DEWS2007 B8-5

多次元データマイニングによるWebアーカイブの構造解析の評価

林　和宏† 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: †{hayashi,omori}@hol.is.uec.ac.jp

あらまし　本研究室では，多次元制約下でデータマイニングを行うアイテムセットキューブの試作を行っている．昨年我々は，このアイテムセットキューブを用いたイントラネット型の Web 空間の構造解析を試行した [1], [2]．具体的には，アイテムセットキューブの多次元制約を用い，Web 空間内のコア計算と，コアをもとにしたグラフ作成を行った．その上でグラフ構造を評価しランキングを行うことで，注目する Web 空間でどの組織が重要視されているか分析を行った．本稿では，まず制約下でのランクアルゴリズムの改良と，それを用いた問い合わせの評価を行う．その上で，アーカイブ化したときの時間領域への対応やキーワードによる制約条件を用いた分析など，詳細な分析を行うための分析属性を取り入れた問い合わせを試行し有効性を報告する．

キーワード　データウェアハウス, データマイニング, Web とインターネット

Evaluation of Web-Structure Mining based on Multi-dimensional Data Mining

Kazuhiro HAYASHI†, 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: †{hayashi,omori}@hol.is.uec.ac.jp

Abstract　In today's web-organized societies, one of key technologies is to find interesting web-communities in a given Web network and to support a personalized search of the communities. For this objective, the authors previously proposed a multi-dimensional data-mining tool called an itemset cube. The itemset-cube system is used for creating a new graph-structure that models relationship among web-communities. This paper improves this graph-structure as well as a ranking method in the graph. The effects and usefulness of various complex queries based on the proposal are examined, by using the Web space of a university domain.

Key words　Data warehouse, Data mining, Web and Internet

1. はじめに

近年 Web 上での仮想組織の活動が活発になっており，Web 空間でどのような活動が行われているかを調べることが重要となってきている．また，個人や状況に応じて情報をパーソナライズして調べることが重要である．

一方本研究室では，多次元データキューブの下でデータマイニングを行う機構であるアイテムセットキューブの試作を行っている．アイテムセットキューブでは各属性で指定されたセルに，条件を満たす高頻度アイテムセットを格納している．これにより，各属性に着目した分析を即座に行うことができる．昨年著者らのグループでは，このアイテムセットキューブを用い多次元制約機構の上で高頻度アイテムセットとして Web 構造解析でいうコア（完全 2 部グラフ）を求め，コア単位のグラフ作成とランキングを行い，イントラネット型の Web 空間の構造解析を行った [1]．この多次元制約機構を用いた分析を行うことにより，制約なしで分析を行った場合には分からなかった結果などが得られ手法の有効性を感じることができた．しかし，詳細に調べたところ [2]，作成されたコアコミュニティグラフにもとづきランキングした結果について，グラフの構造とランクの間の隔たりが大きい場合があるという状況も見られた．そこで本稿では，まずコアコミュニティグラフとランキングに用いる式のモデルの再考を行う．その上で電気通信大学（UEC）ドメインのリンク構造データの分析結果から，この手法によって階層構造以外の構造をどの程度捉えることができているか改めて評価し報告をする．またこれまでは，Web ページの所属するドメインに着目し，どのドメインのページと参照関係にあるかといった視点を用いて分析を行っていた．しかし，このドメイン制約だけでは詳細な分析を行うことが難しいという問題があった．そこで，新たな属性を取り入れた分析機構を構築する．具

体的には，時間軸領域での分析や制約範囲を任意に定めた分析，キーワードによるページ集合を用いた分析の 3 点である．これらの属性を取り入れることにより可能になる分析の特徴と有効性の報告を行う．

本稿の構成は以下である．2 節で関連研究，3 節で多次元制約機構であるアイテムセットキューブを用いた Web 構造解析手法について述べる．4 節でコアコミュニティグラフの作成法について，5 節でランク式とその改善を示す．6 節では提案手法を用いた評価を行う．7 節では新たな属性による分析として，時間軸でのコミュニティ変化への問い合わせ，中間ページを経たページの参照関係を用いた分析，またキーワードによる制約を用いた分析を行った場合の結果について述べる．最後に 8 節でまとめを述べる．

2. 関 連 研 究

これまでに様々な Web 構造解析の研究が行われてきている．その一つとして，あるトピックに関心を持ったページ集合をコミュニティとし，それを導出することで Web 空間の理解に役立てようとするコミュニティ研究 [3] がある．この Web コミュニティにおいて中心となるページ集合をコアと呼ぶ. [3] では，Web 空間におけるページとハイパーリンクをノードとエッジとし，この Web 空間全体を巨大なグラフと見たときに現れる完全 2 部グラフがコミュニティのコアであるとした．このようなコミュニティの研究としては，コミュニティに対してラベル付けをし，コミュニティの検索を行うもの [4] や，コミュニティ間の関連性を導出し，関連のあるコミュニティ同士を結ぶことでコミュニティ間の関係性を表す地図を作成するウェブコミュニティチャート [5] などがある．また，イントラネットにおける構造解析の研究としては [6] などがある．

3. 多次元制約機構

3.1 アイテムセットキューブ

本研究室では，いつ，どこで，誰がといった多次元制約の下で起きている事象の組であるアイテムセットを一定の閾値以上で高頻度アイテムセットとして求め，データキューブモデルのセルに格納したアイテムセットキューブシステムを提案している [7]．これを用いることで，データキューブ機構の持つロールアップ計算などを用いて，分析条件の変化に高速に対応し，そこでなにが起きているのかを分析することが出来る．

