社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－DBS－140（Ⅱ）（84）
2006／7／14

リンク構造を用いたウェブコミュニティ抽出法

大塚　浩司†　　大町　真一郎†　　阿曽　弘具††

†,†† 東北大学大学院 工学研究科
〒 980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-05
E-mail: †{koji6,machi}@aso.ecei.tohoku.ac.jp, ††aso@ecei.tohoku.ac.jp

あらまし　共通のトピックに関するウェブページの集合であるウェブコミュニティを抽出する手法の 1 つとして HITS アルゴリズムが知られている．HITS アルゴリズムは，ウェブコミュニティを“オーソリティ”と“ハブ”と呼ばれる 2 種類のウェブページからなる 2 階層の構造をしていると仮定し，これを 2 部グラフと見なして抽出することを目的としている．しかし，一般にウェブコミュニティはより複雑なリンク構造を持っている．本論文では，3 階層以上の多階層のリンク構造を持つウェブコミュニティを抽出することを目的とし，オーソリティ・ハブに加えて中間ノードを導入したウェブコミュニティ抽出法を提案する．提案手法により，HITS アルゴリズムでは抽出しにくいページをウェブコミュニティのメンバとして抽出することが可能であることを実験により示す．

キーワード　ウェブコミュニティ, ウェブグラフ, HITS アルゴリズム

Web Community Extraction Using Link Structure

Koji OHTSUKA†, Shinichiro OMACHI†, and Hirotomo ASO††

†,†† Gradurate School of Engineering, Tohoku University
6-6-05 Aoba, Aramaki, Aoba-ku, Sendai, 980-8579 Japan
E-mail: †{koji6,machi}@aso.ecei.tohoku.ac.jp, ††aso@ecei.tohoku.ac.jp

Abstract　A web community is a set of web pages about a common topic. HITS (Hyperlink-Induced Topic Search) algorithm is a method for extracting web communities. HITS algorithm presumes that a web community consists of bipartite link structure of authority nodes and hub nodes. However, a web community usually has more complex structure. In this paper, we propose an algorithm for extracting web communities by presuming that a web community consists of n-layer structure and by introducing medium nodes (a page to which many pages links to and from which there are many links to other pages). Experimental results show pages that are hard to be extracted by HITS algorithm can be extracted by the proposed method.

Key words　web community, web graph, HITS algorithm

1. はじめに

インターネットが普及して多くの人が利用している．インターネット上に存在する膨大な情報の中で，ユーザが必要とするあるトピックに関するページ群はウェブコミュニティと呼ばれる．膨大な情報の中からウェブコミュニティを抽出して必要な情報を得るために Google [1] 等の検索サイトが必要とされている．検索サイトでの表示は，一般的にキーワードを入力してそのキーワードに関連のあるページ群をランク付けして表示するという形が取られている．しかし，Google 等の既存の検索サイトにはキーワードを含むページしか抽出していない事が欠点として挙げられる．

ウェブコミュニティを抽出する方法の 1 つに HITS (Hyperlink-Induced Topic Search) アルゴリズム [2] がある．HITS アルゴリズムでは，既存の検索サイトを利用してあるトピックに関するページ群を収集し，集めたページ群の中から“オーソリティ”と“ハブ”と呼ばれる属性を強く持つページを見つけてランク付けすることでウェブコミュニティを発見する．しかし，ウェブコミュニティはより複雑な構造を持っていると考えられるため，オーソリティとハブの 2 階層の構造を想定している HITS アルゴリズムではウェブコミュニティとして抽出すべきページを抽出することができないことがあると考えられる．

また，HITS アルゴリズムは単体で用いた場合，オーソリティやハブに当たらないページをオーソリティ, ハブとして誤って抽出してしまう事があるため，リンクに対する重み付け [3]，行列固有値計算時のフィルタリング [4]，Base 集合作成時のフィルタリング [4] の改良法が提案されている．これらの改良法でより適切なページをオーソリティやハブとして抽出することはできるが，上で述べたようなウェブコミュニティを 2 階層と仮定する点は変わらないため，抽出できるウェブコミュニティには制限があると考えられる．

