jp を頂点として, その下に

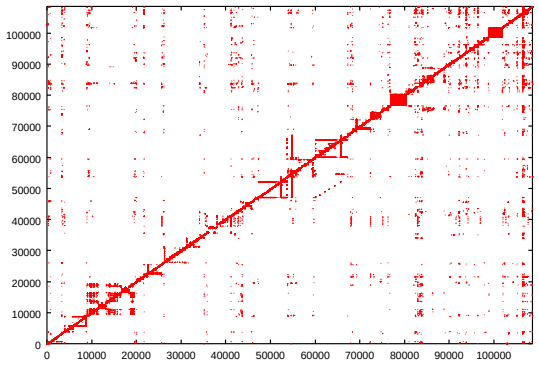

図 5　電気通信大学隣接行列表現 (ページ単位)

IS/C/J, EE, OTHER(IS/C/J, EE 以外の組織) サブドメインが来て, さらにその下に属する組織が続いていくという階層構造を成している(注1).

電気通信大学リンク構造データを隣接行列表現に直したものが図 5 である. 図 5 において, 各ページにはページ番号が与えられており, 属する組織が近いものほどページ番号が近くなるようになっている. 図 5 から, 関連ある組織内で密にリンクを張り合っていることなどを大まかに読み取ることができる.

また, ページ単位での組織間の関連性を求めるために, 対象組織を IS/C/J に限定し, サイト間リンクのみを用いて, 以下のことを行った.

まず, IS/C/J リンク構造データから閾値 0.01% で (i, j) コアを計算し, その後, 似たような組織で構成されるコア同士をまとめるために, コア同士をマージした. 次に, ハブによってオーソリティを説明づけるために, オーソリティ側を 1 ノード化して, ハブがオーソリティに対して辺を持つようにした (図 6).

図 6 からは研究室間の関連性や, ICE とコヒーレント光科学関連が関係していることなどが読み取れる. しかし, サイト間リンクのみを用いているため, 組織間の関連性が粗くなってしまっている. また, 組織単位でしか分析が行えず, どれが重要な組織なのかまでは読み取ることはできない.

3.4　従来手法と提案手法

図 6 のように粗い分析を行うと以下の問題点が起こる.

（1）　組織間の関連性が粗すぎる.

（2）　1 組織がかたまりすぎている.

（3）　どの組織が重要なのかがわからない.

本論文で提案する手法では, 多次元制約下でのコア計算に, コアを単位としたグラフ作成とランク計算を加えた以下の 3 つの操作を考えることで, これらの問題の改善を行う.

- 多次元制約下のコア計算
(サイト間リンクとサイト内リンクを採用する)
- コアを単位としたグラフ作成
- ランク計算

サイト間リンクとサイト内リンクを採用し, また, どの組織にとって重要か, どの組織からみて重要か等の多次元制約を取り入れ, コア単位でグラフを作成することで問題点 1, 2 の改善を行う. また, コア単位からなるグラフに対し, ランキング計算を行い, 重要な組織を目立たせることで問題点 3 を改善する.

(注1)：IS は情報システム学研究科, EE は電子工学科のことである. また, C は ICE(情報通信工学科) の略称, J は CS(情報工学科) の略称である. この他にも MCE(知能機械工学科), FEDU(留学生センター), などがある.

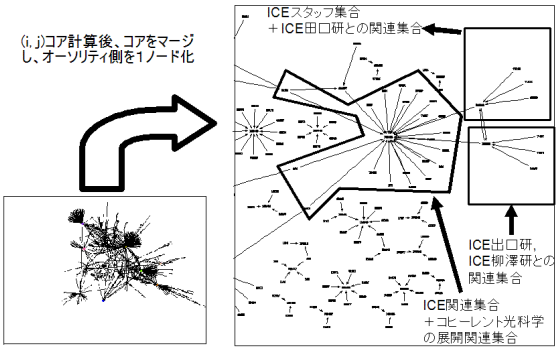

図 6　従来手法での分析 (IS/C/J))

4.　多次元制約下でのコアコミュニティグラフ作成とランク計算

4.1　提案手法の概要

多次元データマイニングモデルを用いて Web 構造解析を行う手法についての概要は以下の通りである. まず多次元データマイニングモデルを用いて, どの集団から見て解析するか, どの集団にとって重要か, などの制約を満たすコア集合を求め, そこから関連性のあるコア同士をマージし, コアのコミュニティ(これ以後コアコミュニティと呼ぶ) からなる集合を作成する. 次にそのコアコミュニティ集合からマップを作成した (これ以後コアコミュニティグラフと呼ぶ). コアコミュニティグラフをもとにコアコミュニティのランキングを行うことで対象 Web 空間の構造理解に役立てる.