3.2 Web 構造マイニングへの適用

現在 Web 構造マイニングの分野では，完全 2 部グラフをコアと呼び，コアに基づいた Web コミュニティ解析の研究が行われている．完全 2 部グラフ計算を，高頻度アイテムセット計算と対応づけることで，一定以上の大きさの完全 2 部グラフを高頻度アイテムセットとして求めることができる．そこで我々は，リンク構造データに対しアイテムセットキューブを用いて，多次元データマイニングによる Web 構造計算を行った [1]．イントラネット型 Web 空間は階層構造を成している．例えば今回分析の対象とした UEC ドメインでは，トップである `uec.ac.jp` の下に，情報分野のドメイン（情報システム学研究科や情報通

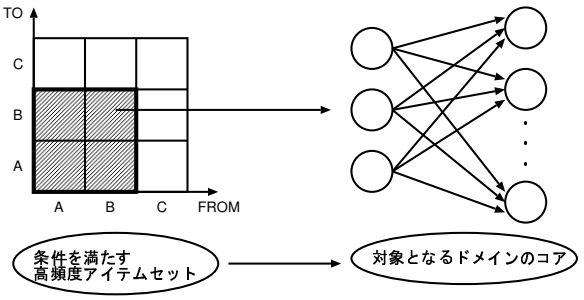


図 1　高頻度アイテムセット計算によるコア抽出

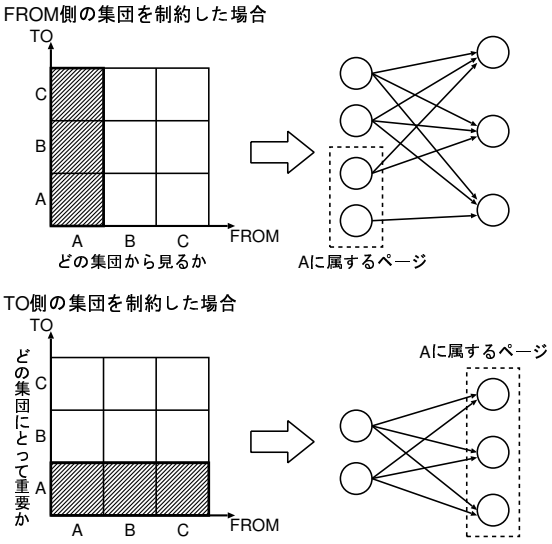


図 2　多次元制約のコアへの影響

信工学科など）や，電気系のドメイン，事務室などのドメインがあり，さらにその下に各研究室ドメインなどが存在するという形になっている．このページの参照関係をもとに図 1，2 に示すように，どのドメインのページ集合からリンクが張られているかに着目した FROM 制約，どのドメインのページ集合に対しリンクを張っているかに着目した TO 制約を行うことで，どのドメインから見るか，どのドメインにとって重要かといった視点を用いて Web コミュニティ構造解析を行っている．

4. コアコミュニティグラフ

リンク構造データから高頻度アイテムセット計算により求められたコアをもとに，**コアコミュニティグラフ**と呼ぶグラフ構造を作成する．コアコミュニティグラフとは，コア同士をマージして作成したコアコミュニティを 1 ノードとして考え，ノード間の関連性を有向辺で表したグラフのことである．このようなグラフを用いることで，組織間の関連性を調べやすくなり，またこのグラフ構造に基づいたランキングを行うことで重要なコアコミュニティを目立たせ，対象となる Web 空間の分析に役立つと考えている．

4.1 グラフの作成

FROM/TO 制約が与えられたときのグラフ作成は次の通り．

4.1.1 対象コア計算

FROM/TO に該当するドメインのコアをアイテムセットキューブキューブを用いて算出する（詳細は [1] 参照）．

4.1.2 コアのマージ

関連性のあるコアをマージしコアコミュニティノードとする．その際の手続きは以下の通り．

Step1: サポート数 60 以上のコアを UEC ドメイン全体から求める．そこからハブページが 2 以上で，極大なものを取り出す（グローバルコア）．そして UEC 全体のリンク構造から，これらグローバルコアの要素を削除する．

Step2: FROM/TO 制約によって求められたページ集合から，ハブページが 2 以上かつ極大なコアを選ぶ．

Step3: グローバルコアから，与えられた FROM/TO 制約を満たすものを抽出する．

Step4: Step2，Step3 の各コア集合内で，共通するオーソリティページを 2 以上もつコア同士をマージしていき．1 つのコアコミュニティノードとする．

4.1.3 辺の付加則

以上の手続きにより作成されたコアコミュニティノード（以下ノード）に対し，下記の 3 つの場合で辺を追加し，グラフ $G_0 = (V_0, E_0)$ を作成する（V_0 をノードの集合，E_0 を辺の集合とする）．

1. ノード $n_1, n_2 \in V_0$ で共通のオーソリティページを持つ，または共通のハブページを持つ場合には，ノード n_1 と n_2 の間には双方向のリンクが存在するとし，$n_1 \to n_2$，$n_2 \to n_1$ とする．

2. ノード n_1 のオーソリティページがノード n_2 のハブページであるような場合には，$n_1 \to n_2$ とする．

3. n_1 に含まれるページから n_2 に含まれるページへのリンクが存在するとき，n_1 から n_2 への有向辺が存在するとし，$n_1 \to n_2$ とする．

4.2 グラフモデルの変更

上記手続きによりコアコミュニティグラフが作成される．このコアコミュニティグラフのモデルとして，[1], [2] では図 3 に示すようにノードに自己ループを付けたモデルを用いていた．これは自己ループにより内包するページ数に応じた滞留を起こさせることで，ノードに含まれるページ数の違いを反映させるためであった．しかし，実際にこのモデルで作成したグラフを用いてノードのランキングを行いランク上位のノードの分布を調べたところ，グラフの構造を適切に反映したランキングが得られているとは言えない状況が見られた．そこで，現在は図 4 に示すようにノードの自己ループを削除したモデルを採用している．これと後述するランク式の変更により，問題点を解決することを目的とする．

