DEWS2004 6-B-05

Max-Flowコミュニティグラフとその特徴分析

今藤　紀子†　　喜連川　優†

† 東京大学生産技術研究所
〒 153-8505 東京都目黒区駒場 4-6-1
E-mail: †{imafuji,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　ウェブコミュニティとは，似たようなトピックを持つウェブページの集合を意味する．最大流アルゴリズムを利用して抽出したウェブコミュニティ（以降，Max-Flow コミュニティと呼ぶ）は，シードとして与えたページを中心として内側に密にリンクが存在するページの集合となる．これまでの研究で，同じシードページを与えるとHITS系手法によるコミュニティではトピックが一般的になり，Max-Flow コミュニティではより詳細になる傾向があることがわかった．本論文では，日本国内のウェブグラフデータを用いて，ウェブ空間全域から抽出した Max-Flow コミュニティをノード，コミュニティ間の関連をエッジとしたコミュニティグラフ（これを Max-Flow コミュニティグラフと呼ぶ）を構築し，HITS 系手法を利用して得られたコミュニティに基づく既存のコミュニティグラフと比較し，それぞれの手法で得られたコミュニティの大域的な特徴差を分析する．

キーワード　ウェブコミュニティ，ウェブグラフ，ウェブコミュニティグラフ，最大流アルゴリズム，HITS

Constructing Max-Flow Community Graph

Noriko IMAFUJI† and Masaru KITSUREGAWA†

† Institute of Industrial Science, University of Tokyo
Komaba 4-6-1, Meguro-ku, Tokyo, 153-8505 Japan
E-mail: †{imafuji,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　A web community is a set of web pages created by individuals or associations with a common interest on a topic. Maximum-flow algorithm extracts web communities which are sets of web pages having more links within the community than the outside. We call the web community *Max-Flow community*. Our previous work has characterized the properties of Max-Flow community (i.e., local properties) by comparing communities obtained by other methods and confirmed that while topics of communities by HITS and its derivations tend to be rather general, those of Max-Flow communities tend to be specific. In this paper, we analyze the global properties of Max-Flow community. First we construct *Max-Flow community graph* whose nodes and edges are Max-Flow communities and relevancies between the communities respectively. The next, we study the differences between Max-Flow community graph and an existing community graph based on HITS related methods.

Key words　web community, web graph, web community graph, max-flow algorithm, HITS

1. はじめに

ウェブコミュニティ（以降，単にコミュニティと呼ぶ）とは，話題が共通するウェブページの集合を意味し，ウェブ検索やウェブの成長分析，ウェブからのトレンド発見など幅広いウェブ技術への応用が期待できる概念として近年非常に重要視されている．ウェブページとその間に張られたハイパーリンクをそれぞれノード，エッジと見なせば，ウェブは巨大な有向グラフ（ウェブグラフと呼ばれる）から成る仮想空間を構築している．コミュニティ特有のグラフ構造を解析することにより，コミュニティを効率よく発見するための種々の手法が提案されてきた [1], [2], [8]．また，HITS [3] などの関連ページアルゴリズムを利用してコミュニティを抽出する方法も存在する [4]〜[6].

[8] では，少なくとも一つ以上の完全 2 部グラフを含む 2 部グラフとしてコミュニティを認識し，ウェブグラフ内に存在する全てのコミュニティの数え上げを行った．当然のことながら入・出次数の大きいノードが完全 2 部グラフを形成し易く，結果として（2 部グラフとしての）辺密度が極めて高い部分がコミュニティとなる．このため，次数の小さいノードは，コミュニティのメンバーとして適切・不適切に関わらず，抽出されない．このコミュニティは，代表的な関連ページアルゴリズムである HITS [3] におけるハブ・オーソリティの概念と基本的な考

え方は同じであるため，ハブ・オーソリティ度の上位ノードと完全 2 部グラフに基づくコミュニティのメンバーは概ね一致している．

[1], [2] では，コミュニティとは,「コミュニティの外のページへの（又は，からの）リンクよりもコミュニティ内のページ同士のリンクを多くもつ」という条件を満たすウェブページの集合と定義している （図 1 参照）．完全 2 部グラフに基づくコミュニティと比べて，コミュニティのグラフ構造としての条件が緩和されており，完全 2 部グラフを形成し得ないようなノードであっても，コミュニティのメンバーと成りうるようになっている．Flake らは，このようなコミュニティが最大流アルゴリズム [7] によって発見できるということを示した [1]．以降，本論文では，最大流アルゴリズムにより抽出されたコミュニティを Max-Flow コミュニティと呼ぶ．

