多次元データマイニングによるWeb空間の構造分析の評価

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

林 和宏[1]　大森 匡[2]　星 守[3]

Kazuhiro HAYASHI　Tadashi OHMORI
Mamoru HOSHI

著者らは, 多次元制約下でデータマイニングを行う機構「アイテムセットキューブ」を試作しており, 昨年, 本機構に基づいて, イントラネット型の Web 空間の構造分析を試行した [1]. 具体的には, 大学ドメインの学科別の視点という多次元制約を用いて, Web 空間内のコア計算と，コアに基づいたグラフモデルを作成し, その中でノードのランキングを行って, 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，本提案方式におけるランク計算式の改良と, その構造分析能力の評価を述べる.

In our previous study[1], the authors applied multi-dimensional data-mining to web structure mining, and proposed a new graph-structure called a core-community graph so as to find interesting web-communities in an intranet domain. This paper improves a ranking algorithm in this graph model and evaluates effects of our web-community mining method.

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 [2][3]．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，データキューブの考えに沿って多次元制約下のデータマイニングを行う機構「アイテムセットキューブ」を試作している [4]．これは, データキューブモデルの最小立方格子 (セル) に, 制約を満たすレコード集合から求まる高頻度アイテムセットを格納したものであり, 実体化やスライス, ロールアップなどの演算によって多次元制約下のアイテムセット分析を効率良く行う機構である. 昨年, 著者らは, この多次元制約下のデータマイニング機構に基づいて, イントラネット型の Web 空間の構造分析 (コミュニティ分析) を行った [1]. すなわち, Web 構造分析でいうコア (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とし, アイテムセットキューブの考えに沿って, 用意した多次元制約下でコアを求め, そこから, コアに基づいたコミュニティ間の関連を表すグラフモデル (コアコミュニティグラフ) を作成し, コミュニティを表すノードのランキングを行った [1]．その結果, 多次元制約下の分析により, 制約なしの場合には分からなかった結果も得られ, 手法の有効性を示すことができた．しかし，詳細に調べたところ，提案したグラフの構造とノードのランク値の間に隔たりがある場合が見られた. そこで本稿では，まず, コアコミュニティグラフとノードランキング式のモデルを再考する. 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，提案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する.

[1]学生会員, 電気通信大学, hayashi@hol.is.uec.ac.jp
[2]正会員, 電気通信大学, omori@hol.is.uec.ac.jp
[3]非会員, 電気通信大学

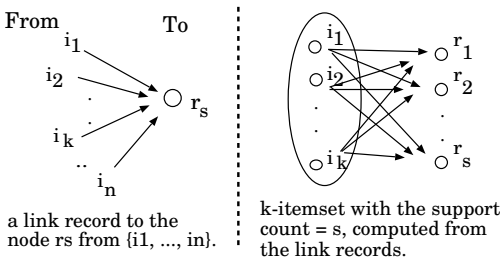


図 1: 高頻度アイテムセットによるコア計算
Fig.1: (k, s)-core as a frequent k-itemset

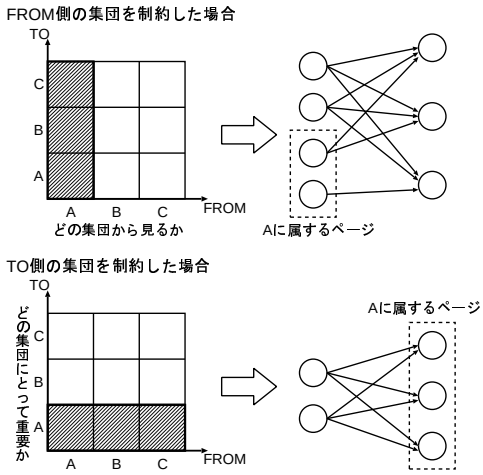


図 2: FROM/TO 制約下で求まるコアの特性
Fig.2: Core under FROM/TO constraints

2. 多次元制約下の Web 構造マイニング

2.1 アイテムセットキューブ

アイテムセットキューブとは, いつ，どこで，誰がといった多次元制約の下で一定頻度以上起きている事象の組 (高頻度アイテムセット) を求めるためのデータキューブモデルであり, 最小立方格子 (セル) に, 対応するレコード集合から求めた高頻度アイテムセットを持つ [4]．これを操作することで, 多次元分析条件に応じた事象の組を効率良く調べることができる.

