グラフ構造によるウェブコミュニティの特徴分析：Max-Flow vs HITS

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
Email; {imafuji, kitsure}@tkl.iis.u-tokyo.ac.jp

要 旨

ウェブコミュニティ（以降，単にコミュニティと呼ぶ）とは似たようなトピックを扱うウェブページの集合を意味する．これまでにウェブから効率よくコミュニティを抽出する手法が多数提案されてきた．それらの手法は，ウェブにおけるハイパーリンク構造の特徴をそれぞれ異なる視点から捉え，それを反映させたグラフ構造でコミュニティを表現する．本稿では，2 手法－ HITS アルゴリズムに基づく手法，最大流アルゴリズムを利用する方法－に着目する．両手法が指針とするグラフ構造を比較分析し，トピック規模という概念を導入し，両手法により得られるコミュニティの特徴差を示す．また，2002 年の日本国内のウェブページから抽出したコミュニティから成る大規模コミュニティセットを利用した実験を行い，この分析の正当性を検証する．

Analyzing Feature of Web Communities: Max-Flow vs HITS

Noriko IMAFUJI†　　Masaru KITSUREGAWA†

†Institute of Industrial Science, University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo, 153-8505 Japan
Email; {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. Various graph theoretical approaches for extracting web communities from the web have been proposed. Each approach specifies a web community by each distinct graph structure, which is well-founded in the context of the link behavior in the web. In this paper, we focus on the web communities extracted by the two approaches; Max-Flow algorithm based and HITS related algorithm based and identify the feature distinction between them. We first analyze a potential feature on the analogy of the graph structure. Next, we verify it by examining two sets of web communities; approximately 1.32 million Max-Flow communities and 1.2 million HITS based communities. Those communities are extracted from the web graph based on Japanese archives in February 2002.

1　はじめに

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．サーチエンジンに代表される情報検索技術の目まぐるしい発展は，混沌としたウェブ空間から様々な情報を引き出すことを可能にした．例えば，実社会においてある話題が注目されると，その話題に関する多くのウェブページがウェブ上に現れ，質の良いページはブックマークやリンク集などからリンクされることにより，非常に密なリンク構造を構築していく．このように，ウェブは如実に実社会の動きを映し出しており，最近ではウェブを巨視的・時系列で捉えて得られる情報の重要性と，それが生み出す様々な可能性に注目した研究が盛んである．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する．コミュニティの存在が，ウェブ上に存在する一つの話題を意味することから，コミュニティは，ウェブにおける現象を捉える一つの指針として利用される．

これまでに，種々のコミュニティ抽出法が提案されてきた．[3] 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[4] 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 [11] によって発見できるということを示した [1]．以降，本稿では，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 では，円で囲まれたノード集合が Max-Flow コミュニティとなっている．

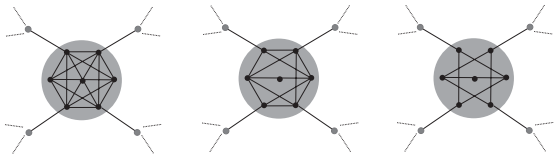

図 1: Max-Flow コミュニティの例

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の例をいくつか比較し Max-Flow コミュニティが持つ特徴を示した [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．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 [8]．また，Max-Flow コミュニティをノードにしたコミュニティグラフを例示し，前述の HITS 系コミュニティグラフと比較することにより両コミュニティグラフのいくつかの相違点を示した [12]．以上より，コミュニティには抽出手法によって局所的にも大域的にも異なる特徴を示すことがわかった．本稿の目的は，これらの議論をグラフ構造に着目した分析により裏付けることにある．具体的には，Max-Flow 手法，HITS 系手法の両手法で得られるコミュニティ内のグラフ構造，および，コミュニティ間のグラフ構造を分析し，それぞれのコミュニティの特徴を示す．また，主として 2002 年の日本国内のウェブページから成る大規模コミュニティセットを利用した実験を行い，この分析の正当性を検証する．

本稿の構成は以下の通りである．第 2 節では，コミュニティの抽出手法－ Max-Flow による手法，HITS 系手法－について簡単に解説する．第 3 節では，両

