D_019

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

Evaluation of Mining Web Structures Using Multi-dimensional Data Mining Model

林 和宏† 大森 匡† 山下 由展† 星 守†
Kazuhiro Hayashi Tadashi Ohmori Yoshinobu Yamashita Mamoru Hoshi

1. はじめに

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．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による多次元制約機構の上で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]．本論文では，電気通信大学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出来ているのか評価し報告する．

2. 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を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．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即座に分析することが出来る．

2.1 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部グラフをコアとしたWebコミュニティ解析の研究が行われている[2][3]．一定以上の大きさの完全2部グラフは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]．ここでは，図1に示すように，リンクのFROMとTOの部分に注目し，どのドメインからみるか，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．

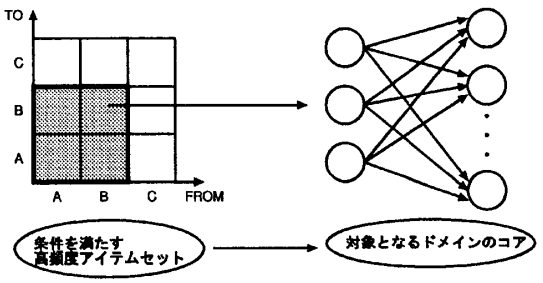

図1: 高頻度アイテムセット計算によるコア抽出

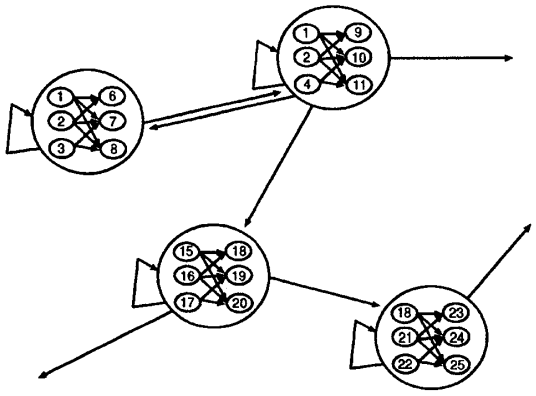

図2: コアコミュニティグラフ

3. コアコミュニティグラフ

リンク構造データから求められたコアをもとに，コアコミュニティグラフの作成を行っている．コアコミュニティグラフとは，図2に示すようにコアコミュニティを1ノードとして考えたグラフのことである．このようなコミュニティ間の関連性の研究としてはウェブコミュニティチャート[3]がある．これを用いることで，組織間の関連性を調べやすくなり，さらに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．

3.1 グラフの作成

3.1.1 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全体から，これらグローバルコアの要素を削除する．

Step2: FROM，TOを制約した集合から，ハブページが2以上かつ極大なコアを選ぶ．

Step3: また，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選ぶ．

Step4: Step2，Step3の各集合内で，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．1つのコアコミュニティノードとする．

†電気通信大学大学院 情報システム学研究科

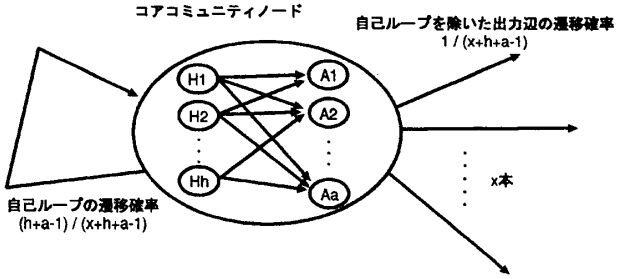

図 3: ノードからの遷移確率

3.1.2 辺の付け方

以上の手続きにより作ら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か，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，$n_1 \rightarrow n_2, n_2 \rightarrow n_1$ 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，$n_1 \rightarrow n_2$ 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，$n_1 \rightarrow n_2$ とする．