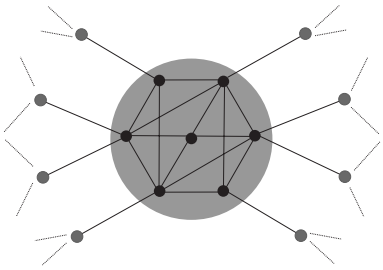


図 1　Max-Flow コミュニティの例

我々は，これまでに，Max-Flow コミュニティと完全 2 部グラフによるコミュニティを比較し Max-Flow コミュニティの特徴を分析した [9]．同じシードページを与えると完全 2 部グラフによるコミュニティではトピックが一般的になり，Max-Flow コミュニティではより詳細になる傾向があることがわかった．また，既存の Max-Flow コミュニティ抽出手法の問題点を明示し，新たな抽出手法を提案した [10]．一方，豊田らによる研究では，日本国内のウェブグラフデータを用いて，HITS 系手法によりウェブ空間全域からコミュニティを抽出し，コミュニティをノード，コミュニティ間の関連をエッジとしたコミュニティグラフを構築し，ウェブ空間に内在する興味深い様々なコミュニティ間の関連を浮かび上がらせることができた [11]．同じシードページを与えると抽出手法によりコミュニティは異なる特徴を持つ．それら個々の特徴の差は，コミュニティを巨視的に捉えた場合，どのような特徴差となって現れてくるかについてはまだわかっていない．そこで，本論文では，Max-Flow コミュニティを利用して Max-Flow コミュニティグラフを構築し，上述の既存のコミュニティグラフと比較する．互いのコミュニティグラフの相違点を明確にし，それぞれのコミュニティ抽出手法の相違から生じる大域的な特徴差を分析する．

本論文の構成は以下の通りである．第 2 節では，Max-Flow コミュニティの抽出手法について簡単に解説する．第 3 節では，Max-Flow コミュニティグラフの構築に関して，構成ノードおよびエッジについて説明し，グラフの表示方法について述べる．第 4 節では，Max-Flow コミュニティグラフの例を示し，第 5 節で既存のコミュニティグラフと比較し両グラフの特徴差を明確にする．最後にまとめと今後の課題を述べる．

2. Max-Flow コミュニティの抽出

以下に Max-Flow コミュニティの抽出手順を示す．手順の詳細，及び，この手順で得られたコミュニティの性能評価については, [10] を参照されたい．

（1） シードノード集合 $S = \{v_{s_1}, v_{s_2}, ..., v_{s_l}\}$ を入力．
（2） 各 $v_{s_i} \in S$ から深さ 2 のサブグラフを抽出する．
（3） 周辺グラフ $G(V, E)$ を後述の手順に沿って構築する．
（4） もともと存在するエッジ $(u, v) \in E$ に対し，後述の式 (1) で得られた辺容量 $c(u, v)$ を与える．$(v, u) \notin E$ のとき，辺容量 $c(v, u) = c(u, v)$ のエッジ (v, u) を E に加える．
（5） 最大流アルゴリズムを実行する．
（6） シードから不飽和辺を辿って到達可能なノード集合を $C = \{v_{c_1}, v_{c_2}, ..., v_{c_m}\}$ とし，各 $v_{c_i} \in C$ に後述する式 (2) でランク付けを行う．
（7） スコア上位数ノードを S に加え上記の手順を C が安定するまで繰り返す．
（8） スコア順に C のノードを出力．

周辺グラフの構築：V，E をそれぞれ各 $v_{s_i} \in S$ から深さ 2 のサブグラフ内のノード，エッジ集合とする．V に仮想ソースノード s と共に，E に s から全てのシードノード $S = \{v_{s_1}, v_{s_2}, ..., v_{s_l}\}$ へ辺容量＝ ∞ のエッジを加える．次に，V に仮想シンクノード t と共に，$V - \{S \cup s \cup t\}$ の各ノードから t へ辺容量＝ 1 のエッジを加える．図 2 は，周辺グラフ $G(V, E)$ を模式的に示したものである．エッジ横の数字は，辺容量を意味している．また，下 2 段のノードから仮想シンクノードへのエッジは省略されている．

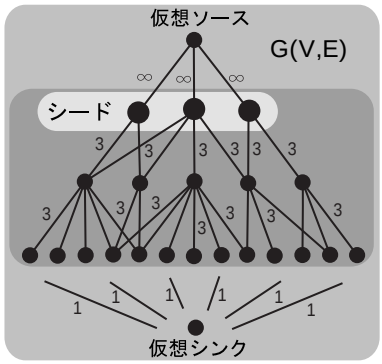


図 2　周辺グラフの例

辺容量の設定：辺容量は，各エッジごとに設定し，重要なエッジ (つまり，両端点のノードがシードノードと関連が深いエッジ) ほど飽和しにくくし，両端点のノードを抽出しやすくする (逆もまた同様に考える）ことが必要である．エッジの重要度は，関連ページアルゴリズムにより求めることができる．PageRank，HITS が代表的な関連ページアルゴリズムであるとされているが，直径の小さいサブグラフにおいてもある程度一定の高い質が期待できるなどの理由から HITS を採用する．HITS アルゴリ

ズムにより各ノードのハブ・オーソリティ値を求め，各エッジの始点のハブ値，終点のオーソリティ値に基づいて辺容量を設定する．ただし，ハブ・オーソリティ値を辺容量に利用するには，[0,1] の実数を自然数に変換する必要がある．変換後の値をそれぞれ h'_{v_i}，a'_{v_i} と書く（これらの値を求めるための詳細手順は[10] を参照.）とき，ノード v から u へのエッジ $(v, u \in V, (v, u) \in E)$ に対する辺容量 $c(v, u)$ は，次式で求める．