5. ランク式とその改良

作成されたコアコミュニティグラフに対し，ノードのランキングを行う．これにより，どこが重要視されているかを調べる．

5.1 ランク式変更前

まず [1], [2] で用いたランク式を示す．自己ループを付加したグラフモデルに対し Pagerank アルゴリズム [8] を用いてランキングを行う．ノード i のランク値 x_i は下式により求められる．

$$x_i = \epsilon \times \frac{1}{\text{ノード数}} + (1-\epsilon) \times \sum_{j \ s.t. \ (j,i) \in E_0} x_j \times A[j,i]$$

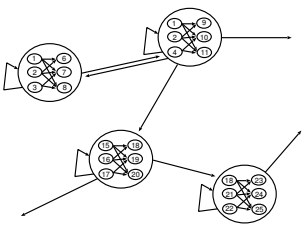


図 3　改良前のグラフモデル

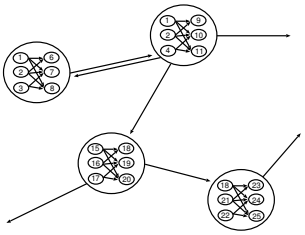


図 4　改良後のグラフモデル

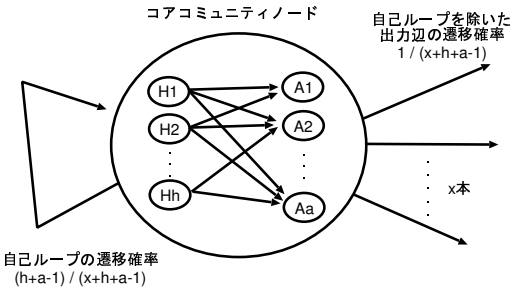


図 5　改良前のノードモデルと遷移確率

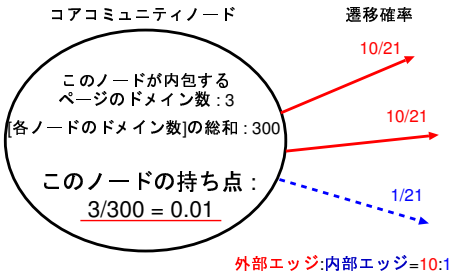


図 6　改良後のノードモデルと遷移確率

ここで $A[j,i]$ は，ノード j からノード i への遷移確率，ダンピングファクタ ϵ は 0.15 とした．また，ランク計算において，異なるノード間の辺に 1 本でもサイト外リンクが含まれていれば外部エッジ，全てサイト内リンクであった場合には内部エッジとした．内部エッジの場合には外部エッジによる参照関係よりも重要度が低いと考え，遷移確率をペナルティとして 1/10 倍している．ここで削減された分の遷移確率は自己ループの遷移に加算することで，全体の遷移確率を 1 としている．しかし前述の通り，図 5 に示すようなノードからの遷移に自己ループを付加したモデルを用いたグラフ構造に対し，このランク式によるランキングを行い，上位ノードをグラフ上にマップし確認したところ，エッジの関係などのグラフ構造にあった結果が得られているとは言えない状態であった．そこで，ランク式の変更を行う．

5.2 ランク式変更後

Pagerank 式の第 1 項にあるように，Pagerank ではダンピングファクタ ϵ により 1/(ノード数) の確率で各ノードに遷移を行うモデルになっている．このアルゴリズムはページ単位のグラフを考えて設定されており，今回対象としているコアコミュニティグラフのノードのように，内部に複数のページのリンク関係を集約したノードモデルは想定していない．そのため今回自己ループをやめ，代わりにノードの持つページ数などを適切に処理する仕掛けを与える．まず最初に考えた方法は，ノード中のページ数をノードの持ち点としたものだったが，データサイ

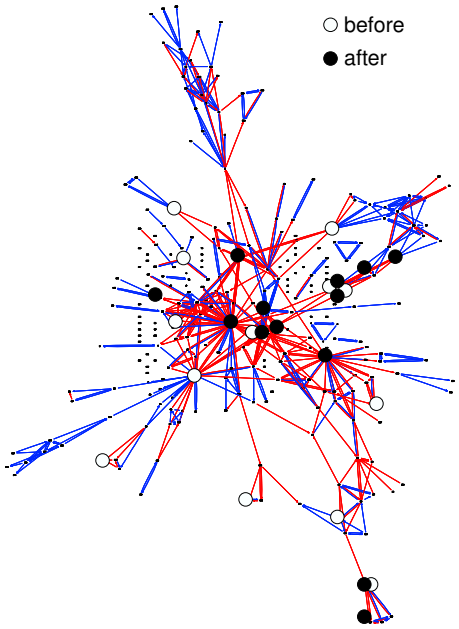


図 7　UEC 全体でのランク式変更前後の上位ノードの位置

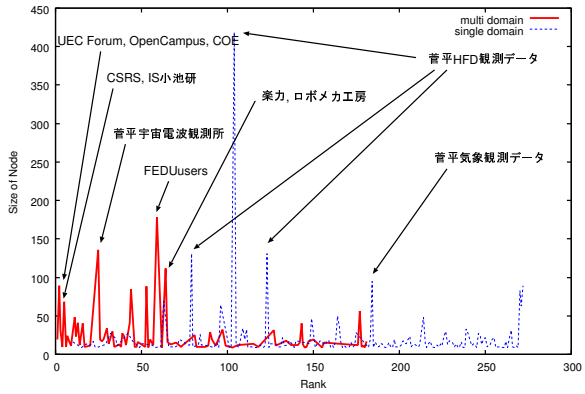


図 8　UEC 全体:ランク式変更後のノードのランクと含まれるページ数の関係

トなどの大量のページを持った孤立点が過度に上位に上がってきてしまうという問題が見られた．そこで，図 6 に示すようにノードが内包するページ数ではなく，ノードの内包するページのサイトドメイン数に応じて遷移するモデルを考えた．つまり，あるノードの持つドメイン数を，各ノードが持つドメイン数の総計で割った値をノードの持ち点とした式とした．

$$x_i = \epsilon \times \frac{s_i}{\sum s_k} + (1-\epsilon) \times \sum_{j\ s.t.\ (j,i)\in E_0} x_j \times A[j,i]$$