そこで，本報告ではウェブコミュニティは 3 階層以上の多階層のリンク構造を持っていると仮定して，オーソリティのページとハブのページの間に位置する中間ノードのページ（複数のページからリンクを受けて，なおかつ複数のページにリンクを出しているページ）を“ミディアム”ページと定義し，オーソリティ，ミディアム，ハブの 3 つの属性を使ってウェブコミュニティを抽出する手法を提案する．特に HITS アルゴリズムで抽出していない重要なページが提案手法で抽出できているか，という点について実際のウェブを使った実験で調べて提案手法の有効性を示す．

2. HITS アルゴリズム [2]

HITS アルゴリズムは，ウェブページのリンク関係から特定のトピックに関するある種のランク情報（ウェブコミュニティ内のページの“オーソリティ”,“ハブ”という 2 属性について）を抽出するアルゴリズムである．オーソリティは図 1(a) の黒いノードのように特定の話題について他の多くのページからリンクされているページの属性を表し．ハブは図 1(b) の白いノードのように多くのオーソリティのページへのリンクを持つページの属性を表す．

HITS アルゴリズムの手順を以下に示す．

（1） トピックを示すキーワードを含むページを既存の検索サイトを使って r 件収集して Root 集合とする．

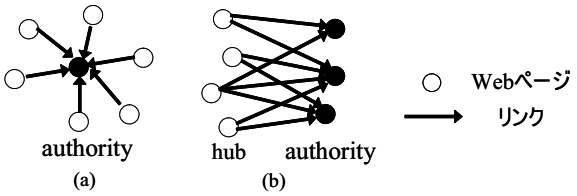

図 1 (a) オーソリティ, (b) ハブ

（2） Root 集合内のページがリンクしている全てのページ，及び Root 集合のページがリンクされているページ最大 d 件を収集し，それらのページ集合を Root 集合に追加して Base 集合を作成する．

（3） Base 集合内のページ間のリンクを全て洗い出す．この際, 同じドメイン下にあるページを結ぶリンクを全て削除し，異なるドメイン間のページを結ぶリンクのみを残す．

（4） Base 集合内の各ページ ρ に対しオーソリティ値 $a(\rho)$，ハブ値 $h(\rho)$ を定める．まず，それらの初期値を 1 とする．$a(\rho)$ と $h(\rho)$ の値を (1),(2) 式で更新して，その都度正規化する（収束したかどうかを判定するため）．$a(\rho)$ と $h(\rho)$ の値が収束するまでこの更新の処理を繰り返す．

$$a(\rho) = \sum_{\delta,\delta\to\rho} h(\delta) \tag{1}$$

$$h(\rho) = \sum_{\delta,\rho\to\delta} a(\delta) \tag{2}$$

（5） 収束したらオーソリティ，ハブの値がそれぞれ上位のページとそれらのオーソリティ値，ハブ値を出力する．

(1),(2) 式について，Base 集合の大きさを n として，ページ間のリンクを示す $n \times n$ 隣接行列を L （ページ δ からページ ρ へのリンクがあれば $L_{\delta\rho} = 1$，リンクが無ければ $L_{\delta\rho} = 0$），各ページのオーソリティ値，ハブ値を並べた n 次元ベクトルを $\vec{a}, \vec{h}$ で表すと，(1),(2) 式はそれぞれ (3),(4) 式で表すことができる．

$$\vec{a} = L^T \vec{h} \tag{3}$$

$$\vec{h} = L\vec{a} \tag{4}$$

これらの方程式より，$\vec{a} = L^T L\vec{a}$ であるので，$\vec{a}$ は $L^T L$ の固有ベクトルの 1 つである．同様に $\vec{h}$ は LL^T の固有ベクトルの 1 つである．