2.2 Web 構造マイニングへの適用

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 コミュニティ分析が行われる [2]. 一般に, 始点 (ハブ) 数 i, 終点 (オーソリティ) 数 j のコア (つまり, (i, j) コア) 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, アイテムセットキューブを使って, 多次元制約下でのコア計算を行った [1].

具体的には, Web リンク構造データとして, 図 1 左に示すように，あるページ r_s へのリンク入力関係を表すレコードを用意する. つまり, r_s への全リンクの始点ページ i_1,i_2,...,i_n をアイテムとしたレコードを作り, このレコード集合から始点ページに関する高頻度アイテムセットを求める. すると, i_1,i_2,...,i_k を k-アイテムセットと考えているから, これらを (始点ページとして)

含むリンクレコード数は当該 k-アイテムセットのサポート数である. その結果, 計算するサポート数の最小値 ($minsupNum$) を与えたとき，リンクレコード集合 D において高頻度となるアイテムセットを I とし, I を含む全てのレコードの終点ページの集合を A とすると，求まる高頻度アイテムセットは,I をハブ, A をオーソリティとした (I, A) コアになる (図 1 右参照). [1] では, 多次元制約下のコア計算は Apriori アルゴリズムを基本としており, 制約を満たすリンクレコードから, 指定した最小サポート数 $minsupNum$ 以上の authority 数のコアを計算する.

分析対象は, イントラネット型の Web 空間であり, 組織別の階層構造を有す. 今回は, 電気通信大学 (UEC) ドメインの Web 空間を用いた. これは, 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ある.

多次元制約としては, 図 2 に示すように,「どのドメインのページからリンクを張られているか」に着目した FROM 制約 (同図上),「どのドメインのページにリンクを張っているか」に着目した TO 制約 (同図下) の 2 次元を考えた. 例えば, 制約条件「FROM 制約=A」は, ドメイン A のページを始点に持つリンクレコードから計算されたコアを求めており, これは, A から見て重要なコアと見なせる.「TO 制約=A」は, ドメイン A のページを終点に持つコアを意味し, A にとって重要なコアと言える. このように, FROM/TO 制約は, どのドメインから見て/どのドメインにとって重要か, という視点に沿ったコアを求めるものである.

以上の考えに基づき, FROM/TO の多次元制約下で該当するコアの集合を求め, コア集団間のグラフ構造を作成すれば, 多次元的な視点からの Web コミュニティの構造分析を行える [1]. 次節以降, 提案したグラフ構造と分析能力の評価を述べる.

3. コアコミュニティグラフ

与えられた FROM/TO 制約下で求まったコアを使って, **コアコミュニティグラフ**と呼ぶグラフ構造を作成する. このグラフは，強く関連するコア同士をマージして 1 ノード (コアコミュニティノード) とし, ノード間の関連性を有向辺で表したものである. これにより, ノードで表されたコミュニティ間の関連性を調べやすくなる. また, グラフ構造に基づいたランキングを行って, 重要なノードを判定できる.

3.1 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.

3.1.1 ノードの計算方法

FROM/TO 制約下のコアを, 最小サポート数 (authority 数)≥ $minsupportNum$ の条件でアイテムセットキューブにより求め, そこから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(前処理): 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しておく.

Step2: FROM/TO 制約を満たすコアを求める. これらのうち, ハブページが 2 個以上で, かつ, 極大 (他のコアに包含されない) 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選ぶ.

Step4: Step2，Step3 の各コア集合を, 同値関係「オーソリティページを 2 個以上共有する」で同値類にわけ, 各同値類についてコア集合をマージして 1 つのコアコミュニティノードとする.