ここで，s_i はノード中のサイトドメイン数，$\sum s_k$ が各ノードに含まれるサイトドメイン数の総和を示す．これにより，内包するページのドメイン数が多いノードをダンピングファクタによる遷移確率が高くなるようにした．また，ダンピングファクタ $\epsilon = 0.5$，他ノードへの遷移確率は外部エッジと内部エッジの比率を 10 : 1 とした．

6.　評　　価

6.1　式変更による効果

自己ループの削除と，ランク式の変更により上位にランクされるノードがどの様に変化するか，2005 年 1 月に収集した UEC ドメインのデータ(注1) で評価を行った．対象を UEC 全体（uec.ac.jp 下の全 Web ページ，ただし被リンク数 8 以上かつ全てがサイト内部からのリンクであるページについては，被リンクを削除し，サイト外部へのリンクのみを残す）としたとき，コアコミュニティグラフを作成するとノード数は 272 であった．このグラフ上でのモデル変更前後のランク上位ノードの分布を図 7 に示す．これらを見るとランク式改良前では白点で示されるようにグラフの端にランク上位のノードが多く現れていたが，改良により黒点で示されるようにエッジが密に張られたノードが上位に現れるようになった．特に外部エッジが集まっているノードがランク上位に現れる様になったことから，グラフの構造を考慮するとより適切なランキングになっていると考えられる．このモデル変更前後の上位ノードの内容は表 1, 2 に示す通りである．変更前の上位に見られた個々の研究室のページのみで構成されたノードがランク式変更後は上位に目立ち過ぎるということがなくなった．また式変更後のノードのランクとサイズの関係を図 8 に示す．意図した通り，複数ドメインからなるノード（マルチドメイン）が上位にランクされている．図中でラベリングしたような内部にページを大量に持つノードの内容を見てみると，マルチドメインノードはフォーラムや Computer Security Research Station(CSRS) など活発に活動していそうなノードを見ることが出来るが，単一のドメインからなるノード（シングルドメイン）では多くがデータサイトのノードであった．このようなシングルドメインノードも価値がないとは言えないが，外部との繋がりを集約したマルチドメインノードの方が情報としては良いと考えられる．

6.2　FROM 制約と TO 制約を用いた分析

ドメインを制約して分析することにより，全体で分析したときには分からない，そのドメインの特色を表す結果が得られると考えている．その効果を見るため UEC ドメインを IS/C/J 学科で構成される情報学科部門と，EE 学科，そしてその他の学科や事務室などから成る OTHER という 3 つのドメインに分け，FROM と TO の関係にあるページのドメインに着目して分析を行った(注2)．

ここでは FROM 制約を OTHER とした場合と TO 制約を OTHER とした場合の分析結果を示す．各条件で作成されたコアコミュニティグラフを，図 9，10 に示す．これらのグラフのノードをランキングすることにより，OTHER から見て重要なノード，OTHER にとって重要なノードがわかる．先ほどの UEC 全体での結果である表 2 と比較すると，ドメインを制約行うことによりランクを上げるノードや，全体では見られなかった新たに現れるノードなどを調べることが出来た．また，FROM 制約と TO 制約の違いを見ていくと，図 9，10 の上位 20 ノードのうち 7 個のノードが入れ替わっている．例えば，FROM を OTHER に制約した場合には，3 位の HC 専攻関連のノードや

(注1)：ページ数 108,631，サイト内リンク数 521,306，サイト外リンク数 11,401
(注2)：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．

表 1　UEC 全体での上位ノードの内容 (変更前)

rank	content
1	ICE 教育計算機室 (ied) 関連
2	菅平宇宙電波観測所関連
3	CSRS, IS 小池研
4	elecon.ee.uec.ac.jp
5	事務室関連 (シラバス等)
6	EE 木村・一色研
7	PC 機能物理化学研究室 (simulation)
8	EE 情報処理教育システム
9	IS シンポジウム
10	EE 実験工学研究室関連
11	IS 箱崎研 (members)
12	IS 弓場研
13	IS 箱崎研 (yuka)

表 2　UEC 全体での上位ノードの内容 (変更後)

rank	content
1	EE 授業ページ関連
2	UEC フォーラム, オープンキャンパス, COE
3	事務室関連 (シラバス等)
4	EE 授業ページ関連
5	CSRS, IS 小池研
6	CSRS, IS 小池研
7	EE 情報処理教育システム関連
8	EE 授業ページ関連
9	EE 実験工学研究室
10	ICE 高橋弘太研関連
11	ICE 三木・来住研関連
12	楽力, MCE 専攻
13	実験工学研究室関連