$$c(v,u) = \lfloor \frac{h'_v + a'_u}{2} \rfloor \quad (1)$$

スコア付け：各ノードに対するスコアは，そのノードのオーソリティ値，ハブ値によって重み付けされた入・出リンクの総和によって求める．$v_{c_i}(In)$，$v_{c_i}(Out)$ をそれぞれ $v_{c_j} \in C$ から v_{c_i} へのリンク数，$v_{c_k} \in C$ から o v_{c_i} へのリンク数とする．v_{c_i} のスコア $Sc(v_{c_i})$ は次式で求める．

$$Sc(v_{c_i}) = a_{v_{c_i}} v_{c_i}(In) + h_{v_{c_i}} v_{c_i}(Out). \quad (2)$$

3. Max-Flow コミュニティグラフの構築

Max-Flow コミュニティグラフとは，Max-Flow コミュニティをノードとし，関連するコミュニティ間に重み付きのエッジを張ったグラフである．エッジの重みは，類似度に基づくコミュニティ間の関連度を示すものとする．Max-Flow コミュニティグラフの構築は，ウェブ全域の Max-Flow コミュニティの抽出（ノード抽出），及び，抽出した Max-Flow コミュニティ間のエッジ生成の 2 ステップで実現する．また，コミュニティに重複がある場合は，それらをマージしてグラフ表示する．

3.1 Max-Flow コミュニティグラフ構成ノード

十分に多くのシードノードを用いてウェブ全域から抽出した Max-Flow コミュニティをグラフのノードとする．ここで言うウェブ全域とは，日本国内のウェブを意味する．以下に Max-Flow コミュニティの全域抽出の詳細を示す.

利用データ：2002 年 2 月にクロールした日本国内のウェブアーカイブによるウェブグラフデータベースを利用した．これは，アーカイブの全てのページから URL とリンク情報（どの URL がどの URL へのリンクを持つか）のみを取り出し，ウェブをグラフ表現によりデータベース化したものである．実際にクロールしたウェブページ数は，約 4500 万ページ，これらのページが保持するリンク情報を元に得られた URL の総数は約 8400 万，また総リンク数は，約 3 億 7500 万である．

シードノード：シードノードの入リンク数が少なすぎる場合，周辺グラフは極めて小さくなり，意味のある Max-Flow コミュニティの抽出が期待できない．そこで，上記のウェブグラフデータベースにおいて，3 つ以上の異なるサーバからリンクを張られているページをシードノードとして選択した．このとき総シードノード数は，約 132 万である．

前節で示した Max-Flow コミュニティ抽出手法を用いて，上記の各シードノードに対して，Max-Flow コミュニティを求める．ウェブ内から十分に多くのシードノードを利用し Max-Flow コミュニティを抽出するため，それらコミュニティのメンバーには重なりが生じていることが予想される．各コミュニティの上位 20 ノードについての重なり度を検証したところ，約 8 割のノードが単一のコミュニティに属しており，95%強のノードは，3 つ以下のコミュニティに属していることがわかった．また，10 以上のコミュニティに属するノードは，約 2%（約 55000 ノード）存在した．図 3 のグラフは，属するコミュニティ数（5〜50：横軸）に属しているノードの数（縦軸）をプロットしたものである．約 32300 のノードが異なる 5 つのコミュニティに属している．

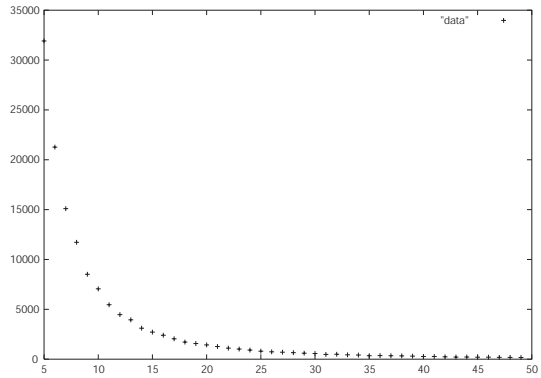
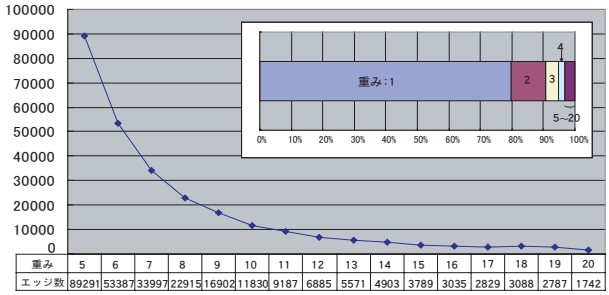


図 3　属するコミュニティ数（横軸）とノード数（縦軸）分布

3.2 Max-Flow コミュニティグラフ構成エッジ

Max-Flow コミュニティグラフでは，コミュニティ間の関連度をコミュニティの類似度で求め，それらコミュニティ間に重み付きのエッジを張る．コミュニティの類似度は，それぞれのコミュニティにおける上位 n ノード（ただし，後述の例は全て $n = 20$）の重複メンバー数で求める．エッジは無向であり，その重みの最大値は，n となる．図 4 は，属しているコミュニティ数が 50 以下のノードのみで得られたエッジに関して，その重み（横軸）と対応する重みのエッジ数（縦軸）を示したものである（ただし，$n = 20$）．グラフ右上の棒グラフは，重みが 1 から 4 およびそれ以上のエッジの割合を示している．重み 1 から 4 のエッジ数は，それぞれ，約 669 万，93 万，34 万，16 万であった．また折れ線グラフは，重み 5 から 20 のエッジ数を示している．重みが大きくなるにつれ，コミュニティの類似度は増す．よって，コミュニティ間の重みが 20 の場合は，両コミュニティの上位 20 ノードは完全に一致していることから，コミュニティのメンバーがほぼ全て重複していると予想される．