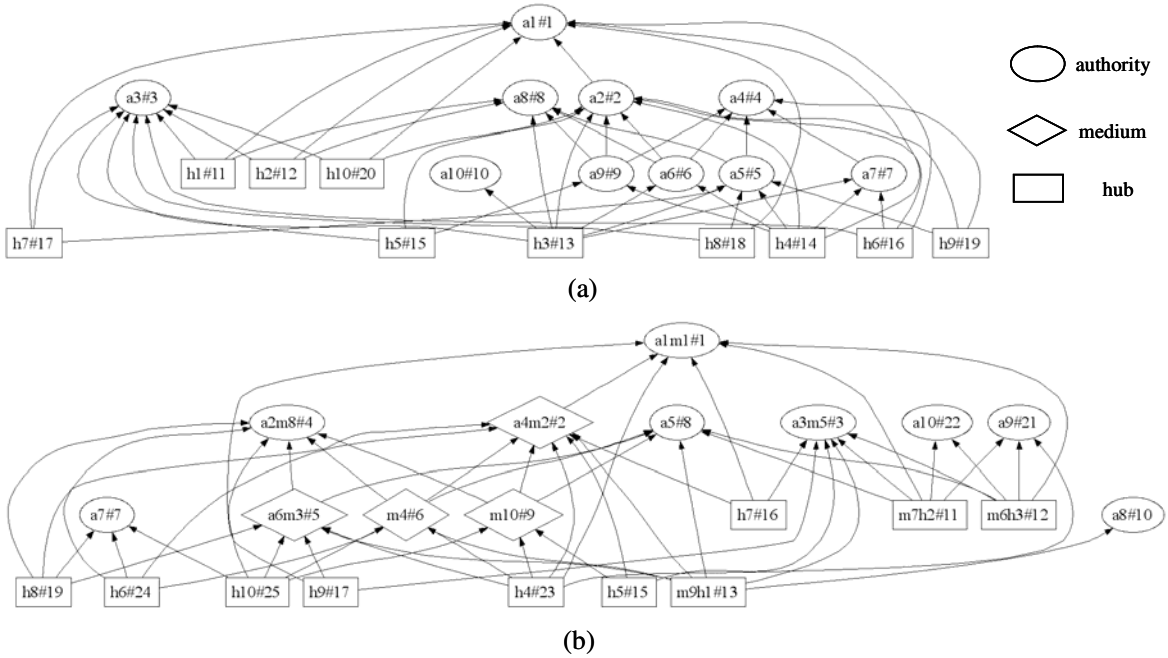

図 4 “東北大学”のウェブコミュニティのグラフ（(a) HITS アルゴリズム，(b) 提案手法）（書式例:a2m8#4，a の 2 位，m の 8 位，表 1 の番号#4，同じページがランキングに重複して入った場合は順位の高い方の形にした）

3. n 階層の構造を仮定したウェブコミュニティの抽出手法

3.1 ウェブコミュニティの構造の仮定

ウェブコミュニティの構造のモデルについては，様々なものが提案されている [5]．これらを踏まえて，我々はウェブコミュニティを示すグラフ構造のモデルとして図 2 の構造を仮定する．すなわち，最上層に複数のオーソリティのページがあって，その下に複数のノードが複数の階層に渡って複雑にリンクしている構造である．ただし，ウェブコミュニティの中にリンクのループは存在しないものとする．

3.2 中間ノードを考慮した抽出手法

図 2 のようなウェブコミュニティの構造に含まれているもので図 3 のようにオーソリティのページから 2 つ以上のリンクをたどった先にあるページまでをウェブコミュニティのメンバとして抽出する方法を提案する．

まず，オーソリティとハブの間にある中間ノードを示す属性ミディアムを図 3 の黒で示したページのように，ハブやミディアムのページからリンクを受け，かつミディアムやオーソリティのページにリンクを出しているページの属性と定義する．さらにオーソリティとハブの定義を以下のように変更する．オーソリティはミディアムとハブのページからリンクされているページの属性，ハブはオーソリティとミディアムのページにリンクしているページの属性とする．