12 位の SE 人間・知識システム学講座のノードが上がってきている．これらのノードはいずれも，TO を OTHER に制約した場合には 40 位以下であった．そのため制約を変えたことにより目立つようになったノードであるといえる．言いかえると OTHER ドメインにとっては重要なノードであるが，他のドメインからは参照されにくいノードであるといえる．ランクが上がった要因をグラフから考えると，これらのノードは FROM を OTHER に制約した場合には，1 位である OpenCampus らのノードから 1 ホップ，または 2 ホップのところに位置していて，TO を OTHER に変更することによりこのノードが分離したことの影響を受けたことなどが考えられる．特に SE 関連のノードは TO を OTHER に制約した場合には孤立点になっているためその影響が大きく出ている．また，TO を OTHER に制約すると順位を上げてくるノードには大きな変化を示したものはないが，グラフの変化を受けランキングは変化している．このように，FROM 側から見るか，TO 側から見るかの違いでグラフが変化し，それにより分析結果が変わってくる．制約条件を与えることで，事なる視点でリンク構造の分析を行うことが可能であることが示された．

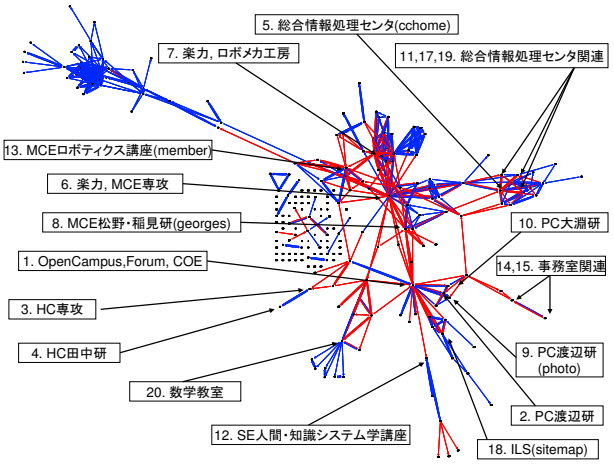


図 9　FROM を OTHER に制約したときの結果

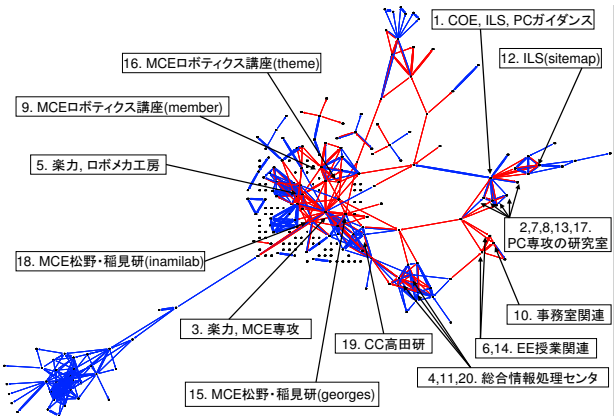


図 10　TO を OTHER に制約したときの結果

7.　柔軟な問い合わせの機構の試作

ここまでは，Web ページの属するドメインに注目し，どのドメインのページと参照関係にあるか FROM/TO という制約を用いた分析を行ってきた．しかし，このドメイン情報を用いた分析だけでは，Web をアーカイブ化したときの分析属性として充分でない場合が考えられる．そこで，従来の分析属性に加え，1) 時間軸領域での変化，2) 制約範囲の調整，3) キーワードを用いた制約といった視点を取り入れて分析を行った．これらを用い，分析条件の調整を行い詳細な分析を目的とする．

7.1　時間軸領域での問い合わせ

Web 空間は日々変化している．この変化を捉えることで，コミュニティ解析で組織の変化の調査を行うことは重要であると考えている．これより Web リンク構造のスナップショットをアーカイブ化しておき，時間軸領域での問い合わせを行うための手法の導入を行った．

7.1.1　Jaccard 係数を用いた調査

時間が経過することにより，構成されるコアコミュニティノードの状態がどう変化しているかを知りたい．具体的には，新たに出現したノードがあるか，消えたノードはあるか，またノードを構成するページの入れ替わりなどの変化を見たいという要求を考えている．このような問い合わせに対し，時間経過前後のノードの類似性を Jaccard 係数を用いて調査を行う．す

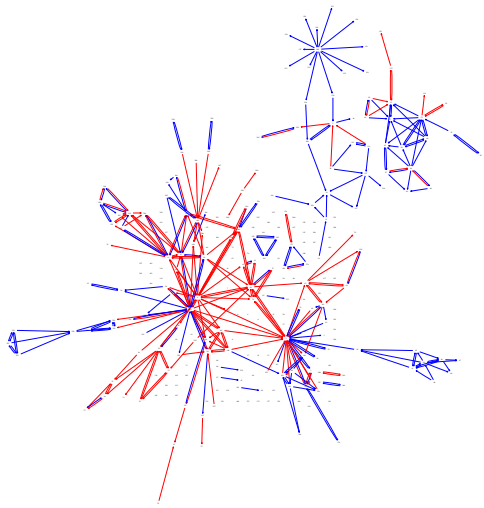


図 11　UEC 全体で作成したコアコミュニティグラフ (2006 年 10 月)

表 3　2006 年 10 月時点での UEC 全体の上位ノードの内容

rank	content
1	数学教室
2	菅平宇宙電波観測所
3	EE オープンキャンパス関係
4	数学教室 (jant)
5	EE 授業関連, EE 桑田研
6	IS 前川研
7	数学教室, 事務 (profile)
8	IS 前川研
9	EE 桑田研関連
10	EE 電子工学工房
11	MCE 森重研 (5axis)
12	MCE 森重研 (6axis)
13	COEcoherent
14	MCE 森重研 (5axis)
15	電気通信大学地域・産学官連携推進機構

なわち，時刻 t_1，t_2 に存在しているノード集合 X、Y の間で Jaccard Similarity Join を行い，総当たりで Jaccard 係数を求める．このとき Jaccard 係数はノード x，y に内包されるページの関係より．