重み	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
エッジ数	89291	53387	33997	22915	16902	11830	9187	6885	5571	4903	3789	3035	2829	3088	2787	1742

図 4　エッジの重み（横軸）とエッジ数（縦軸）

3.3　Max-Flow コミュニティグラフ表示方法

現在のところは，着目したいコミュニティ近辺のグラフのみを取り出し，Max-Flow コミュニティグラフの部分グラフとしてそれを表示する．具体的には，以下の手順で行う．

（1） URL を一つ指定し，その URL が含まれているコミュニティセット $C = \{C_1, C_2, ..., C_n\}$ を取り出す．

（2） さらに広範囲のグラフを検証する場合は，取り出した $C_i \in C$ の各コミュニティを抽出したシードページ s_i が含まれているコミュニティセット $C_{s_i} = \{C_{s_i1}, C_{s_i2}, ..., C_{s_in'}\}$, を取り出す．

（3） $\mathbf{C} = C \cup C_{s_1} \cup C_{s_2} \cup, ..., \cup C_{s_n}$ をベースノードセットとし（図 5 参照），この中でメンバーの類似度が設定閾値以上のコミュニティ同士をマージする．

例えば，閾値=0.6 とした場合，重複メンバー数が双方のコミュニティメンバー数の平均の 6 割以上の場合はマージを行う．

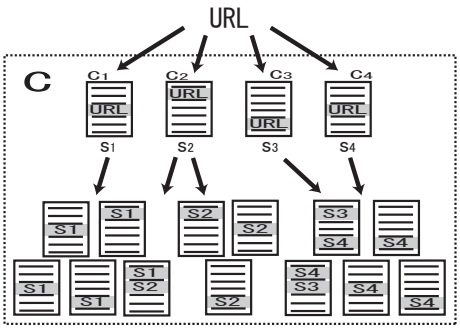


図 5　Max-Flow コミュニティグラフのノード取得方法

4.　Max-Flow コミュニティグラフの例

本節では，前節で述べた Max-Flow コミュニティグラフ構築方法によって得られたグラフの具体例を，マージスを行う場合および行わない場合，また，コミュニティセットを指定 URL から 2 段階まで辿って抽出した後，マージを行った場合の 3 種について示す．

4.1　マージを行わない場合

図 6 は，URL:kurashiki.or.jp/が含まれている 4 つのコミュニティを取り出した部分グラフを示している．図中のエッジ横の数字は該当エッジの重みを表している．右の表は，該当コミュニティ内のページを指すリンクのアンカーテキストを基に抽出されたキーワードを示している．この例の指定 URL は，倉敷の観光情報等に関するポータルサイトである．キーワードからは分かり難いが，Comm1 は，岡山県所在企業（サンワサプライ），岡山県児島等，Comm2 は，旅行情報等，Comm3 は，倉敷所在のチボリ公園等のページ，Comm4 は，高速・交通情報，大原美術館等のページから成っており，どのコミュニティも何らかのかたちで倉敷に関連している．

4.2　マージを行う場合

図 7 は，URL:www.miso.or.jp/を指定した．この URL が含まれているコミュニティは 30 存在し，重複コミュニティが多数存在するのでコミュニティのマージステップを行った．その際の閾値は，0.6 とした．マージステップにより，抽出グラフ内のノード数は 16 となっている．エッジの重みが 2 未満の部分は省略してある．この例の指定 URL は，味噌に関するページである．この例の場合は，コミュニティキーワードおよびメンバー URL からわかるように，指定 URL の「味噌」というトピックと現実的にも関わりが深いと思われるような醤油，酒，米等をトピックとしたコミュニティ群が，実際にコミュニティグラフ上でも互いに関連していることが確認できる．

4.3　2 段階までグラフを広げた場合

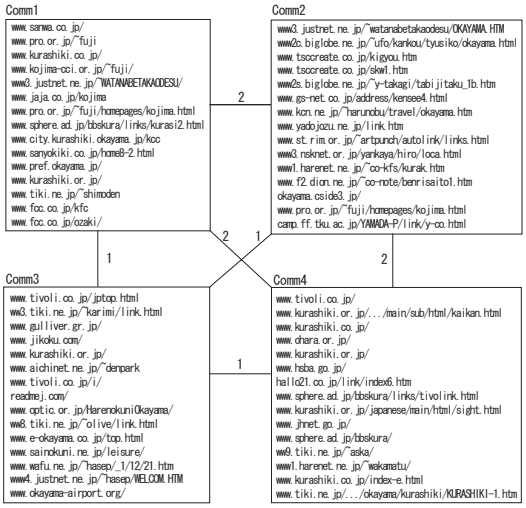
図 8 における指定 URL は，バレーボール協会サイトの URL:www.jva.or.jp/であった．この URL が含まれているコミュニティは 28 存在し，コミュニティの 2 段階抽出をした後のコミュニティ数は 122 となっており，そのうち 74 のコミュニティが閾値 0.6 でマージされ 16 のノードとなっている．図 8 は，マージされた 16 ノードについてのグラフを表している．表中の数字が一つのコミュニティを意味しており，右表が該当コミュニティのメンバーにおけるキーワードとなっている．また，色づけられたノードは，指定 URL が含まれているコミュニティである．図中点線で示されたエッジのみが重み 1 であり，その他は，重み 2 以上のエッジのみを表示している．指定 URL が含まれている 6 つのコミュニティは，大小様々な規模（社会人チーム，地域のチーム，大学のチーム等）のバレーボールチーム，バレーボールに関する総合情報サイト等からなるコミュニティである．Comm8 にバレーボールとスイミングが共存するスポーツ全般に関するコミュニティを介して，Comm14，Comm16 と水泳関係のコミュニティが抽出されている．また，薄い繋がりではあるが，ラグビーに関するコミュニティ群が，バレーボールコミュニティの近辺に存在することがわかる．