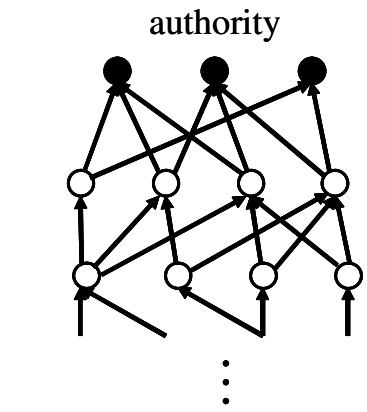

図 2 ウェブコミュニティのモデル

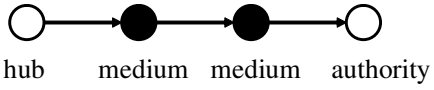

図 3 オーソリティ，ミディアム，ハブ

ここで HITS アルゴリズムの式 (3),(4) と同様の考えで属性ミディアムを考慮した基本の更新式を考える．各ページのミディアム値を表す n 次元ベクトルを $\vec{m}$ とする．L を掛けた項は他のページにリンクを出している時に値を大きくし，L^T を掛けた項は他のページからリン

クを受けている時に値を大きくする効果があるので，属性ミディアムの持つ性質を考慮して以下の (5),(6),(7) 式を 3 つの属性値を求める際の更新式として考えることができる．

$$\vec{a} = L^T(\vec{m} + \vec{h}) \quad (5)$$

$$\vec{m} = L(\vec{a} + \vec{m}) + L^T(\vec{m} + \vec{h}) \quad (6)$$

$$\vec{h} = L(\vec{a} + \vec{m}) \quad (7)$$

しかし，これらの (5),(6),(7) 式を使ってウェブコミュニティの抽出を行なうと，他のページからのリンクを多く受けているページはオーソリティ値とミディアム値の両方が大きくなり，他のページに多くリンクを出しているページはミディアム値とハブ値の両方が大きくなってしまう（オーソリティとミディアム，ミディアムとハブの更新式に同じ項があるため）．そこで，各ノードがオーソリティ，ミディアム，ハブのどれであるかを明確に区別できるように抑制項を含む次の (8),(9),(10) 式で $\vec{a}, \vec{m}, \vec{h}$ を更新するこれらは，$\vec{a}$, $\vec{m}$, $\vec{h}$ に関する方程式になっている．ミディアムの式はそのままとする．

$$\vec{a} = L^T(\varepsilon\vec{h} + \vec{m}) - \alpha(L\vec{a} + \vec{m}) \quad (8)$$

$$\vec{m} = L(\vec{a}+\vec{m}) + L^T(\vec{m}+\vec{h}) \quad (9)$$

$$\vec{h} = L(\varepsilon\vec{a} + \vec{m}) - \beta(L^T\vec{h} + \vec{m}) \quad (10)$$

ただし，$\varepsilon, \alpha, \beta$ は正の定数である．

(8) 式の第 2 項は，あるページがミディアムやハブの属性値が大きい場合にオーソリティ値を小さくする効果がある．(10) 式の第 2 項は，あるページがオーソリティやミディアムの属性値が大きい場合にハブ値を小さくする効果がある．定数の値は実験的に $\alpha = \beta = 1.0$ とした．

また，(8),(10) 式の中の ε は，オーソリティ–ハブという 2 層の関係よりもオーソリティ–ミディアム–ハブという 3 層以上の関係を重視するために導入している．すなわち，$\vec{a}$ に対して $\vec{h}$ からの影響を低く，$\vec{h}$ に対して $\vec{a}$ からの影響を低く評価している．定数の値は実験的に $\varepsilon = 0.1$ とした．なお，(8),(10) 式により $a(\rho) < 0, h(\rho) < 0$ となった場合は $a(\rho) = 0, h(\rho) = 0$ と再設定して次の反復計算を行なうことにした．

4. 実　　験

HITS アルゴリズムと提案手法を使って実際にウェブコミュニティを抽出する実験を行った．

実験条件は以下のようにした．

- キーワード:“ 東北大学 ”,“ 東京大学 ”
  “ データベース ”,“ J リーグ ”,“ ベガルタ仙台 ”
- $r = 50, d = 30$
- 追加ページの収集では 2 ステップ先のリンクのページ
  まで収集

第 1 章で示した HITS アルゴリズムの改良法の中で HITS アルゴリズムでは 3 種類全部，提案手法では重み付けと Base 集合作成時のフィルタリングを使用した．