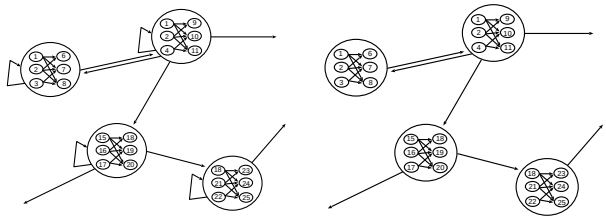


図 3: 改良前のグラフモデル
Fig.3: graph model (old)

図 4: 改良後のグラフモデル
Fig.4: graph model (new)

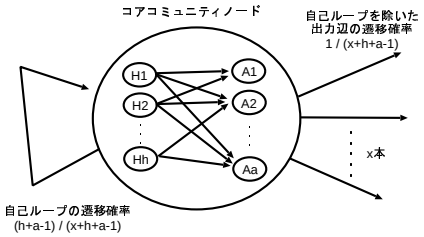


図 5: 改良前のノードモデルと遷移確率
Fig.5: node model (old)

3.1.2 辺の付加則

コアコミュニティノード (以下, ノード) に対し，下記の 3 つの場合で有向辺を与え, グラフ $G_0 = (V_0, E_0)$ を作成する（V_0 をノード集合，E_0 を辺集合とする）.

1. ノード $n_1, n_2 \in V_0$ が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，n_1 と n_2 の間に双方向リンク $n_1 \rightarrow n_2$，$n_2 \rightarrow n_1$ を与える.

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場合には，有向辺 $n_1 \rightarrow n_2$ を与える.

3. ノード n_1 に含まれるページからノード n_2 に含まれるページへのリンクが存在するとき，有向辺 $n_1 \rightarrow n_2$ を与える.

3.2 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] 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場合が見られた. そこで，今回, 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用いる. そして後述するランク式の変更により，グラフ構造にあったランキングをえられるようにする.

4. 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調査する.

4.1 ランク式変更前

まず [1] 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 (図 3 と図 5) に Pagerank アルゴリズムを適用し, ノードのランキングを行った. ノード i のランク値 x_i は次式で与えた:

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

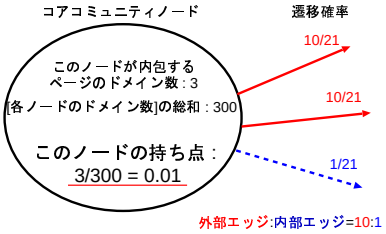


図 6: 改良後のノードモデルと遷移確率
Fig.6: node model (new)

ここで, $A[j,i]$ はノード j からノード i への遷移確率，ダンピング係数 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して, 全体の遷移確率を 1 とした．

3.2 節で述べたように, このランク式は, 図 3 のグラフモデル上では, グラフの辺の密度に応じたランク値にならない場合があった (5.1 節で示す). そのため, グラフを図 4 に変え, ランク式も変更することとした．

4.2 ランク式変更後

Pagerank 式の第 1 項にあるように，Pagerank ではダンピング係数 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．そのため今回は自己ループをやめ，代わりにノードの内包するページ数などを考慮し，適切に処理するモデルを考えた．ノードの持つ情報としては，あるサイトのみで構成されたノードよりも，複数のサイトで構成されたノードを情報として良いと考えている．そこで，図 6 に示すように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，$\sum s_k$ は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係数による遷移確率が高くなるようにした．また，ダンピング係数 $\epsilon = 0.5$，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．

5. 評価

5.1 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ように変化するか，2005 年 1 月に収集した UEC ドメインのデータ[4] を用いて評価した．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，被リンクを削除してサイト外部へのリンクのみを残す) としたとき，作成されたコアコミュニティグラフ（ノード数 272, minsupNum=7）を図 7 に示す．このグラフ上で，今回提案したランク式を用いた場合の上位 13 ノードの位置を図中に黒点で示した. (図中, いくつかのノードには, url 集合から人手で判断した組織名とランク値を添えた)．ランク式変更前では，白点で示されるようにグラフの端に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ところに上位ノードが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．これより, グラフ構造を考慮したより適切なランキングになっていると考えられる．