手法がコミュニティ抽出に用いるグラフ構造に着目し，潜在的な特徴差を分析する．第 4 節では，大規模ウェブデータを用いて行った実験について述べ，第 3 節で行った分析に対する検証結果を示す．最後にまとめと今後の課題を述べる．

2 コミュニティの抽出手法

本節では，コミュニティの抽出手法について簡単に説明する．ただし，ここで説明する二つの手法は，検証に用いたコミュニティセット構築のために利用された手法であることに注意されたい．

2.1 Max-Flow 手法

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. シードページ v_s を入力．
2. v_s を中心として深さ 2 のサブグラフ $G'(V,E)$ を抽出する．V は，v_s から前後方向に 2 リンク以内で辿れるノードの集合，E は，V 内のノードを両端点とする有向辺の集合とする．
3. 周辺グラフ $G(V,E)$ を以下の手順に沿って構築する．仮想ソースノード s を V に，また辺 (s,v_s)（辺容量=∞）を E に加える．また，仮想シンクノード t を V に，また $V-\{S\cup s\cup t\}$ の各ノードから t への辺（辺容量＝1）を E に加える．各辺 $(u,v)\in E$ に対して後述する方法で辺容量 $c(u,v)$ を設定する．ただし，$(v,u)\notin E$ のとき，辺 (v,u) を辺容量 $c(v,u)=c(u,v)$ として E に加える．
4. 最大流アルゴリズム [11] を実行する．
5. 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 $v_{c_i}\in C$ にスコア付けを行う．
6. 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．
7. スコア順に C のノードを出力する．

辺容量は，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．

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

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_i} から $v_{c_k}\in C$ 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．

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

2.2 HITS 手法

HITS 系手法の多くは，与えたシードページに対して，ハブ・オーソリティスコアを求め，それらのスコアが上位のページ集合を，コミュニティとしている．ここで説明する手法は，ウェブグラフ全体からメンバーが重複することなくコミュニティを効率よく抽出することを主眼として，豊田らにより提案された [8]．以下に手順を示す．詳細は，[8] を参照されたい．

1. シードページの集合を入力する．
2. 各シードページに対し，HITS 系関連ページアルゴリズムを施す．
3. オーソリティ導出グラフ (ADG) を構築する．
4. ADG から対照導出グラフ (SDG) を構築する．
5. SDG 内で連結なノード集合をコミュニティのメンバーとして出力する．

オーソリティ導出グラフ (ADG) とは，各シードが，どのように *Companion-* [8] と呼ばれる HITS 系関連ページアルゴリズムによって他のシードを導出するかを示す有向グラフである．ノードは，シードセット内のシードページから成る．また，ノード u から v への有向辺は，u が Companion-で求められた関連ページのうちの一つとして v を導出していることを表している．ADG において u，v 間に双方向にエッジが存在するとき，u，v 間に無向辺を生成し，これらの無向辺とその両端点のノード集合により得られた無向グラフが，対照導出グラフ (SDG)

である．SDG において密に連結しているサブグラフをコミュニティの核として抽出し，これらに残りのノードを加えることによりコミュニティを得る．

3　グラフ構造による特徴分析

コミュニティの抽出原理を端的に表現すると Max-Flow 手法は,「連結度の (コミュニティの) 内外差」, HITS 系手法は,「ハブとオーソリティから成る密連結 2 部グラフ」となり，抽出するコミュニティのグラフ構造は，両者で異なる．本節では，Max-Flow 手法，HITS 系手法の両手法で得られるコミュニティのグラフ構造，および，コミュニティ間のグラフ構造を分析することにより，必然的に生じるコミュニティの特徴差を示す．

あるシードページの入力に対して得られる周辺グラフとそこから得られる Max-Flow コミュニティおよび HITS コミュニティについて考える．周辺グラフは，シードページから前後方向に 2 リンク以内で辿れるページ及びリンクからなるグラフを意味する．図 2 に一つのシードページの周辺グラフから抽出される Max-Flow コミュニティと HITS コミュニティの例を示す．同一の周辺グラフ内で，異なるコミュニティが抽出される要因を図中に描かれたページが実際にどのようなページに対応するか考えながら説明する．