5.　コミュニティグラフの特徴比較

本節では，既存のコミュニティグラフと Max-Flow コミュニティグラフを比較し，Max-Flow コミュニティの有する特徴を分析する．最初に，比較対象となる既存のコミュニティグラフについて説明した後，両グラフの相違が顕著な例を挙げ，Max-Flow コミュニティグラフの特徴について述べる．

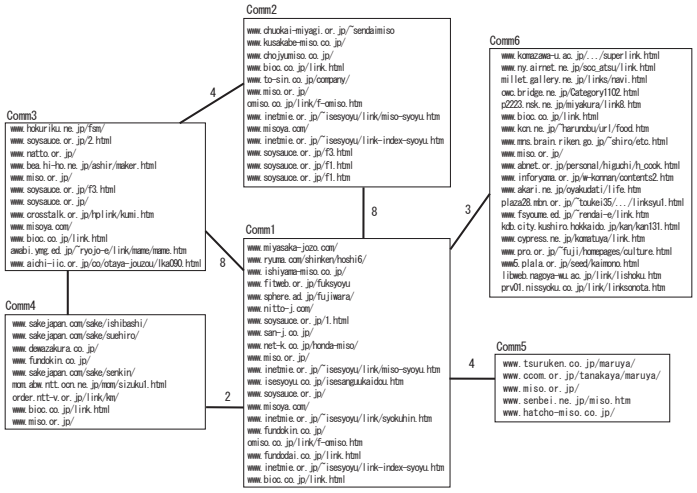
5.1　既存のコミュニティグラフ概要

豊田らによるコミュニティグラフ（[11] においては，ウェブ



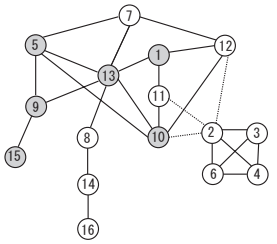
Comm1	サンワサプライ 児島 大学 短期 県庁 sanwa 競艇 電 岡 山 supply
Comm2	全国 国内 旅 sights 旅行 児島 1cv ライン 藤原 博文 やど
Comm3	ガリバー チボリ 王国 公園 富士 box 公式 くらしき ホテル 組合 旅館
Comm4	公団 四国 道路 連絡 橋 本州 jh 高速 美術館 大原 交通

図 6　Max-Flow コミュニティグラフの例 1：マージを行わない場合



Comm1	醤油 協会 醸造 米 五
Comm2	米 五 醤油 宮城 朴
Comm3	醤油 米 五 協会 福井 納豆
Comm4	酒造 桜 出羽 フンドーキン 醤油
Comm5	八丁味噌 カクキュー まる 資 郷
Comm6	飲食 part-1 薦め 藤原 博文 料理

図 7　Max-Flow コミュニティグラフの例 2：マージを行う場合



Comm1	小学生 バレー クラブ 高岡 大塚 調布 新宿 少女 柏木 長崎
Comm2	ラグビー ラグビーフットボール 協会 rugby 関東 union 大学 football 関西 神戸製鋼
Comm3	cool ラグビーフットボール ラグビー 協会 関東 オンライン union rugby football 大学
Comm4	ラグビーフットボール ラグビー 協会 関東 union 九州 football 関西 高校 神戸製鋼
Comm5	バレーボール 大学 連盟 連 東海 協会 volleyball volley 学 表
Comm6	ホーム ラグビーフットボール ラグビー ライコス ライコスジャパン 協会 関東 union rugby
Comm7	ドメイン ebookoff 転職 auction お店 温泉 hoops ヘルプ span 雁 医
Comm8	worldwide 堀内 jafsa wide volleyball www 水泳 フィンスイミング finswimming vwwhome
Comm9	バレーボール 協会 volley vi volleyball 星座 早見 排球 盤 association
Comm10	バレーボール volleyball バレーボールガイド 協会 東洋 旭 化成 オーキス ブレイザーズ 横浜
Comm11	vi 栄 volley みん クラブ ball 今成 バレーボール volleyball ken
Comm12	volley ball volleyball valleyball バレーボールオフィシャルサイトリンク バレーボール
Comm13	バレーボール 協会 ソフトバレーボール association 財団 volleyball jva 法人 連盟 財 大阪
Comm14	水泳 スイミングクラブ fina 国際 協会 連盟 社団 swimming internationale natation
Comm15	東北 大学 アシックス ヨネックス 日本野球連盟 asics tohoku yonex university バレーボール
Comm16	水泳 interwoman swimming ゴルフ インターウーマン 連盟 inter 王国 水 woman