4.2　コアコミュニティグラフ作成

コアコミュニティグラフ作成に用いられるため, コアコミュニティグラフ作成の前に以下のマージ処理について解説する.

4.2.1　マージ処理

以下の処理はコアをオーソリティ側共通項が n 個ある場合に, コア同士をマージする.

[3] ではオーソリティ側共通項 2 でマージしたコアをコミュニティの 1 つとして考えており, 共通項 2 でコア集合をマージすることにより, コアコミュニティ集合を作成できる.

処理:

コア Co' =($\{a_1,a_2,...,a_{|F|}\}$,$\{b_1,b_2,...,b_{|C|}\}$) と
コア Co'' =($\{c_1,c_2,...,c_{|F'|}\}$,$\{d_1,d_2,...,d_{|C'|}\}$) の
オーソリティ側に共通項が n 個あるとき
$Co'+Co''$ =($\{a_1,a_2,...,a_{|F|},c_1,c_2,...,c_{|F'|}\}$,
$\{b_1,b_2,...,b_{|C|},d_1,d_2,...,d_{|C'|}\}$)
とすることを再帰的に繰り返す.

4.2.2　コアコミュニティグラフ作成

ここでは, コアコミュニティグラフを作成する.

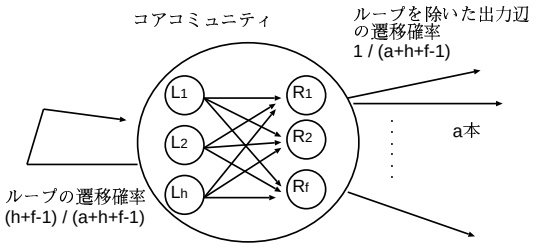

図 7　遷 移 確 率

1 巡目処理　サポート数 60 以上のコアを UEC 全体のリンク構造データから求め, その求めたコア (*) から, ハブの要素数 2 以上でかつ極大であるものを抽出する (**).

2 巡目処理　(*) を構成する要素を UEC 全体のリンク構造から取り除き, 対象とする組織間のリンクの参照関係を元にした多次元データマイニングモデルを作るものとする.

TO FROM の範囲を指定した時に該当するコアのうち, ハブの要素数 2 以上で, かつ極大となっているコアをオーソリティ側共通項 2 でマージしたもの (コアコミュニティに対応する) をそれぞれ 1 ノードと考え, これに (**) のうちハブ側に FROM, オーソリティ側に TO の要素を含むものをオーソリティ側共通項 2 でマージした後 (コアコミュニティに対応する) に 1 ノード化したものを加え, 以下の規則で辺を付加し, コアコミュニティをノード単位とするコアコミュニティグラフを作成した.

- ノード $n_1, n_2 \in V(G_0)$ のオーソリティまたはハブ間で共通項があるならば, (n_1,n_2), $(n_2,n_1) \in E(G_0)$.
- ノード $n_1, n_2 \in V(G_0)$ を構成するノード間にリンク集合での辺が存在するならば, 辺を付加する.

4.3　ランキング手法について

コアコミュニティグラフに対して Pagerank アルゴリズム [7] をかけ, コミュニティのランキングを行う.

Pagerank アルゴリズムは Web 空間上のページをページランクと呼ばれる値で評価するためのものであり, 良質なページはより多くの良質ページから参照されているという考えをもとにしている.