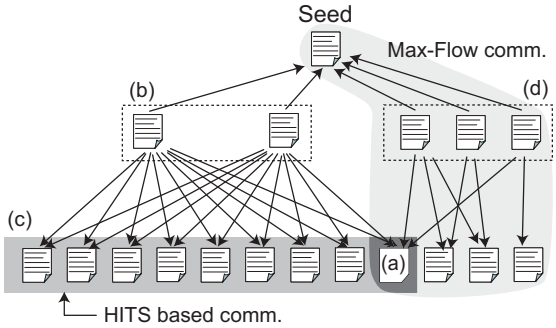

図 2: 周辺グラフから抽出される Max-Flow，HITS コミュニティ

例えば，シードページが，SONY の代表 PC である”VAIO”に関するページであったとする．このシードページをリンクするページには，(b) に相当する主としてコンピュータメーカーのリンク集や，(d) に相当する VAIO ユーザーなどが作成した小規模リンク集などが存在する．前者は，シード以外へのページ (c) へのリンクを多く持つため，シードと (b) 間のエッジは容易に飽和辺となり，(b) と (b) からリンクされる (c) は，Max-Flow コミュニティのメンバーとしては抽出されない．一方，(d) は規模の小さいリンク集に相当するので，シードと (d) 間のエッジは飽和辺と成りにくく，(d) と (d) からリンクされるページ群は，Max-Flow コミュニティのメンバーとして抽出される．ここで，(c) に着目する．これらのページは，IBM や，TOSHIBA といった各種コンピュータメーカのページが相当し，SONY のページ (a) も含まれる．この周辺グラフ外には，(c) のページを共参照するページ，つまり，コンピュータメーカーへのリンク集を持つページは多く存在する．その結果，それらのページがハブ，(c) 内のページがオーソリティとなり，非常に密に連結した 2 部グラフを構築していることになる．よって，(c) のページ群が，HITS 手法ではコミュニティとして抽出される．ただし，(a) は SONY のトップページであり，右側のサブグラフ内からのリンクも多く持つため，Max-Flow コミュニティのメンバーとしても抽出される．

次に，異なる視点からコミュニティ間の関係について考える．トピックが階層構造をしていると仮定すると，メジャー (上位) ↔ マイナー (下位) や，general(上位) ↔ specific(下位) など，トピック規模を形容する軸が考えられる．HITS コミュニティを形成する上位のハブ・オーソリティは，共参照の関係で互いに密に連結しているため，HITS 系コミュニティは，同レベルのトピック規模のページから成ると考えられる．図 3 は，トピック規模の軸と対応する HITS コミュニティの例を示している．

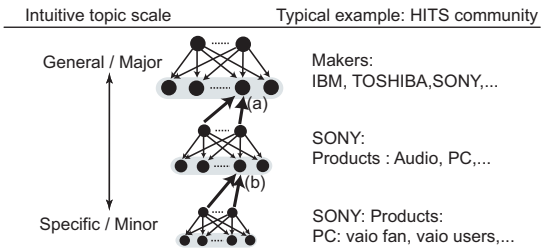

図 3: コミュニティとトピック規模

図中のページ (a) が，SONY のサイトのトップページとすれば，(a) は，多くのハブページからリンクされると同時に，よりマイナーなコミュニティ

(つまり，図における下方のコミュニティ) からもリンクされる．また，(b) が VAIO のオフィシャルページであるとすれば，(a) と同様のことが (b) にも当てはまる．(a) や (b) と下位コミュニティ間のリンクは，図 2 においては，右側のサブグラフからページ (a) へのリンクに相当する．図 2 で見たように，これらのリンクは HITS 系手法では意味を成さないが，Max-Flow 手法では重要な役割を果たす．結果として，図 3 中のトピック規模の軸において，HITS コミュニティが横方向に伸びるページ集合であるとすると，Max-Flow コミュニティは，縦方向に伸びるページ集合になる．

以下に，Max-Flow コミュニティと HITS 系コミュニティの相違をまとめる．

- Max-Flow コミュニティと HITS 系コミュニティは異なるページ集合から成る．
- Max-Flow コミュニティは，異なるトピック規模のページから成り，HITS 系コミュニティは，同じトピック規模のページから成る．