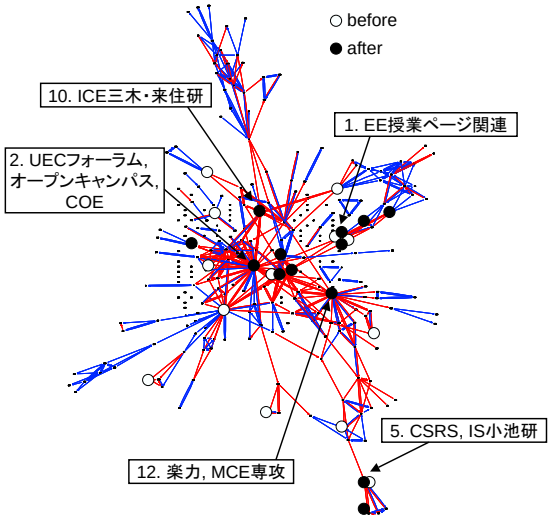


図 7: UEC 全体でのランク式変更前後の上位ノードの位置
Fig.7: Core-community graph (UEC) with top-13 nodes

5.2 FROM 制約と TO 制約を用いた分析

ドメインを制約してコアコミュニティグラフを作りランクづけすると, 制約なしの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はずである. 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[5].

例として, FROM 制約を OTHER とした場合と TO 制約を OTHER とした場合のコアコミュニティグラフ ($minsupNum = 4$) を，各々, 図 8 と 9 に示す. (図中のコメントは上位 20 位までのランク値と組織名である). 例えば, 図 8 は, OTHER ドメインからみた時の条件下での重要なコミュニティを (図 7 より小さいコアまで入れて) 表していて, UEC 全体のランク上位組織からは分からない情報である. また, FROM 制約と TO 制約により OTHER ドメインの重要なノードの違いを見ると, 図 8，9 の上位 20 ノードのうち 7 個が異なっている．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 12 位の SE 人間・知識システム学講座のノードが TO 制約 =OTHER の場合にない上位ノードである. これらは, OTHER から見たときに固有な重要性を持つ組織といえる．

5.3 キーワード制約を用いた分析

最後に, 今回考えたグラフモデルが Web 空間の分析機構と

[4]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. サイト内リンクも使うことに注意.

[5]IS=情報システム学研究科，EE=電子工学科. C=情報通信工学科, J=情報工学科．OTHER には MCE(知能機械工学科) や各種研究所が入る.

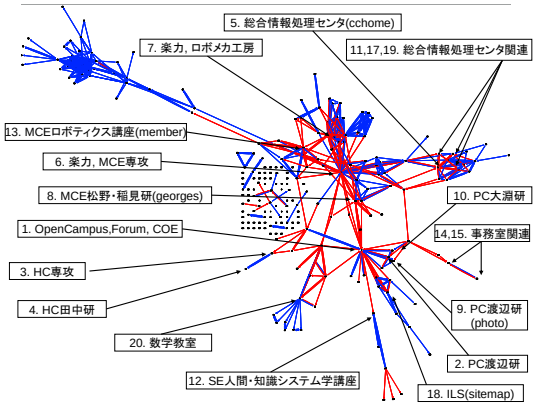


図 8: FROM を OTHER に制約したときの結果
Fig.8: The graph of FROM=OTHER

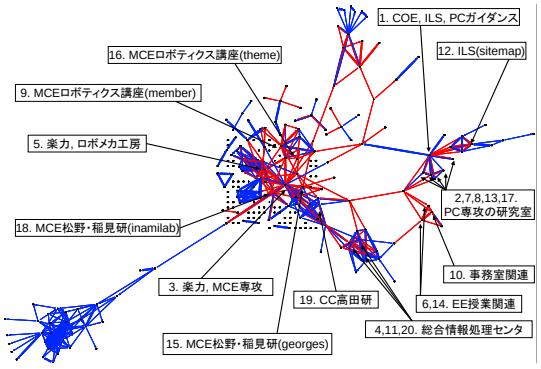
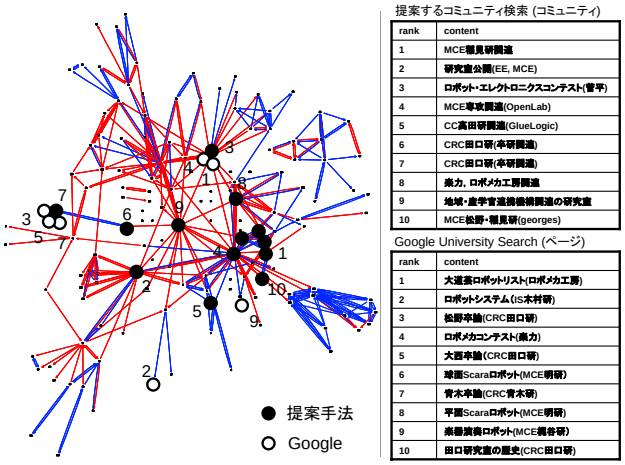