Pagerank アルゴリズムにおけるノード i の Pagerank x_i は以下で定義される ($A[i,j]$ はノード i からノード j への遷移確率.).

$$x_i = \epsilon \times (1 / (\text{ノード数})) + (1-\epsilon) \sum_{(j,i) \in E(G)} x_j \times A[j,i]$$

ただし, $\epsilon = 0.15$ ([7] においてこの数値を採用している), 各ノードの初期値は 1 とした. ここで第一項 が付け加えられている理由は, ユーザーが Web ページを巡回する際, 多くの場合は現在のページに存在するリンクをたどって移動するが, 時々は全く無関係なページにジャンプするという考えを Pagerank では採用しているからである.

また, コアコミュニティグラフにおいて, 1 コアコミュニティを 1 ノードと考えているため, コアコミュニティ内部に遷移する確率と, コアコミュニティ外部に遷移する確率を考慮する必要が

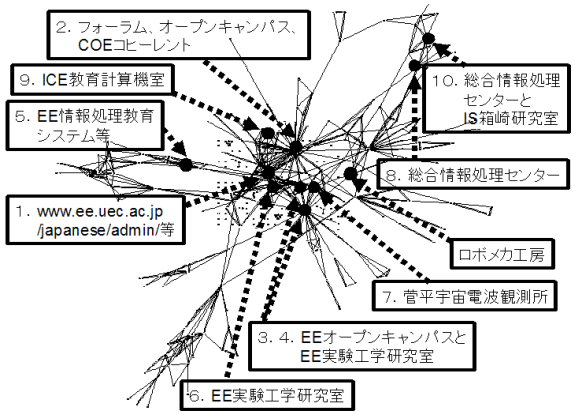

図 8　テスト 1: UEC 全体のコアコミュニティグラフ

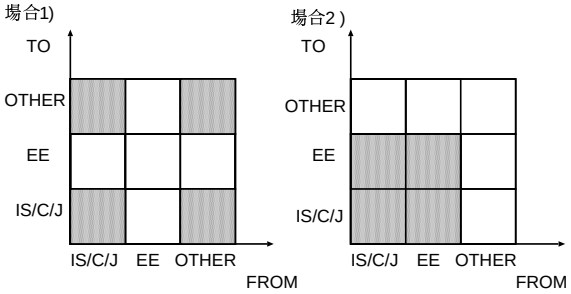

図 9　テスト 2: ドメイン限定

出てくる.

そのため各ノードの出力辺の遷移確率は今回, 以下の通りにしている.

ノード A が自己ループを除いて a 本の出力辺を持ち, ノード A のハブ数が h, オーソリティ数が f であるとする. この時, ノード A の出力辺の遷移確率を $1/(a+h+f-1)$, ループの遷移確率を $(h+f-1)/(a+h+f-1)$ とする (図 7).

ただしノード A の出力辺が同一サイト間のリンクである場合は PENALTY として今回 1/10 倍をしている (サイト内で閉じているコミュニティよりも, サイト間でつながっているコミュニティの方が情報量があると今回考えているため).

5.　Web リンク構造への適用

5.1　概　　要

ここでは, 2005 年 1 月にクローラーで収集した電気通信大学リンク構造データに対し, 多次元制約下で構造分析を行い, 多次元分析処理の評価を行う.

今回は, クローラーで収集した電気通信大学リンク構造データから学科トップ等の主要ページと, サイト内被リンク数 8 以上でかつサイト間リンクをもたないページを削除したリンク構造データを考える (ページ数:86161).

今回, このリンク構造データに対し以下の 3 つのテストを行った.

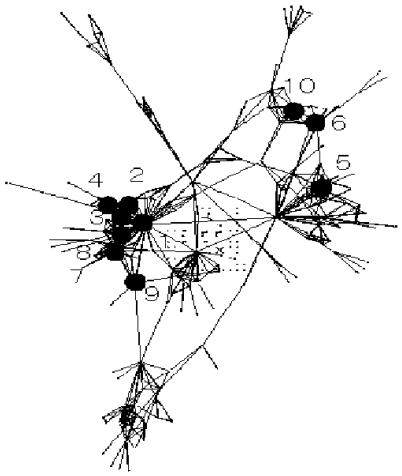

図 10　テスト 2: FROM, TO ともに IS/C/J, OTHER に制約した場合

テスト 1　多次元制約を考慮せずに, コアコミュニティグラフを作成し, ランキングを行う場合について評価を行う.