4 実験による特徴検証

前節では，グラフ構造の相違がもたらすコミュニティの特徴相違を示した．本節では，これらの特徴を実際のウェブデータを用いて検証する．検証には，ウェブアーカイブから抽出した Max-Flow，HITS コミュニティからなる 2 つのコミュニティ集合を用いる．以下に利用したデータ及び，コミュニティ集合の概要を示す．

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．

コミュニティ集合：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シード集合を構築する．シード集合内の各ノードに対して，第 2 節で述べた手法を用いてコミュニティを求め，それらの集合をコミュニティ集合として用いる．総シードノード数は，約 132 万であり，各コミュニティ集合のコミュニティ総数は，Max-Flow：約 132 万，HITS:約 120 万であった．Max-Flow コミュニティの総数は，シードページ数に相当し，HITS コミュニティを構成するメンバーは，シード集合内のページのみから成ることに注意されたい．

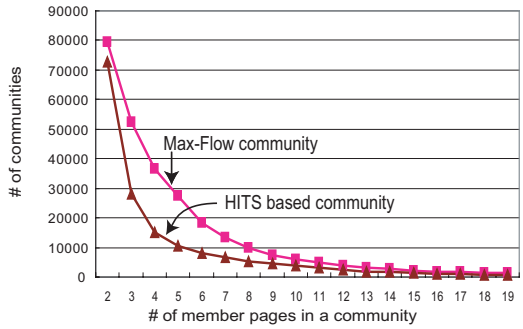

図 4: サイズ分布：Max-Flow コミュニティ，HITS コミュニティ

図 4 のグラフは，横軸が，各コミュニティのサイズ (つまり，メンバーページの数)，縦軸がコミュニティ数を示している．Max-Flow コミュニティ全体のうち約 65.24%(=858523comms.)，HITS コミュニティ全体のうち，約 85.39%(=1025937comms.) が単一ページであった．そのうち 269956 ページが両者に共通していた．20 以上のメンバーページから成るコミュニティ数は，Max-Flow：183788，HITS：6391 であった．

4.1 検証概要

検証 1：一つ目の特徴「Max-Flow コミュニティと HITS 系コミュニティは異なるページ集合から成る」を検証する．前述のように，HITS コミュニティのメンバーは，シードセット内のページ (つまり，3 つ以上の異なるサーバーからリンクされているページ) のみから成る．一方，Max-Flow コミュニティのメンバーは，シードセット以外のページも多く含む．これに注意し，以下の手順で検証を行う．

1. 各 Max-Flow コミュニティに含まれるシードセット内のページ数を調査．
2. 一つの Max-Flow コミュニティが何種類の HITS コミュニティに対応するか調査．

手順 2 は，一つの Max-Flow コミュニティに含まれるシードセット内のページが，HITS では全て同じ一つのコミュニティに属しているのか，全て異なるコミュニティに属しているのかを調査することを意味している．当然のことながら，前者の場合は，Max-Flow コミュニティと HITS コミュニティでは，同じページ集合から成り，後者の場合は，異なるページ集合から成るということが言える．

検証 2：二つ目の特徴「Max-Flow コミュニティは，異なるトピック規模のページから成り，HITS 系コミュニティは，同じトピック規模のページから成る」を検証する．例えば，メジャーリーグチームとある地方に存在する高校野球チームの 2 つのページを考えた時，前者のページへのリンク数の方が後者よりもずっと多いであろうことが容易に予想できる．このように，概して，トピックがより一般的，メジャーに成ればなるほど，多くのページからリンクされるようになる．よって，コミュニティのトピック規模を把握するためにメンバーページの平均入リンク数を用いる．図 5 に HITS コミュニティの平均インリンク数を示す．横軸が，HITS コミュニティの平均インリンク数，縦軸がそれに該当するコミュニティ数を示している．