結果について，まずキーワード“ 東北大学 ”で実験を行なった時のランキング上位のページとそれらのページから成るグラフをそれぞれ表 1，図 4 に示す．図 4 より提案手法の方が HITS アルゴリズムより複雑で多階層の構造を抽出できている．また，オーソリティのページをグラフ上の一番上に，ミディアムのページをグラフ上の中間に，ハブのページをグラフ上の一番下のページに位置するように提案手法のパラメータ $\varepsilon, \alpha, \beta$ を調整した．次に，HITS アルゴリズムのみで抽出できたページ数と

表 2　HITS アルゴリズムと提案手法の結果の比較
（片方の結果のみで抽出できたページ数）

	提案手法 ページ数			HITS ページ数		
キーワード	auth.	med.	hub	auth.	hub	計
東北大学	4	1	1	3	3	6
東京大学	5	1	0	3	3	6
データベース	2	1	2	3	2	5
J リーグ	23	8	15	22	24	46
ベガルタ仙台	1	0	2	0	3	3

注）auth.:オーソリティ，med.:ミディアム，hub:ハブ

提案手法のみで抽出できたページ数を調べた．HITS アルゴリズムのオーソリティ，ハブの上位のページ群 50 ページと，提案手法のオーソリティ，ミディアム，ハブの上位のページ群 50 ページを比べて片方の手法のみで抽出されているページ数を表 2 に示す．この結果より，まず提案手法と HITS アルゴリズムで異なるページを抽出していることが分かる．特にキーワード“ J リーグ ”では 50 ページ中 46 ページが異なっていて，既存手法と提案手法で全く異なるページ群を抽出していると言える．

さらに，キーワード“ 東北大学 ”,“ J リーグ ”について提案手法，HITS アルゴリズムの一方だけで抽出できた 6 ページ及び 46 ページ中上位 20 ページの内容を表 3，表 4 に示す．この結果について，キーワード“ 東北大学 ”では両手法で異なって抽出されたページは東北大学に関連するページだった．キーワード“ J リーグ ”ではオーソリティに関しては HITS アルゴリズムではスポーツ関連の報道機関のページが多く抽出され，提案手法では J リーグのチームのページが多く抽出された．“ J リーグ ”というコミュニティのメンバを考えた時に後者の J リーグのチームの方がメンバとしてふさわしいので提案手法には有効性があると言うことができる．ミディアム，ハブに関しては両手法ともに J リーグに関するリンク集

表 1　(a) HITS アルゴリズムでのランキング, (b) 提案手法でのランキング（# は図 4 の番号）

(a)

順位	オーソリティ	ハブ
1	#1　東北大学 Top	#11 東北大学理学研究科 地震・噴火予知観測センター リンク集 (index_html)
2	#2　東北大学工学部・工学研究科	#12 東北大学理学研究科 地震・噴火予知観測センター リンク集
3	#3　東北大学理学研究科	#13 東北大学組織一覧
4	#4　東北大学 Top(index-j.html)	#14 電気通信研究所 矢野研究室 リンク集
5	#5　東北大学附属図書館	#15 関西学院大学理工学部 物理学科 リンク集
6	#6　東北大学電気通信研究所	#16 Z 会大学院リスト
7	#7　東北大学大学院情報科学研究科	#17 理学研究科 大気海洋変動観測研究センター リンク集
8	#8　東北大学情報シナジーセンター	#18 東北大学物性理論研究室
9	#9　東北大学電気・情報系 (ECEI)	#19 電気通信研究所図書館
10	#10 東北大学流体科学研究所	#20 カナモト 大学へのリンク

(b)