4. ランキング

グラフ G_0 に対し，Pagerank アルゴリズムを用いてノードのランク計算を行う．これにより，どこが重要視されているかを調べる．このときノード i の Pagerank x_i は下記で定義される [4]（$A[j,i]$ は，ノード j からノード i への遷移確率）．

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

ここで $\epsilon = 0.15$，各ノードの初期値は1とした．今回はコアコミュニティをノードとしているため，図3に示すように内部への遷移も考慮している．また，ランク計算において，相異なるノード間の辺にひとつでもサイト外リンクが含まれていればその間の遷移確率を1倍，全てサイト内リンクによる辺の場合にはペナルティとして1/10倍している．このとき減らされた9/10の遷移確率は自己ループに足している．これにより，ノードからの遷移確率全体を1にしている．

5. 評価

以上の手順で求められた，コアコミュニティグラフを用いて分析を行った．今回用いたデータは，2005年1月にクローラで収集した電気通信大学のリンク構造データである．ここでは，UEC全体の場合，及びFROM, TOをIS/C/J OTHERの学科サブドメイン群に制約した場合について分析を行った．各条件の下でコアコミュニティグラフを作成したところ，ノード数はそれぞれ270，276であった．ノードには単一サイトのページのみで構成されるノード（single）と，複数サイトのページから構成されるノード（multi）がある．図4，5に，これらのノードのランクと，ノードを構成するページ数の関係を示した．それぞれの場合でのランキング上位5%にあたる，13位までの内容を表1，2に示した．また，図6にFROM, TOをIS/C/J OTHERに制約したときのコアコミュニティグラフを示した．図4，5を見ると，上位に現れているノードはmultiが多いことがわかる．これは，ランク計算のところも述べた，ノード間のリンクにひとつでも異なるサイト間でのリンク関係がある場合にはペナルティを与えないとしたことが影響していると考えられる．

また，ランキング結果を見てると，UEC全体では菅平宇宙電波観測所や電気通信大学歴史資料館，ISシンポジウムなどが上がってきた．これに対し，FROM, TOをIS/C/J OTHERに制約した場合には，全体としては目立っていなかった，楽力，ロボメカ工房関連や

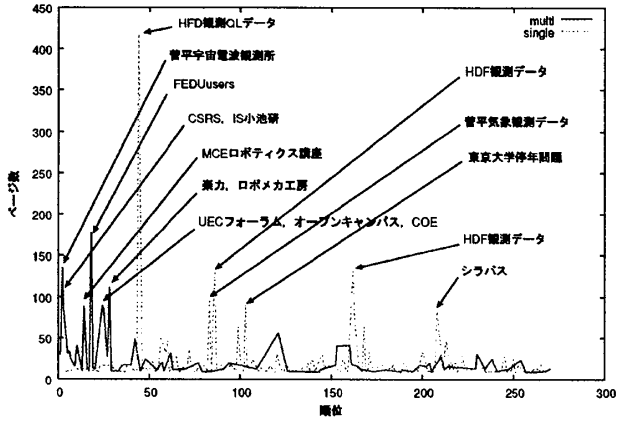

図 4: single, multi のノードに含まれるページ数（UEC全体）

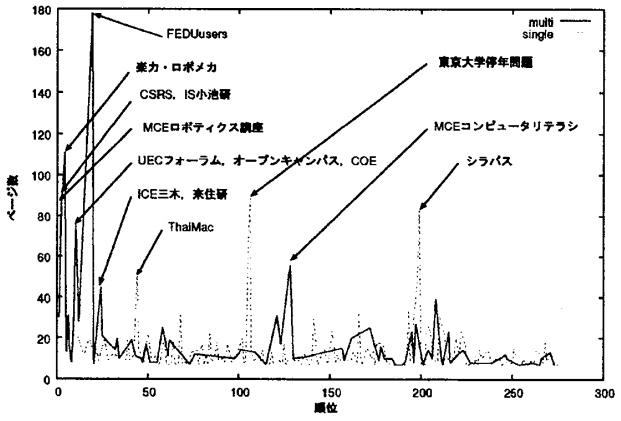

図 5: single, multi のノードに含まれるページ数（FROM, TO を IS/C/J OTHER に制約）

表 1: UEC 全体での上位 5%

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	電気通信大学歴史資料館
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: FROM，TO IS/C/J OTHER での上位 5%