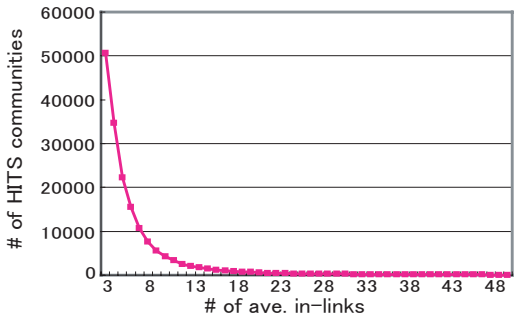

図 5: HITS コミュニティの入リンクの平均数

検証手順を以下に示す．

1. 各 HITS コミュニティの平均入リンク数を求め，全ての HITS コミュニティを 4 つのトピックレベルに分類．
2. 各 Max-Flow コミュニティが対応する HITS コミュニティのトピックレベルを求め，後述するレベルクロス数を調査．

検証で利用する 4 つのトピックレベルは，レベル 1 が平均入リンク数が 50 以上，レベル 2 が 10~49，レベル 3 が 5~9，レベル 4 が 3~4 とする．図 6 にこのように設定されたトピックレベルに分類される HITS コミュニティの割合を示す．レベルクロス数とは，一つの Max-Flow コミュニティが対応している HITS コミュニティが属するレベルの数を意味している．例えば，ある Max-Flow コミュニティのメンバーがレベル 3,4,2,2 の HITS コミュニティに分散しているとき，3 種類のレベルの HITS コミュニティに対応しているので，クロス数は 3 となる．クロス数が 1 の場合は，Max-Flow コミュニティにおいても同じトピックレベルのページ集合であると言え，逆にクロス数が複数の場合は，異なるトピックレベルのページ集合であると言える．

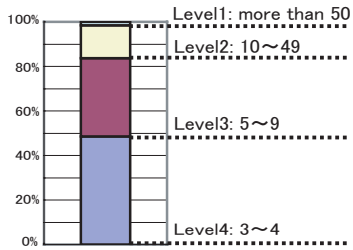

図 6: 入リンク数によるトピックレベル

4.2 検証結果

検証結果 1：図 7 のグラフは，それぞれのサイズの Max-Flow コミュニティ(横軸) に対し，それらの Max-Flow コミュニティに含まれるシードセット内のページ数の平均値 (上側)，及び，一つの Max-Flow コミュニティに対応している HITS コミュニティの平均数 (下側) を示している．

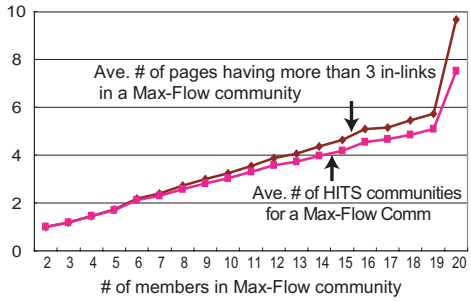

図 7: 検証結果 1

例えば，メンバー数が 16 の Max-Flow コミュニ

ティに属する平均シードページ数は，5.08 で，対応する HITS コミュニティ平均数は，4.54 となっていた．平均シードページ数に対する対応 HITS コミュニティ平均数の割合は，メンバーページ数が 20 以上の Max-Flow コミュニティについては，約 77.81%，19 以下における平均値は，約 93.80%であった．これは，Max-Flow コミュニティに含まれるシードページの殆どが重複すること無しに，異なる種類の HITS コミュニティに属していることを意味し，Max-Flow と HITS コミュニティは，異なるページ集合を成していることがわかった．

検証結果 2：図 7 に，Max-Flow コミュニティのレベルクロス数の割合 (下)，HITS コミュニティのメンバーページにおけるレベルクロス数の割合 (上) を示す．HITS コミュニティのメンバーページの入リンク数のレベルクロス数の約 80%が，1 または 2 であった．また，2 に属するうちの多くに，入リンク数が 4 と 5 のメンバーページを持つものが観察された．これより，HITS コミュニティは同じトピックレベルのページから成ることが確認され，実験で設定したレベル分類は妥当であったと言える．これに対し，同じ一つのレベルの HITS コミュニティに対応している Max-Flow コミュニティは，全体の約 7%に過ぎず，対応する HITS コミュニティが 3 つ，或いは，全てのトピックレベルに跨っている Max-Flow コミュニティが約 60%を占めた．これは，一つの Max-Flow コミュニティに複数のレベルの HITS コミュニティが対応していることを意味し，Max-Flow コミュニティが異なるトピックレベルのページから成ることが言える．