順位	オーソリティ	ミディアム	ハブ
1	#1　東北大学 Top	#1　東北大学 Top	#13 東北大学 組織一覧
2	#4　東北大学 Top(index-j.html)	#2　東北大学工学部・工学研究科	#11 東北大学理学研究科 地震・噴火予知観測センター リンク集 (index_html)
3	#3　東北大学理学研究科	#5　東北大学 附属図書館	#12 東北大学理学研究科 地震・噴火予知観測センター リンク集
4	#2　東北大学工学部・工学研究科	#6　東北大学電気通信研究所	#23 東北大学理学研究科地球物理学専攻リンク集
5	#8　東北大学情報シナジーセンター	#3　東北大学 理学研究科	#15 関西学院大学理工学部 物理学科 リンク集
6	#5　東北大学附属図書館	#12 東北大学理学研究科 地震・噴火予知観測センター リンク集	#24 東北大学電気情報・物理工学科教務
7	#7　東北大学大学院情報科学研究科	#11 東北大学理学研究科 地震・噴火予知観測センター リンク集 (index_html)	#16 Z 会 大学院リスト
8	#10 東北大学流体科学研究所	#4　東北大学 Top(index-j.html)	#19 電気通信研究所図書館
9	#21 仙台市ホームページ	#13 東北大学 組織一覧	#17 東北大学理学研究科 大気海洋変動観測研究センター リンク集
10	#22 宮城県ホームページ	#9　東北大学電気・情報系 (ECEI)	#25 電気通信研究所 伊藤研究室

やブログのページを抽出していた.

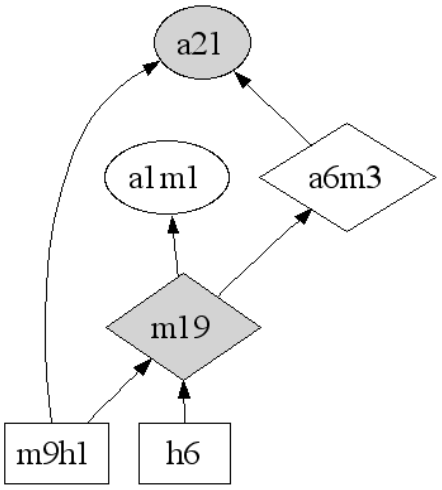

図 5　提案手法“東北大学”での結果の一部
（a21,m19 に注目，全体のグラフから a21,m19 のページと
直接リンクを張っているページのみ表示）

ここで，提案手法で抽出したキーワード“東北大学”の結果のオーソリティ21 位，ミディアム 19 位のページに注目したグラフを図 5 に示す．これを見ると分かるように $[h1, h6 \rightarrow m19 \rightarrow a6m3 \rightarrow a21]$ と多階層の構造としてページ群を抽出できており，その中のオーソリティ 21 位に当たるページとミディアム 19 位に当たるページ（a21 東北大学資料館，m19 東北大学 高等教育開発推進センター，“東北大学”のコミュニティとして有効なページ）を提案手法では上位で求めることができている（HITS アルゴリズムではオーソリティ，ハブ合わせた順位で 50 位以下になっている）．この事からは提案手法の有効性を示していると言える．

5. おわりに

本報告では，オーソリティから 2 階層以上離れているため HITS アルゴリズムでは抽出しにくいページをウェブコミュニティとして抽出するために，ウェブコミュニティの構造を n 階層と定義して，中間ノードの性質を示すミディアム属性を導入した．そして，ミディアム属性を加えたウェブコミュニティ抽出のための属性値計算規

表 3　HITS アルゴリズムで抽出できたページのリスト
(a) キーワード“東北大学”

属性・順位	ページ内容
a17	東北大学情報シナジーセンター 計算システム
a19	東北大学附属図書館 工学分館
a20	東北大学 HP
h22	東北大学理学研究科地球物理学専攻
h23	東北大学 複雑系流動システム分野 リンク集
h24	東北大学高等教育開発推進センター 情報教育室

(b) キーワード“J リーグ”

属性・順位	ページ内容
a2	J リーグ ファンサイト
a4	toto サイト
a5	J リーグ選手協会 —JPFA—
a6	日刊スポーツ HP
a7	Super Soccer HP
a8	サポティスタ (blog)
a9	サンケイスポーツ HP
a10	WEB サッカーマガジン
a11	スポニチ HP
a12	ホームページ制作　Neutral
h1	よく使うリンク集
h3	サッカー関係 リンク集
h4	sports spirit - soccer site - (index.html)
h5	sports spirit - soccer site -
h6	SANFRECCE Supporters' Blog
h7	SANFRECCE Supporters' Blog 内ページ
h8	SANFRECCE Supporters' Blog 内ページ
h9	SANFRECCE Supporters' Blog 内ページ
h10	SANFRECCE Supporters' Blog 内ページ
h11	SANFRECCE Supporters' Blog 内ページ