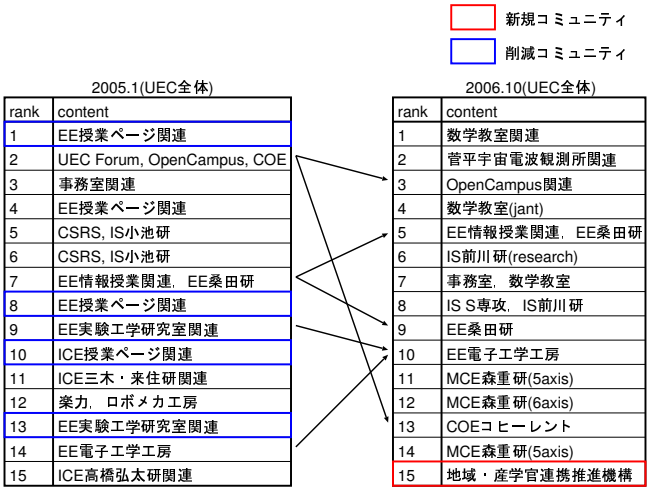
$$Jaccard(x,y) = \frac{|x \cap y|}{|x \cup y|}$$

で求められる．求められた Jaccard 係数より，類似ノードと見なす基準を 4% 以上として調査を行った．

7.1.2　分析例:UEC 全体での変化

ここでは，2006 年 10 月に収拾したデータ(注3) で分析を行った結果を用い，2005 年 1 月からの変化の調査を行った．2006 年 10 月のデータで UEC 全体で分析を行った結果を図 11 と表 3 に示す．2005 年 1 月の分析結果の表 2 と比較してみると，時間の間隔があったこともあり，上位ノードの内容は大きく変化している．このことから，この期間に多くのページが作成，削除されているのではないかと推測することができる．ここ

(注3)：ページ数 111,646，サイト内リンク数 582,206，サイト外リンク数 23,385



rank	content
1	EE授業ページ関連
2	UEC Forum, OpenCampus, COE
3	事務室関連
4	EE授業ページ関連
5	CSRS, IS小池研
6	CSRS, IS小池研
7	EE情報授業関連，EE桑田研
8	EE授業ページ関連
9	EE実験工学研究室関連
10	ICE授業ページ関連
11	ICE三木・来住研関連
12	楽力，ロボメカ工房
13	EE実験工学研究室関連
14	EE電子工学工房
15	ICE高橋弘太研関連

rank	content
1	数学教室関連
2	菅平宇宙電波観測所関連
3	OpenCampus関連
4	数学教室(jant)
5	EE情報授業関連，EE桑田研
6	IS前川研(research)
7	事務室，数学教室
8	IS S専攻，IS前川研
9	EE桑田研
10	EE電子工学工房
11	MCE森重研(5axis)
12	MCE森重研(6axis)
13	COEコヒーレント
14	MCE森重研(5axis)
15	地域・産学官連携推進機構

図 12　Jaccard 係数を用いた比較

で，この間にどのような変化があったのか調べるため，各時点に存在するノード間で Jaccard Similarity Join を行った．これより，この時間経過の間に新たに出現したノード (created) と，消えてしまったノード (deleted)，ノード同士がマージされたノード (merged)，逆に分割されたノード (splitted) の数は以下に示す通り．

- created 96
- deleted 101
- merged 30
- splitted 31

これらのノードの内容を見ていけば，この期間の間に活動を開始，活性化させたノードや，逆に活動を縮小させていったノードを見ることができる．特に上位ノードの間で比較を行った結果を図 12 に示す．図中で赤で囲ったノードは時間経過により新たに出現したノード，青で囲ったノードは消滅したノード，また矢印はマージ，スプリットの様子を示している．これより，表 3 で 15 位の地域・産学官連携推進機構のノードが新たに出現したこと，また，オープンキャンパス関連のノードがスプリットしていった様子を見ることが出来た．

このように，Jaccard 係数による時間軸での変化を捉えることで，仮想組織の状態の変化を捉えることが容易に行えることが示された．

7.2　ホップ数を用いた制約範囲の調整

ここまでは，ページ間の直接の参照関係をもとにドメイン制約を行い，分析対象を求めていた．しかし，直接のリンク関係はなくても，中間となるページを介して参照関係にあるページも重要である場合が考えられる．そこで，ページの参照関係を元に，N ホップまでの参照関係にあるリンクレコードを対象として分析を行うことで結果がどのように変化するか調査を行った．

ここでは，2006 年 10 月のデータを用い FROM 制約を情報部門としたときに，これまで通り直接のリンク関係（1 ホップ）にあるページを対象した場合と，情報部門から 5 ホップまでを対象にした場合を比較する（ただし，グローバルコアはともに 1 ホップで作成）．このときの結果を図 13 に示す．この結果よ

情報部門から1ホップ

rank	content
1	CS専攻
2	ICE離散数学授業ページ
3	IS研究科
4	菅平系の研究室，事務室
5	ICE離散数学授業ページ
6	ICE福田研関連
7	IS木村研(research)
8	ICE太田研関連
9	ICE暗号理論授業ページ
10	CSRS，IS小池研

情報部門から5ホップ

rank	content
1	旧電子情報学経営システム学講座
2	ILS関連
3	ILS(archive center)
4	EE知能システム学講座
5	EE知能システム学講座
6	CS専攻
7	菅平系の研究室，事務室
8	ICE福田研関連
9	IS研究科
10	ICE離散数学授業ページ