図 8　Max-Flow コミュニティグラフの例 3：2 段階までグラフを広げた場合

コミュニティチャートと呼んでいる）では，HITS を発展させた関連ページアルゴリズム [11], [12] が利用されている．このコミュニティグラフにおいても，ノードは，コミュニティであり，重み付きエッジは，コミュニティ間の関連度を意味している．ただし，Max-Flow コミュニティグラフとは異なり，エッジは有向である．コミュニティグラフの構築手順は概ね以下のようになる．詳細に関しては [11] を参照されたい．尚，比較の際に利用したコミュニティグラフは，前節で述べた Max-Flow コミュニティグラフ構築の際に利用したデータと同一のものにより作成されている．

（1） 全てのシード（異なる 3 つ以上のサーバからリンクされているページ）について関連ページアルゴリズムを適用し，各シードが他のシードをどのように導出するか調べる．

（2） シードが互いに導出しあうという対称関係で密に結合されたシード同士をコミュニティとして抽出する．

（3） 2 つのコミュニティのメンバー間に導出関係がある場合に，そのコミュニティ間にエッジを張る．

比較には，ウェブコミュニティブラウザツール [13], [14] により表示したグラフを用いる．ノードは，コミュニティを示しており，アンカーテキストを基に生成されたラベルが張られている．このツールでは，URL を入力するとその URL が含まれているコミュニティが画面上に現れる．厳密には，指定 URL と一部分が一致する URL が含まれているコミュニティも現れるが，実際に指定 URL が含まれているコミュニティのみを残す．まず，そのコミュニティの周辺コミュニティを追加表示する機能（in-link・out-link 展開）を用いて周辺コミュニティを全て表示する．表示されるノード（コミュニティ）が多すぎる場合は，ノードの大きさ，エッジの重みによる閾値操作により適度にノードおよびエッジを除去する．

5.2 コミュニティグラフの比較例

比較例 1：前節における 3 例目の指定 URL：www.jva.or.jp（バレーボール協会サイト）は，既存コミュニティグラフにおいては，各種スポーツ協会のコミュニティに属している．図 9 は，このコミュニティを in-link 展開し，エッジの重みの閾値を 5，コミュニティサイズを 10 以上に設定して表示させたグラフと，指定 URL が含まれているコミュニティメンバーの URL（右表）を示している．

一方，前節（図 8）で示したように，Max-Flow コミュニティグラフでは，各種スポーツ協会のページが集まったコミュニティは存在せず，指定 URL は複数のバレーボールに関するコミュニティに含まれていた．既存コミュニティグラフでは，指定 URL が含まれているコミュニティを介して他の種々のスポーツに関するコミュニティが連結されているが，Max-Flow コミュニティグラフでは，色々なスポーツ全般のリンク集や，それらのページからリンクされているページからなるコミュニティ（例えば，Comm8）を介して異種スポーツのコミュニティ間が連結していた．

比較例 2：URL に well-mannered.org/（個人が設立している犬を飼う際の情報サイト）を指定する．図 10 は，マージ操作をせずにエッジの重さが 2 以上部分を抽出した Max-Flow コミュニティグラフである．ただし，各コミュニティはそれぞれ 5 つずつ URL を表示し，右表は，これまでと同様，該当コミュニティのアンカーテキストを基にして抽出したキーワードを示している．犬の飼い主の私的サイト集（Comm2，Comm5，Comm7，Comm8），犬の総合情報サイト・リンク集（Comm3，Comm4，Comm6）その他犬関連のページ集（Comm1，Comm6）等がコミュニティ群として抽出された．

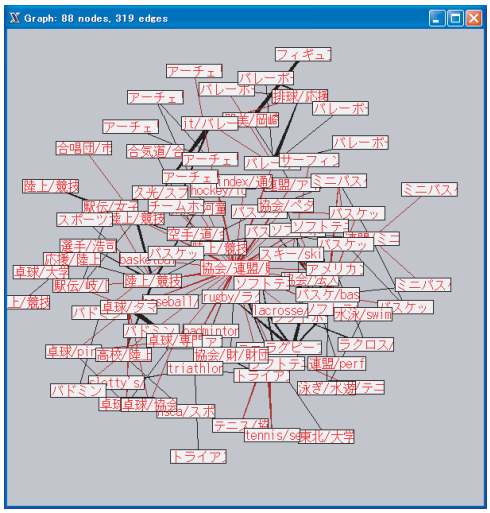
一方，既存コミュニティグラフでは，指定 URL が種々の犬情報サイトから成るコミュニティに属している．図 11 は，このコミュニティを in-link 展開し，エッジの重みを 10，コミュニティサイズを 10 以上に設定して表示させたグラフと，指定 URL が含まれているコミュニティメンバーの URL（右表）を示している．Max-Flow コミュニティグラフでは，複数のコミュニティにまたがって存在していたページが一つのコミュニティに集約されており，色々なペット，動物，犬の雑貨等関するコミュニティが周辺に存在している．