rank	content
1	ICE 教育計算機室 (ied) 関連
2	MCE ロボティクス講座 (下条・明研) 関連
3	CSRS, IS 小池研
4	楽力, ロボメカ関連
5	PC 渡辺研, 教務課
6	IS 韓・長岡研, ICE 小林, ICE 川端研
7	総合情報処理センタ関連
8	ICE 専攻, シラバス
9	PC 林研, CC コンピュータ演習
10	UEC フォーラム, オープンキャンパス, COE
11	FEDU SAIYAI
12	IS 曽和研，IS N 専攻
13	ICE 高澤研

6 位の情報理論関係の研究室の集まりといったものや，UEC フォーラム，オープンキャンパス，COE というノードを見ることが出来た．さらには，SAIYAI といった留学生のプロジェクトが見られた．特に情報理論関係のノードは UEC 全体では 173 位であるため，全体の活動からは捉えることは難しい．この様に，制約条件を与えることで，異なる視点でリンク構造の分析が行え，これにより詳細な分析が可能であることが分かった．

自己ループを付けることによる効果を見るため，図 7 に FROM，TO を IS/C/J OTHER に制約し，自己ループなしでランク計算を行った場合のグラフを示した．図 5 と比較してみると，構成するページ数の多いノードの多くが，自己ループを付けることにより，ランクを上げていることが分かる．自己ループの性質上，これは予想されることである．ある程度の大きさのノードが，活発な活動を表しているとすれば，自己ループを付けたことにより，分析結果はより良いものになっていると考えられる．

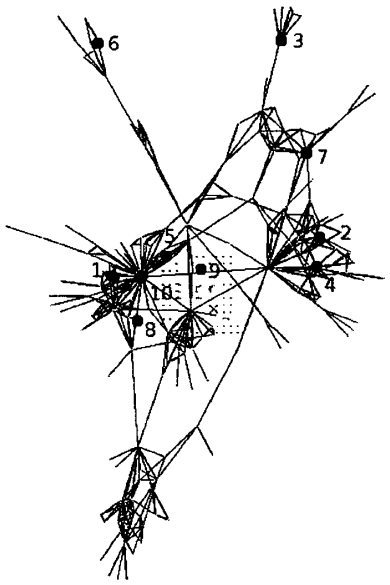

図 6: FROM，TO を IS/C/J OTHER に制約した場合（数値は順位を表している）

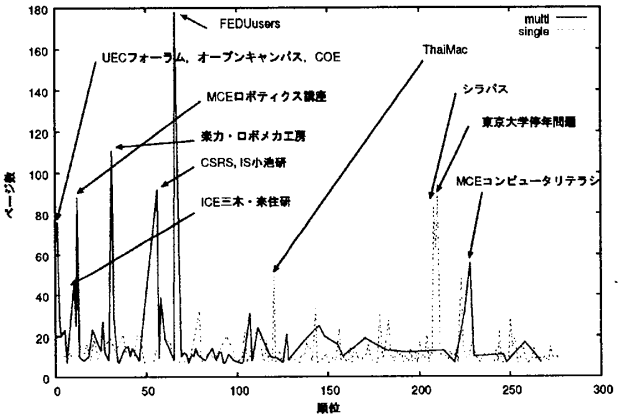

図 7: FROM，TO を IS/C/J OTHER に制約した場合（自己ループなし）

参考文献

[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] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, “Trawling the Web for emerging cyber-communities,” WWW8/Computer Networks, Vol.31(11-16), pp1481-1493, 1999.

[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-1 Vol. J87-D-1 No.2, pp256-265, 2004.

[4] 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.