テスト 2　多次元制約を変化させた場合の, コアコミュニティグラフとランキングの変化について評価を行う.

テスト 3　FROM を制約した場合の特徴と, TO を制約した場合の特徴について評価を行う.

ここでは, これら 3 つのテストを通じて, 多次元制約下での Web 構造分析の有用性を示す.

5.2　テ ス ト 1

ここではまず, 多次元制約を考慮せずに, コアコミュニティグラフを作成し, ランキングを行う場合について評価する.

図 8 は, 多次元制約を考慮せずに電気通信大学全体のリンク構造データから作成したコアコミュニティグラフ (閾値 0.0098%, ノード数 270) である. また, 図 8 において, いくつか, コアコミュニティのランクを掲載してある.

本論文ではサイト間リンクとサイト内リンクの両方を用いており, また, コアコミュニティ単位で分析を行っているため, サイト内リンクのみを用いてページ単位の分析を行う場合に比べ, 組織間の関連性を細かく分析できる. またランク計算を取り入れているため, 対象組織において, どの組織のどんな活動が目立っているのかを理解することができる.

図 6 に含まれているコアコミュニティを観察すると, COE コヒーレントやフォーラム, ロボメカ工房など, UEC 組織のような階層組織を観察するだけでは読み取れない活動を発見することや, EE 関連組織や総合情報処理センター, ICE など, どの組織が UEC 全体で目立っているのかなどを読み取ることができる.

以上のように, コアコミュニティグラフとランク計算を組み合わせた解析を行うことで, 多くの情報が得られていることがわかる.

5.3　テ ス ト 2

次に, 多次元制約を考慮にいれた場合を考える. ここでは, 電気通信大学リンク構造データを元にした多次元データマイニングモデル (FROM, TO ともに IS/C/J, EE, OTHER の 3 つの組織で分割されている) 上で TO FROM 関係を以下の 2 つの場合にして, それぞれコアコミュニティグラフ作成とランク計算を行った (図 9. 閾値 0.0098%).

表 2　テスト 2: FROM, TO ともに IS/C/J, OTHER に制約した場合のランク上位 1 位から 20 位

ランク	コミュニティ
1	UEC フォーラムや UEC オープンキャンパス等
2	ICE 教育計算機室
3	ICE 教育計算機室と ICE 渡辺研究室
4	ICE 渡辺研究室 RTB
5	MCE 下条・明研究室関連
6	総合情報処理センター関連
7	ICE スタッフ関連
8	量子計算研究会と ICE 西野研究室
9	ICE 専攻紹介と ICE 来住研究室, ICE 三木研究室関連
10	総合情報処理センター関連と IS 箱崎先生
11	「楽力」教育メンバーと MCE スタッフ
12	ICE 西野研究室と関連研究
13	ICE 西野研究室と関連研究
14	ICE 田口研究室と ICE 角田研究室
15	FEDU 政府奨学金と IS,CS,MCE,PC
16	CS 大学院と計算科学講座関連
17	IS U 専攻と IS 箱崎研究室
18	楽力」教育とロボメカ工房
19	総合情報処理センターユタカ先生
20	ICE 高澤研究室と ICE 高澤先生

場合 1 FROM, TO ともに IS/C/J, OTHER に制約.
場合 2 FROM, TO ともに IS/C/J, EE に制約.

上の 2 つの場合について作成されるコアコミュニティグラフのノード数はそれぞれ 276, 406 であった.

図 10 は FROM, TO ともに IS/C/J, OTHER に制約した場合にできるコアコミュニティグラフ, 表 2 はランキング結果である. また, 図 11 は FROM, TO ともに IS/C/J, EE に制約した場合にできるコアコミュニティグラフで, 表 3 はランキング結果である.

場合 1 と場合 2 のコアコミュニティグラフを比べてみると, グラフの形に変化が起こっており, グラフの密な部分も変化している. このコアコミュニティグラフの違いがランキングに影響を及ぼしている. 場合 1 と場合 2 のランクを比べてみると, 場合 1 ではその多くを IS/C/J 関連コアコアコミュニティが占めているのに対し, 場合 2 では IS/C/J 関連のコアコミュニティはあまり入っておらず, 2 つの場合で IS/C/J の影響力が変化している.