図 13　情報部門を起点としてホップ数を調整した時の変化

り，1 ホップの場合に現れるノードは情報部門に所属しているノードしかなかったが，5 ホップまでのリンク関係を用いて分析することで図 13 右の 2，3 位に現れている ILS（レーザー研究センター）関連のノード，また 4，5 位に現れている EE 知能システム研など情報部門から見て関係が深そうな他ドメインのノードなども見られるようになった．また，情報部門のドメイン制約を緩和することにより上位に上がって来たノードとして，11 位に IS シンポジウムというノードも新たに見られた．

7.3　キーワードによる制約

目的に応じたページ集合で構成されるコミュニティの分析を行いたい場合がある．ドメイン制約でも細かくドメインを分けることにより，ある程度対象を絞った分析を行うことは出来るが，目的に応じて対象を任意に定める方法としてはあまり適していない．そこで，より目的に応じたページ集合に対して問い合わせを実現するため，キーワードヒットするページ集合を Seed ページとした分析を行った．その際，Seed ページの直接のリンク関係のみを用いた場合には，ページ数が少なすぎてコミュニティ分析がうまく機能しない場合があると考えられるため，前小節同様 Seed ページを起点として N ホップリンクを辿ることで到着するページまでのリンクレコードを対象とし分析を行った．

7.3.1　分析例 1:“コヒーレント光科学” または “COE プログラム”

電気通信大学では “コヒーレント光科学の展開” が 21 世紀 COE プログラムとして選出されている．ここでは “COE プログラム” または “コヒーレント光科学” というキーワードを用いて分析を行った．FROM 制約によりキーワードを含むページ集合から 5 ホップまでの Web 空間を対象にして分析した結果を図 14 に示す．ノードの内容を見ていくと，図 14 左のプロジェクトのノードを中心とした，関連した研究を行っているノードの集団，またオープンキャンパスなどの研究公開のノードを見ることができた．右下には大学による COE プログラムを取得したことを報告するページや，シンポジウム等の開催報告のページを持つノードが見られた．ドメイン制約における分析でも，COE コヒーレントプロジェクトのノードは出てきており漠然とその活動を見ることはできたが，今回明示的に Seed ページとして指定することで，関連する活動としてどのようなノードがあり，どう繋がりを持っているか調査を行うことができた．

また，このキーワードヒットするページ集合に関連する活動

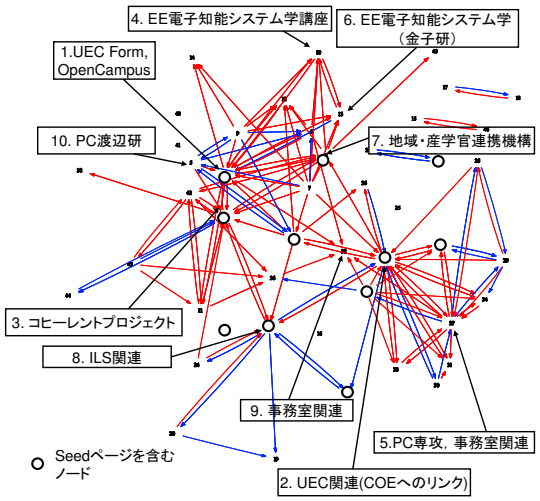


図 14　キーワードを用いた制約での結果（“COE プログラム” または “コヒーレント光科学” にヒットするページ集合から 5 ホップ）(2006 年 10 月)

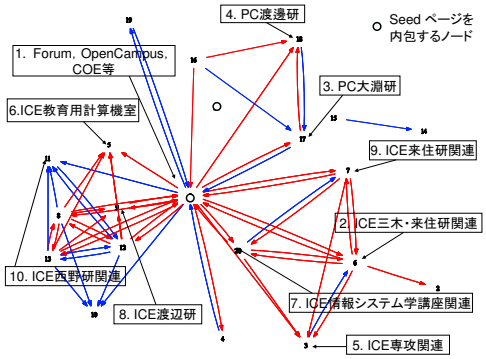


図 15　“COE プログラム” または “コヒーレント光科学” での分析 (2005 年 1 月)

が時間経過により，どのように変化したか調査を行った．2005 年 1 月に収集したリンク構造データを用いて，同様の制約で作成したコアコミュニティグラフを図 15 に示す．2006 年 10 月の図 14 と比較し，コミュニティノード数の増加やノード間の繋がりの変化をみることでコミュニティが時間経過によって発展していった様子を読み取ることができた．

この様に，キーワードを用いた分析により，ドメイン制約とは異なるページ集合を対象とした分析を行うことができ，より詳細なコミュニティ解析を行うことが可能になった．

7.3.2　分析例 2:“ロボット” によるキーワードランキング

キーワード “ロボット” にヒットするページを Seed とし，FROM 制約を用いて 5 ホップまでのリンク関係にあるレコードを対象として分析を行った．このとき作成されるコアコミュニティグラフは “ロボット” のノードと，それに関連する活動で作成されている．これを利用して Seed ページを内包する活動の調査を行った．つまり作成されたコアコミュニティグラフを利用し，Seed を含むノードのランキングを行った．ここではこのランキング結果を一般的な技術との比較として，Google University Search [10] を利用して UEC ドメインの “ロボット” にヒットする上位ページと比較を行った．その結果を図 16，17