表 4　提案手法で抽出できたページのリスト
(a) キーワード“東北大学”

属性・順位	ページ内容
a11	TAINS (計算機センター) Home Page
a12	東北大学病院
a20	東北大学工学研究科・工学部 技術部
a21	東北大学資料館
m19	東北大学 高等教育開発推進センター
h19	東北大学 大学院理学研究科 数学専攻

(b) キーワード“J リーグ”

属性・順位	ページ内容
a2	F.C.TOKYO
a3	ガンバ大阪オフィシャルサイト
a4	CEREZO OSAKA OFFICIAL SITE
a5	URAWA RED DIAMONDS
a6	アルビレックス新潟 HP
a7	S-PULSE OFFICIAL WEB SITE
a8	-VISSEL KOBE- OFFICIALSITE
m2	Ｊリーグ関連サイト・リンク集
m3	中坊コラム〜J リーグコラム集サイト〜
m4	サッカーファンサイト
m6	サッカー情報ナビ
m7	J.B.Antenna-J2-
m8	J.League Drive
h2	Ｊリーグ関連サイト・リンク集
h3	J.B.Antenna-J2-
h4	中坊コラム〜J リーグコラム集サイト〜
h5	J.League Drive 内ページ (横浜)
h6	J.League Drive 内ページ (大分)
h7	J.League Drive 内ページ (M 選手)
h8	J.League Drive 内ページ (N 選手)

則を提案した．実際のウェブを対象にした抽出実験より HITS アルゴリズムで抽出しにくいページを提案手法で抽出できていることが確かめられた．

今後の課題としては以下のものが挙げられる．まず，実際のウェブでは互いに参照しあうページがあり，リンクがループしていることが多くある．これに再現するため，ループを含んでいるグラフを入力として与えて実験を行った所 HITS アルゴリズムではループの中でオーソリティ値，ハブ値が一番強くなるリンクを抽出して弱いリンクは無視する結果になったが，提案手法では属性値が振動してしまい収束しなかった．この点について改良する必要がある．

リンクを使った検索アルゴリズムではリンクファーム（密にリンクを張り合っているページ群）に高いスコアが与えられる性質があり，スパムページ（広告ページ等）がメンバとなっているリンクファームが存在すると適切なコミュニティが求められなくなる．リンク構造を利用してリンクファームを検出・リンクをカットする手法 [6] が提案されているのでこれを我々の手法に利用したい．

最後に，第 4 章で述べたように HITS アルゴリズムで抽出しにくいページを提案手法で抽出することができたが，これらのページの中で重要なページを絞り込むことが必要である．

文　献

[1] http://www.google.com

[2] J.Kleinberg :“Authoritative Sources in a Hyperlinked Environment” Research Report RJ 10076(91892), IBM, 1997.

[3] G.Chang, M.J.Healey, J.A.M.McHugh, J.T.L,Wang,“Mining the World Wide Web 〜An Information Search Approach〜”Kluwer Academic Publisher, 2001.

[4] 野村 早恵子, 小山 聡, 早水 哲雄, 石田 亨“WEB コミュニティ発見のための HITS アルゴリズムの分析と改善”電子情報通信学会論文誌 D-I Vol.J85-D-I, No.8, pp.741-750, 2002.

[5] Wen-Jun Zhou, Ji-Rong Wen, Wei-Ying Ma and Hong-Jiang Zhang“A Concentric-Circle Model for Community Mining in Graph Structures”Technical Report MSR-TR-2002-123 Microsoft Corporation, 2002.

[6] Baoning Wu and Brian D.Davison,“Identifying Link Farm Spam Pages”in Proceedings of the 14th International World Wide Web Conference, 2005.