多次元制約を変えることで, コアコミュニティグラフとランキングに変化が起こり, 制約した条件下でどの組織が目立っているのかを理解することができることがわかる.

5.4　テ ス ト 3

ここでは, TO を制約した場合と FROM を制約した場合の違いについて考える.

TO を制約した場合, (i, j) コアのオーソリティ側には, 制約した TO ドメインの要素しか含まれないようになる. また, FROM

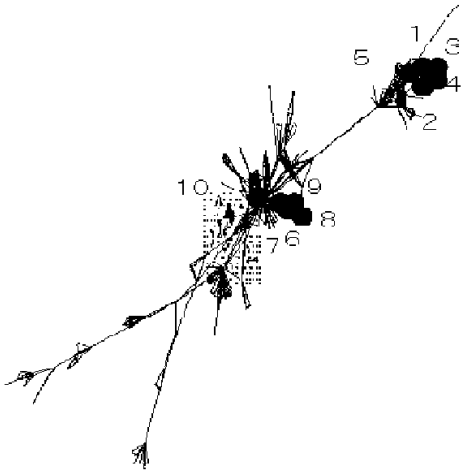

図 11　テスト 2: FROM, TO ともに IS/C/J, EE に制約した場合

表 3　テスト 2: FROM, TO ともに IS/C/J, EE に制約した場合のランキング結果上位 1 位から 20 位

ランク	コミュニティ
1	電子工学基礎セミナーと EE 斎藤先生その 1
2	電子工学基礎セミナーと EE 斎藤先生その 2
3	電子工学基礎セミナーと EE 斎藤先生その 3
4	電子工学基礎セミナーと EE 斎藤先生その 4
5	EE 斎藤先生授業ページ
6	EE 電子知能システム学講座
7	EE 金子研究室と金子研究室高橋さん
8	EE 電子知能システム学講座と kurelab
9	EE 電子知能システム学講座とナガイ先生
10	フォーラムやオープンキャンパス関連
11	オープンキャンパスと EE 実験工学研究室
12	オープンキャンパスと EE 実験工学研究室
13	EE 木村・一色研究室研究関連
14	www.ee.uec.ac.jp/japanese/admin/と EE 熊田研究室等
15	菅平宇宙電波観測所富井研究室関連
16	EE 実験工学研究室関連
17	EE 実験工学研究室関連
18	EE 電子知能システム学講座と kurelab 関連
19	ICE 教育計算機室
20	EE 早川研究室関連と IWSE

を制約することにより, FROM ドメインの要素にとって重要な要素がオーソリティ側に含まれることになる.

TO を制約した場合と FROM を制約した場合の違いをみるために, TO を OTHER に制約した場合と, FROM を OTHER に制約した場合とでそれぞれコアコミュニティグラフを作成して, ランキングを行った.

図 12 は, TO を OTHER に制約した場合と, FROM を OTHER に制約した場合にそれぞれコアコミュニティグラフを作成し, ランキングの分布を比べたものである (閾値 0.0098%).

TO を OTHER に制約した場合と FROM を OTHER に制約した場合とを比べてみると, コアコミュニティグラフの形が異なることがわかる. このグラフの形が異なることが影響し, ランキング上位 20 の要素の分布に変化がでていることが見て取れる.