5.3 考　　察

比較例 1 より，トピック A に関して極めて重要度が高いページ B（例えば，トピック「バレーボール」に対してバレーボール協会，トピック「倉敷観光」に対して倉敷市サイト内の観光ページ等）が含まれるコミュニティは，Max-Flow コミュニティグラフでは，トピック A と階層的に関連しているページ集合（例えば，「バレーボール」に対して，地域のバレーボールチームのページなど，「倉敷観光」に対して，私的な倉敷紹介サイト，倉敷市所在の公園に関するサイトなど）となっており，それらがグラフの一部分を形成していた．一方，既存コミュニティグラフでは，B と並列的に関連しているページ（例えば，バレーボール協会のページに対して，その他各種スポーツの協会・団体ページ，倉敷市サイト内の観光ページに対して，市町村が運営する観光ページ）が，一つのコミュニティを形成していた．このことから，Max-Flow コミュニティグラフからは，一つのトピックに関して，階層的なコミュニティの分布を検証するのに適していると言える．

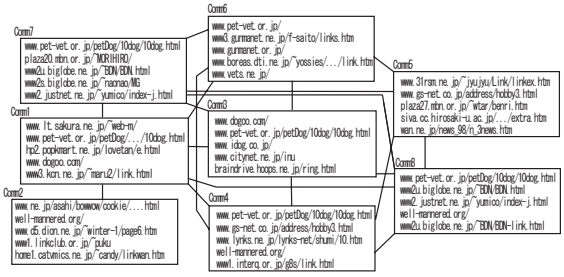
比較例 2 のように，既存コミュニティグラフでは単一のコミュニティに属しているページ群が，Max-Flow コミュニティグラフでは，複数のコミュニティに分散しグラフの一部を形成している場合が多く見られた．また，前述のように比較例 1 の既存コミュニティグラフにおいて指定 URL が含まれていたコミュニティは，各種スポーツ協会・団体のページであったが，Max-Flow コミュニティグラフでは，そのようなコミュニティは存在しない．しかしがなら，各種スポーツリンク集と共に種々のスポーツのページが集まったコミュニティは存在する．これらのことから，既存コミュニティグラフに比べて Max-Flow コミュニティグラフは，細かい粒度のページ集合となっていると言える．両コミュニティグラフの関係は模式的に図 12 のように表すことができる．

また，Max-Flow コミュニティグラフでは，全く関連のないと



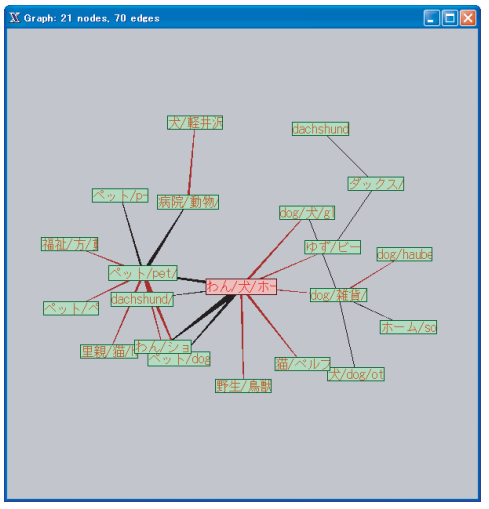
jtta.ge.niigata-u.ac.jp/
www.u-netsurf.ne.jp/nichiba
www.soft-tennis.com/
www.handball.or.jp/
www.jva.or.jp/
www.rikuren.or.jp/
www.kendo.or.jp/
www.jihf.or.jp/
www.jabba-net.com/
www.joc.or.jp/
www.archery.or.jp

図 9　既存コミュニティグラフ：www.jva.or.jp



Comm1	広報 板 獣 医師 好き 太 一行 歓迎 集合 ゴン pet
Comm2	らん うえ chinami bob's ruddy's ウエスティランド いちご キッチン c-nuts チロル house
Comm3	free queen's copyright queen&#39 わん ホーム リング 説明 ring jazz 登録
Comm4	わん ホーム forum lovers with pet トップ living-with-dogs living lover's 十戒
Comm5	ビアディー デイジー bdn 目次 じゅん フロントページ daisy's bnd beardie beardies front
Comm6	獣 群馬 医師 わん 広報 板 pet forum ホーム lovers with
Comm7	cool わん ひろ ビアディー オンライン top アート デイジー daisy's
Comm8	デイジー わん ビアディー ホーム daisy's bdn front 募集 里親 back ペット

図 10　Max-Flow コミュニティグラフ：well-mannered.org/



ww.wan.ne.jp/
www.vets.ne.jp/
www.pet-vet.or.jp/
www.idog.co.jp/
www.dogoo.com/
w3.poporo.ne.jp/~bxro7302/cgi-bin/dog/navi.html
www.wnlll.com/
www.petoffice.co.jp/daisuki
www.papipopo.com/pet/
www.ne.jp/asahi/dog/crew
www.jungle-pet.com/
www.ilovepet.net/
www.citynet.ne.jp/inu
www.pets-city.co.jp/
www.petpet.ne.jp/
www.j-pet.com/
www2s.biglobe.ne.jp/~sheltie/navi/
www.pets-j.com/
www.peppynet.com/
www5a.biglobe.ne.jp/~pet-mart/top.htm
www.j-bic.com/
www.dog-web.com/
www.colorzoo.com/
living-with-dogs.com/
cgi2.osk.3web.ne.jp/~dtoku/cgi-bin/yellowpg.cgi
www.pwl.co.jp/
www.jspca.or.jp/
www.pet-vet.or.jp/petDog/10dog/10dog.html
www.mlf-aid.com/
www.kanapure.net/cgi/search-ani
well-mannered.org/
www2s.biglobe.ne.jp/~exiady/wanwan
www1.neweb.ne.jp/wa/mikamoto
www.sapporo104.net/lornyan
www.pet-zoo.com/chameleon/
www.kanapure.net/yokomama
www.interdream.co.jp/search/search02/pet.cgi
petbottle.bridge-net.co.jp/
www.wander.ne.jp/
www.p-well.com/
www.ai-pets.net/