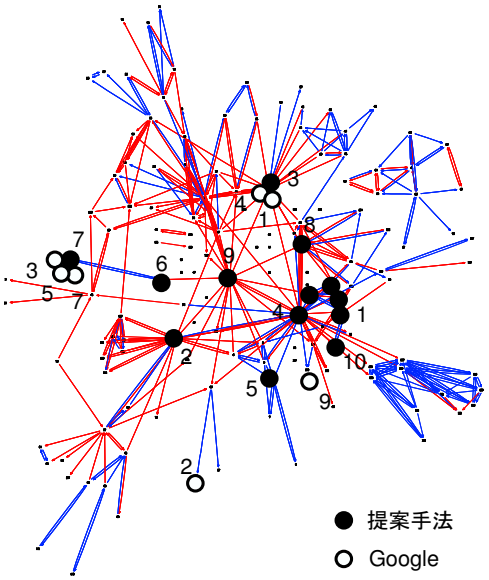


図 16 “ロボット” をキーとしたときの分析

Google

rank	content
1	大道芸ロボットリスト(ロボメカ工房)
2	ロボットシステム（IS木村研)
3	松野卒論(CRC田口研)
4	ロボメカコンテスト(楽力)
5	大西卒論（CRC田口研)
6	球面Scaraロボット(MCE明研)
7	青木卒論(CRC青木研)
8	平面Scaraロボット(MCE明研)
9	楽器演奏ロボット(MCE梶谷研)
10	田口研究室の歴史(CRC田口研)

Google University Search (電気通信大学)
http://www.google.co.jp/univ/ja/uec?hl=ja

我々のコミュニティ検索

rank	content
1	MCE稲見研関連
2	研究室公開(EE, MCE)
3	ロボット・エレクトロニクスコンテスト(菅平)
4	MCE専攻関連(OpenLab)
5	CC高田研関連(GlueLogic)
6	CRC田口研(卒研関連)
7	CRC田口研(卒研関連)
8	楽力, ロボメカ工房関連
9	地域・産学官連携機構関連の研究室
10	MCE松野・稲見研(georges)

（※上位の稲見研のミラーは 1位に併合）

図 17　Google との比較

に示す．図 16 中の黒点が我々の提案手法での上位ノード，白点が Google University Search で上位だったページを含むノードの位置である．ただし，グラフ中に該当するページがなかった場合は図中には示していない．我々のコミュニティ検索では MCE 稲見研やロボット関係の研究室を持つ MCE や EE 学科の研究室公開などの活動が上位に見られた．一方，Google の結果では，共同研究センター（CRC）に所属する研究室の卒業研究に関するページなどが上位に現れている．これら Google での上位ページの分布を図 16 で見ると，ほとんどがグラフの端のノードの活動であり，ロボットのコミュニティとして中心的なものではなかった．この結果に見られるように，UEC ドメインにおけるロボット関連の重要度の分析という点で，提案手法の提供するランキングの価値は高いと言える．

ここで示したような視点を変えたパーソナライズランキングは，これからの情報があふれる時代において重要性が高まって行くと考えられる．本提案手法では，各制約は多次元キューブで実装されるためパーソナライズ化の制約の変化に対して高速に対応することが可能である．これらの属性を組み合わせることにより, [9] のように詳細に条件を定めて Web 空間への Query を実行することが，コミュニティ構造解析の場合でも可能になると考えている．

8. おわりに

本稿では，電気通信大学ドメインでの分析をもとに，多次元データマイニングによる制約を用いたイントラネット型 Web 空間の分析の有効性評価と，属性追加による詳細な分析を用いた評価を報告した．

まず，従来のランキングアルゴリズムが抱えていた問題の改善を行い，本手法による分析によりどの程度の情報を得られるか再評価を行った．次に，ドメイン制約による分析だけでは，詳細な分析が難しいことから 1) 時間軸の追加，2) 制約範囲の調整，3) キーワード制約の 3 属性を用いた分析を行いその有効性を確認した．時間軸の追加では，Jaccard 係数を用いてノードを構成するページの変化を調べることで，時間経過によるコミュニティ変化の調査を行えることを示した．制約範囲の調整では，直接の参照関係からはわからなかった組織の繋がりなどを調べられることを示した．キーワード制約では，ドメイン制約では難しかった任意の Seed ページ集合を元にした分析を行った．これにより，調査を行いたい対象の絞りこみが行い易くなり，組織の重要性や，繋がりを詳細に分析を行うことが可能であることを示した．また，ドメイン制約のような組織構造の事前知識が必要なくなり，様々な仮想組織に対し分析を行うことが容易になった．

今後の課題としては，コア計算により分析対象から外れてしまったページを含めた分析や，Blog などへの対応，さらなるランクアルゴリズムの改良などを考えている．

文　献

[1] 山下 由展, 大森 匡, 星 守, “多次元データマイニングを用いた Web 空間の構造解析,” 電子情報通信学会 DEWS2006, 3B-o3, 2006.
[2] 林 和宏, 大森 匡, 山下 由展, 星 守, “多次元データマイニングを用いた Web 空間の構造解析の評価,” FIT2006, D-019, 2006.
[3] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, “Trawling the Web for emerging cyber-communities,” WWW8/Computer Networks, Vol.31(11-16), pp.1481–1493, 1999.
[4] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, “Extracting large-scale knowledge bases from the web,” In Proc. of the 25th VLDB Conference, pp.639–650, 1999.
[5] 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-1 Vol. J87-D-1 No.2, pp.256–265, 2004.
[6] 大塚 浩司, 大町 真一郎, 阿曽 弘具, “ウェブコミュニティ内のページ・リンクから成る階層構造の抽出,” 電子情報通信学会, WI2-2006-33, pp.123–128, 2006.
[7] 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,” 電子情報通信学会, DEWS2005, 3C-i10, 2005.
[8] 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, 2003.
[9] Srairam Raghavan, Hector Garcia-Molina, “Complex Queries over Web Repositories,” In Proc. of the 29th VLDB Conference, pp.33–44, 2003.
[10] Google ユニバーシティ検索, URL:<http://www.google.co.jp/intl/ja/options/universities.html> .