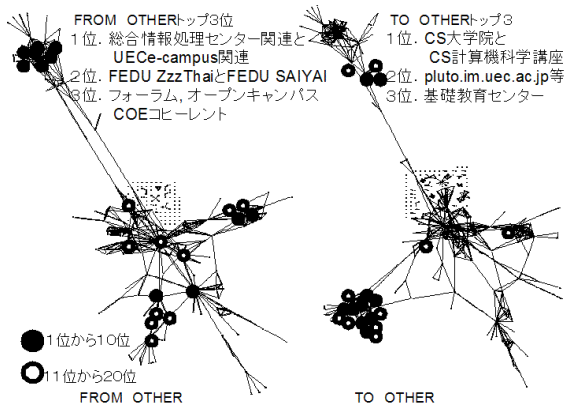

図 12　テスト 3: FROM OTHER と TO OTHER の特徴の比較

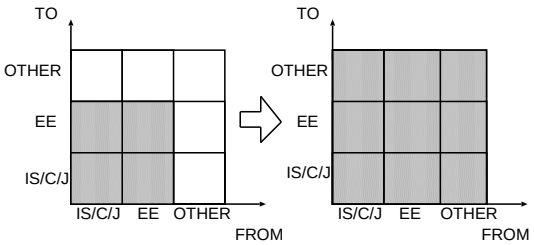

図 13　複合問い合わせ

以上のことから, FROM を制約した場合と, TO を制約した場合にそれぞれ得られる (i, j) コアが持つ特徴が, コアコミュニティグラフとランキングの分布に違いを与えることを示すことができた.

6.　複合問い合わせの記述

TO と FROM の制約を変化させると, それに伴いコアコミュニティグラフとランキング結果も異なってくる. ここでは TO と FROM の制約を変化させた場合に得られるコアコミュニティ間の関連性を考える.

ここでは制約を変化させた場合のコアコミュニティのランクの上昇について考えた.

ランキング結果間の関連性を考えるために, FROM, TO ともに IS/C/J, EE に制約した場合から, 何も制約しない場合に制約を変化させた時に, どのようなコアコミュニティのランクが上昇するのかを調べた (図 13). 結果として得られたランクの上昇を表 4 に掲載してある.

表 4 を見ると, フォーラムや菅平宇宙電波観測所, 教育計算機室など, より多くの組織と関連性のあるコアコミュニティのランクが特に上昇している. 今回のような制約の変化を考えると, IS/C/J, EE 組織の中での, より多くの組織と関連性のあるコアコミュニティを理解することができる.

今回のように, 制約を変化させた場合のランクの上昇など, 制約を変化させた場合に得られるランキング結果間の関連性を調べることで, さまざまな制約下で得られるコアコミュニティの関連性を理解することができ, 多次元制約を行った場合に得ら

表 4　ランクの上昇

コアコミュニティ	ランクの上昇
フォーラム 2004 とオープンキャンパス	10 位から 2 位
EE 実験工学研究室とオープンキャンパス等	11 位から 3 位
www.ee.uec.ac.jp/japanese/admin/等	14 位から 1 位
菅平宇宙電波観測所富井研究室等	15 位から 7 位
EE 実験工学研究室	16 位から 6 位
ICE 教育計算機室	19 位から 9 位
ICE 教育計算機室と ICE 渡辺研究室	24 位から 14 位
ICE 渡辺研究室と RTB	29 位から 22 位
ICE 福田研究室等と LENA プロジェクト	43 位から 28 位
ICE 西野研究室	44 位から 30 位
ICE 西野研究室と調布際研究室公開	45 位から 31 位

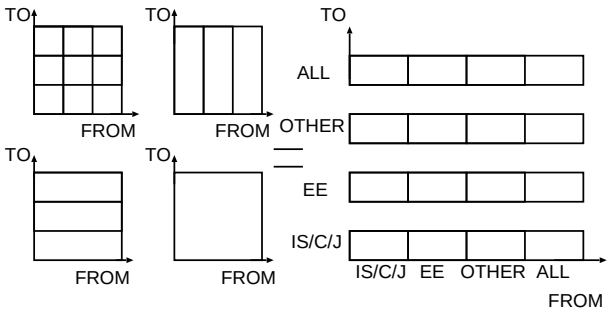

図 14　同時実体化

れる情報に加え, 新たな情報を得ることができる.

7. アイテムセットキューブ上の実装

レコードからアイテムセットキューブを作成する演算を「実体化」と呼ぶ. ここではアイテムセットキューブ上で多次元データマイニングを行う際の, 実体化手法による効率性について述べる. 実体化の方法には, 個別の制約条件ごとに計算する方法と, 相異なる条件下の計算を同時に実行する方法がある. 後者の手法として, 著者らは, 冗長な計算を削除して効率よく計算できる算法 Cubic Apriori [2] を提案している.