図 11　既存コミュニティグラフ：well-mannered.org/

思われるトピックのコミュニティ同士が近隣コミュニティとして，抽出されることがある．これらは，それらのトピックを結びつけるリンクが偶然的に何らかの理由であるページに張られることにより生じる．実際に確認できた例としては，あるスポーツ選手のオフィシャルホームページとそのページの制作会社を結ぶリンクにより，そのスポーツに関するコミュニティとネット系デザイン会社のコミュニティが隣接していた．コーヒーのコミュニティの近辺に JAVA 言語に関するコミュニティが存在しているといった興味深い例も確認できた．よって，Max-Flow コミュニティグラフでは，表だってはその存在がわからないようなトピック間の隠れた関係を導き出すことができる．

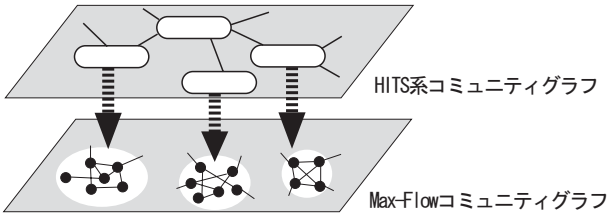


図 12　コミュニティグラフの相違

6. おわりに

本論文では，最大流アルゴリズムを用いて抽出したウェブコミュニティを Max-Flow コミュニティと呼び，Max-Flow コミュニティをノードとし，コミュニティ間の類似度に基く重み付きエッジを張った，Max-Flow コミュニティグラフを構築する方法を述べた．また，豊田らによる HITS 系手法を利用して得たコミュニティをノードにした既存のウェブコミュニティグラフと比較し，Max-Flow コミュニティグラフが有する特徴を示した．これまでの研究では，同じシードを与えると，Max-Flow コミュニティは，シードトピックよりもより詳細なトピックになり，逆に，HITS 系手法によるコミュニティでは，より一般的になるといった，一つ一つのコミュニティレベルでの特徴の相違が明らかになっていた．本研究により，それらのコミュニティを巨視的に見た場合，Max-Flow コミュニティグラフは，HITS 系手法によるコミュニティグラフに比べて，細かい粒度のグラフとなっていることがわかった．

現在のところ，コミュニティ間のエッジには，コミュニティの重複に基づく類似度を利用しており，重複のあるコミュニティに関するグラフのみが得られるようになっている．しかしながら，この方針では，ウェブグラフ上では隣接関係にあるようなコミュニティであっても，重複がなければ表示することはできない．よって，基のウェブグラフにおけるリンク関係を反映させるなどして，ウェブグラフにおいて周辺にあるコミュニティをも抽出可能にする方法を考えなければならない．それと同時に，より簡単に，Max-Flow コミュニティグラフを表示し特徴を検証するために，URL の指定，キーワード及びグラフの表示などの一連の操作を可能にするツールを開発することが今後の大きな課題である．

文　献

[1] G. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, 2000.

[2] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[3] J.M.Kleinberg: *Authoritative Sources in a Hyperlinked Environment*, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677, 1998.

[4] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[5] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*, 1998.

[6] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In *The VLDB Journal*, pages 639–650, 1999.

[7] L. Ford.Jr. and D.R.Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

[8] R.Kumar, P.Raghavan, S.Rajagopalan, and A.Tomkins: *Trawling the web for emerging cyber-communities*, Proc. 8th International WWW Conference, pp. 1481–1493, 1999.

[9] N. Imafuji and M. Kitsuregawa. Effects of maximum flow algorithm on identifying web community. *4th International Workshop on Web Information and Data Management*, pages 43–48, 2002.

[10] N. Imafuji and M. Kitsuregawa. Finding Web Communities by Maximum Flow Algorithm using Well-Assigned Edge Capacities. *To be appeared in Information Processing Technology for Web Utilization, IEICE*, Feb. 2004.

[11] M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. *12th ACM Hypertext*, pages 103–112, 2001.

[12] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Proceedings of the 8th International World Wide Web Conference*, pages 1467–1479, 1999.

[13] 福地健太郎, 豊田正史, 喜連川優. Web Community Browser: 大規模 Web コミュニティチャートの可視化. 第 13 回データ工学ワークショップ (DEWS2002) 論文集 A1-4, 2002.

[14] 福地健太郎, 豊田正史, 喜連川優. Web Community Browser: Web コミュニティ構造の可視化と探索機構の実現. 第 1 回情報科学技術フォーラム (FIT2002) 論文集, 2002.