図 9: TO を OTHER に制約したときの結果
Fig.9: The graph of TO=OTHER

してどの程度有効な結果を示すことができるかを調べるため, より一般的な制約条件としてキーワード制約を用いた場合を試行した. すなわち, 特定のキーワードを与え, それにヒットするページ (SEED ページ. 当然, 複数ある) から 5 ホップまでに存在するリンクレコード集合を対象にしてコアコミュニティグラフを作成し, ランキングを行った. キーワードとしては「ロボット」を与えた. 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」にヒットするノード (=SEED ページを含むノード) と, それに関連する活動で構成されている.

2006 年 10 月の収集データ上でのグラフ作成結果を図 10 に示す (最小サポート数 5). 図中の黒点は, SEED ページを含むノードの中でランキング上位 10 位までのものを示す. このランキング結果を, Google University Search (URL:`http://www.google.co.jp/intl/ja/options/universities.html`) による UEC ドメインの「ロボット」にヒットする上位 10 ページと比較した. 図 10 中の白点が 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. 提案手法を用いて分析した場合には, Google のページ単位の分析とは異なる視点での重要なノードが得られていることがわかる. 例えば上位には MCE 関連の研究室で構成されるノードや, オープンキャンパスでの研究室公開, またロボメカ工房などのノードといった UEC ドメインで重要と考えられるロボット関連の活動を捉えられていた. この結果からも, UEC ドメインにおけるロボット関連の組織の重要度を調査するという場合には, 提案手法の



提案するコミュニティ検索 (コミュニティ)

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(本研関連)
7	CRC田口研(本研関連)
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)

Google University Search (ページ)

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	松野本論(CRC田口研)
4	ロボメカコンテスト(楽力)
5	大西本論(CRC田口研)
6	球面Scaraロボット(MCE明研)
7	青木本論(CRC青木研)
8	平面Scaraロボット(MCE明研)
9	垂器演奏ロボット(MCE横谷研)
10	田口研究室の歴史(CRC田口研)

図 10: 「ロボット」をキーとした時のグラフ
Fig.10: the case where a keyword = 'robot'

ようなコミュニティ単位のランキングも有効と考えられる.

6. おわりに

本稿では, 電気通信大学ドメインの Web 空間を例として, 多次元的な制約条件下でのイントラネット型 Web 空間におけるコミュニティ構造の分析評価を行った. 具体的には, 文献 [1] で提案した技法を元に, コアコミュニティグラフの作成方法とランクアルゴリズムを改良し, FROM/TO 制約と呼ぶ多次元的制約下での構造分析における有効性を示した. 現在, アイテムセットキューブによるこうした問い合わせの処理効率化にはまだ制限がある [1][4]. この改良や, キーワード制約など多様な多次元制約の組込みが現在の課題である.

[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] 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.

[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, pp.256–265, 2004.

[4] 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.

林 和宏 Kazuhiro HAYASHI
2007 電気通信大学大学院修士課程了, 工修. Web/XML-DB に関心を持つ. DBSJ 学生会員. 現 (株) 富士通.

大森 匡 Tadashi OHMORI
1990 東京大学大学院博士課程了, 工博. 1994 より電気通信大学大学院助教授. 高性能・高機能 DB を研究. DBSJ 正会員.

星 守 Mamoru HOSHI
東京大学大学院修士課程了, 工博. 1992 より電気通信大学大学院教授. データ構造等を研究. ACM, IEEE 等, 正会員.