今回, アイテムセットキューブの実体化に関して, 個別の制約条件ごとに計算する場合と, Cubic Apriori を用いて同時に計算する場合とで実行時間の評価を行う. ここでは多次元データマイニングにおいてよく用いられる制約条件 (図 14 左) をもつアイテムセットキューブの実体化時間について評価を行う.

図 14 左のアイテムセットキューブの実体化は, 図 14 右のアイテムセットキューブを実体化することで実現できる. 図 14 右のアイテムセットキューブの実体化に関して, 個別実体化を行った場合の実行時間と, 同時実体化を行った場合の実行時間をそれぞれ表 5 に掲載している. 表 5 を見ると, TO をどの組織に制約した場合においても同時実体化を行った場合の方が個別実体化を行った場合よりも早いことが読み取れる. これより, アイテムセットキューブを実体化する際には同時実体化を行った方が個別実体化を行う場合よりも効率的であると言える.

8. おわりに

本稿では, 多次元データマイニング機構アイテムセットキューブの応用として, 電気通信大学リンク構造を対象に, サイト間リンクとサイト内リンクを用いて, どの集団から見て解析するか, どの集団にとって重要か, などの多次元制約下でコアコミュニティグラフを計算することに加え, ランク計算を行った結果を報告した.

まず始めに, サイト間リンクのみを用いて組織単位で分析を行う場合よりも, サイト間リンクとサイト内リンクを用いて, コアコミュニティグラフ作成とランク計算を組み合わせた手法を用いた場合の方が, より細かく分析できることを示した.

次に, 多次元制約を変化させた場合の違いを見るために, FROM, TO ともに IS/C/J, OTHER に制約した場合と, FROM, TO ともに IS/C/J, EE に制約した場合とを比較し, 多次元制約を変化させることで, コアコミュニティグラフが変化し, ランキングに変化が起こることを示した. また, 制約を変えた場合におけるコアコミュニティグラフの関連性の変化を調べることで, 多次元制約で得られる情報に加え, 新たな情報が得られることを示した.

また, アイテムセットキューブの実体化に関して, 個別実体化を用いた場合と同時実体化を用いた場合とで実行時間を比較することで, 多次元データマイニングをアイテムセットキューブ上で行う際の効率性について示した.

今回, コアコミュニティ間の辺にはコアコミュニティを構成するノード間の辺しか考慮に入れられていないが, ノード間を直接結ぶ場合以外にも, 別のノードを経由してノード間がつながっている場合も検討している.

表 5　同時実体化と個別実体化 (秒)

TO	個別	同時
IS/C/J	164.307101	110.975846
EE	7.830407	4.667393
OTHER	14.684115	9.651938
ALL	11.148614	10.129073

文　献

[1] 助川 貴信, 大森 匡, 星 守, 蔦谷 雄一, “Web ログ分析における高頻度アクセスパターン検出を支援するデータキューブモデル”, DEWS2003, 1-A-01, (2003).

[2] 成瀬 正英, 大森 匡, 星 守, “多次元的なログデータマイニングを実現するデータキューブ機構の提案と評価”, DEWS2005, 3C-i10, (2005).

[3] 豊田 正史, 吉田 聡, 喜連川 優, “ウェブコミュニティチャート – 膨大なウェブページを関連する話題を通して閲覧可能にするツール –” 電子情報通信学会論文誌, D-I Vol. J87-D-I No.2, pp.256-265, (2004).

[4] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, “Extracting large-scale knowledge bases from the web.” Proc. of the 25th VLDB Conference, pp.639-650, (1999).

[5] Ravi Kumar, Prabhakar Ragavan, Sridhar Rajagopalan, Andrew Tomkins, “Trawling the web for emerging cyber-communities,” In Proc. of the 8th WWW Conference, pp. 403-416, (1999).

[6] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Wei-Ying Ma, Hong-Jiang Zhang, Chao-Jun Lu, “User Access Pattern Enhanced Small Web Search,” In Proc. of the 12th WWW Conference, pp. 278, (2003).

[7] Sriram Raghavan, Hector Garcia-Molina, “Complex Queries over Web Repositories,” In Proc. of the 29th VLDB Conference, pp. 33-44, (2003).