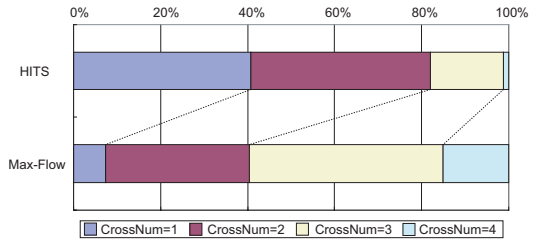

図 8: トピックレベルクロス数：HITS vs Max-Flow

4.3 考察

以上の検証結果より Max-Flow コミュニティは，異なるトピックレベルのメンバーページから成ることがわかった．上で挙げたコンピュータメーカーの他に実際に観察できた具体例としては，各種スポーツ協会・団体 (サッカー，野球，バレーボール，etc...) の HITS コミュニティに対応する Max-Flow コミュニティは，例えばバレーボール協会，地域のバレーボールチームのコミュニティなど，一つのスポーツ協会ページとそのスポーツに関わるページ集合となっている．また，地方自治体が作成する各地の観光協会 (倉敷市，広島市，熊本県，etc...) の HITS コミュニティに対応する Max-Flow コミュニティは，例えば倉敷観光協会のページと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から成るコミュニティなど，一つの地域の観光協会のポータル的なサイトと個人やその地域所在の商店やホテルなどのサイトからなるコミュニティとなっている．

5 まとめと今後の課題

本稿では，Max-Flow 手法および HITS 系手法がコミュニティ抽出の指針とするグラフ構造に着目して，それらのコミュニティが示す特徴差を分析した．その結果，Max-Flow コミュニティと HITS 系コミュニティは異なるページ集合から成り，前者は，異なるトピック規模のページから成り，後者は，同じトピック規模のページから成るということがわかった．また，2002 年の日本国内のウェブページから抽出したコミュニティから成る大規模コミュニティセットを利用した実験を行い，この分析の正当性を検証した．

HITS 手法によりコミュニティを抽出するためには，顕著なハブ・オーソリティの傾向を示すページの存在が必須となる．一方，Max-Flow 手法は，リンク連結度の内外差のみからコミュニティを抽出するため，ハブに相当するリンク集が存在しなくてもコミュニティを抽出できるという利点がある．今後の課題として，Max-Flow 手法でのみ得られるコミュニティに着目し，これらのコミュニティの時系列変化を追い，ウェブにおけるトピックの時間軸上でのグラフ変化の特徴付けに取り組む予定である．

参考文献

[1] G.Flake, S.Lawrence, C.L.Giles.: Efficient identification of web communities. 6th ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, pages 150–160 (2000)

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

[3] R.Kumar, P.Raghavan, S.Rajagopalan, A.Tomkins.: Trawling the Web for emerging cyber-communities. Proc. of the 8th International World Wide Web Conference, pages 1481–1493 (1999)

[4] J.M.Kleinberg.: Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632 (1999)

[5] S.Chakrabarti, B.Dom, D.Gibson, J.Kleinberg, P.Raghavan, S.Rajagopalan.: Automatic resource list compilation by analyzing hyperlink structure and associated text. Proc. of the 7th International World Wide Web Conference, pages 65–74 (1998)

[6] D.Gibson, J.M. Kleinberg, P.Raghavan.: Inferring web communities from link topology. UK Conference on Hypertext, pages 225–234 (1998)

[7] S.R. Kumar, P.Raghavan, S.Rajagopalan, A.Tomkins.: Extracting large-scale knowledge bases from the web. The VLDB Journal, pages 639–650 (1999)

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

[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. IEICE transactions on Information and Systems, Vol.E87-D No.2 pages 407–415 (2004)

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

[12] 今藤紀子，喜連川優：Max-Flow コミュニティグラフとその特徴分析．DBSJ Letters Vol.3, No.1 掲載予定　(2004)