F-010

# 多階層のリンクを考慮したWebコミュニティの抽出
# Web Community Extraction Considering Hierarchical Links

大塚 浩司† 大町 真一郎† 阿曽 弘具†
Koji Ohtsuka Shinichiro Omachi Hirotomo Aso

## 1. はじめに

キーワードでWebページを検索するとき，希望するトピックに関連するページとは異なるページが多数含まれることが多い．ある希望するトピックに関するWebページの集合は“Webコミュニティ”と呼ばれ，キーワードに関する適切なWebコミュニティを抽出できることが望まれている．

Web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 [1] を使ったものがある．HITSアルゴリズムとは“authority”と“hub”の2つのタイプのページを考え，収集したページがその2つの性質をどれくらい持っているかを求めることでWebコミュニティを抽出する方法である．

しかし，ツリー構造などのようにauthorityのページから多階層下のページまでWebコミュニティとして抽出することを想定した場合，HITSアルゴリズムを用いたWebコミュニティ抽出ではauthorityとhubの2つのページ間の関係を使っているため，複数のリンクを介して繋がっているページの抽出が素直でない．本研究ではauthorityとhubの間にある中間ノードとなるはずのページの性格付けを行なうためmedium値を導入して，3つのパラメータを使ってWebコミュニティを抽出する方法を提案する．

## 2. HITSアルゴリズム

HITSアルゴリズムとは,Webページのリンク関係から検索ワードに対する適切な情報（ページの“authority”と“hub”への2タイプへの分類）を抽出するものである．authorityを図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と定義する．hubを図1(b)の白いノードのように多くのauthorityへのリンクを持つページと定義する．

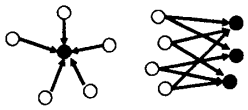


図 1: (a)authority, (b)hub

HITSアルゴリズム [1] の手順は以下のようになる．

1. 検索サイトを使ってキーワードを含むページを$r$件収集してRoot集合とする．
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，Root集合に追加してBase集合を作成する．
3. Base集合内のページで異なるドメインのページ間のリンクを全て抽出する．
4. Base集合内の各ページ$\rho$に対しauthority値$a(\rho)$,hub値$h(\rho)$を定義し，初期値を1とする．$a(\rho)$と$h(\rho)$の値を以下の式で更新し正規化する．この更新の処理を$a(\rho)$と$h(\rho)$の値が収束するまで繰り返す．

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

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

5. 収束したらauthority，hubの値がそれぞれ上位のページのURLとauthority値，hub値を出力する．

ここで，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$L$，authority値，hub値のリストをベクトル$\vec{a},\vec{h}$で表すと，式(1)(2)は以下のように表すことができる．

$$\vec{a} = L^T\vec{h} \quad (3)$$

$$\vec{h} = L\vec{a} \quad (4)$$

HITSアルゴリズムを単体で行なった場合authorityやhubの誤認を生じることがあるため，以下に示す改良法が提案されている．

1. リンクに対する重み付け (式(3)(4)の右辺に重みをつける)[2].
2. 反復計算の式(3)(4)を行列の固有値計算と見立てた場合の固有値計算時のフィルタリング [3]
3. Base集合作成時のフィルタリング [3].

## 3. 中間ノードの性格付け

Webコミュニティとしてはauthorityやhubだけでなく図2に示すような中間ノードも抽出対象とすべきと考えた．この中間ノードの性格付けのためmedium値を導入する．medium値が高いノードをmediumとし，このノードは図2の黒いノードのようにhubやmediumからリンクを受け，かつmediumやauthorityにリンクを出しているページであることが期待される．また，HITSアルゴリズムのauthority値とhub値の計算を以下のように修正する．authorityはmediumとhubのページからリンクされているページであり，hubはauthorityとmediumのページにリンクしているページとなる．medium値の

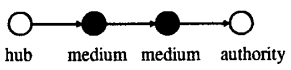


図 2: authority, medium, hub

†東北大学 大学院工学研究科

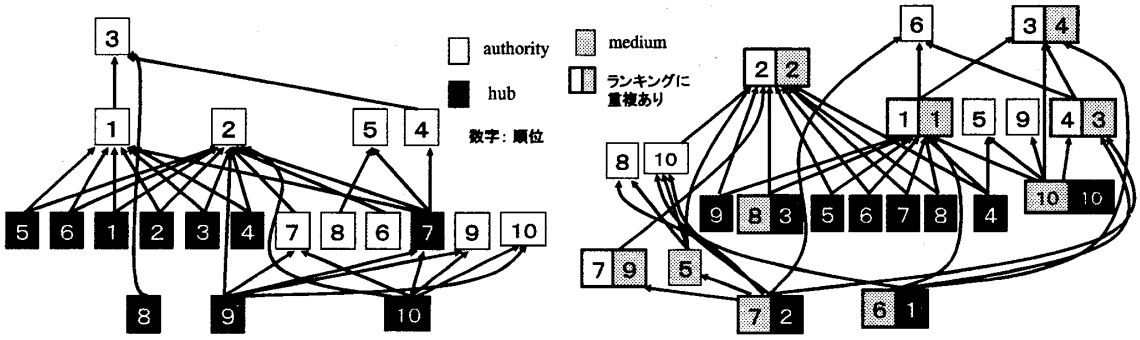


図 3: (左) HITS アルゴリズムでの結果，(右) 提案手法での結果 (グラフ)

リストを $\vec{m}$ と置いて，$\vec{a}, \vec{m}, \vec{h}$ を求める反復計算の式を以下のように定める．

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

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

$$\vec{h} = L(\vec{a} + \vec{m}) - pL^T\vec{h} - q\vec{m} \quad (7)$$

ただし，$p, q$ は定数である．また，反復計算中に $a(\rho) < 0, h(\rho) < 0$ となった場合は $a(\rho) = 0, h(\rho) = 0$ とする．式 (5) の第 2 項，第 3 項は hub 値,medium 値による抑制を，式 (7) の第 2 項，第 3 項は authority 値,medium 値による抑制をしている．

これらの式で求められた authority 値,medium 値,hub 値によるランキングの上位のページをもとに Web コミュニティを構成する．

## 4. 実験

HITS アルゴリズムと提案手法について実験を行なった．キーワードは“東北大学”，$r = 50$, $d = 30$，また $p = 0.2$, $q = 0.5$ とした．リンクは 2 階層収集した．HITS アルゴリズムでは 3 種類の改良法を，提案手法では重み付け，Base フィルタリングの 2 種類の改良法を使用した．

### 4.1 考察

実験結果を表 1，図 3 に示す．図 3 ではそれぞれ 10 位までのページについて表示したが，表 1 ではスペースの関係上それぞれ 5 位までを表示した．図 3 より，両者とも 4 階層までの抽出することができたが，提案手法の方がより複雑な構造を抽出している．ただ，今回の提案手法では表 1(b) のように authority と medium の上位のページに同一のページが入ってしまうということもあって明確な差は出てこなかった．この点は今後の課題である．

## 5. おわりに

本研究では，Web コミュニティの抽出において，HITS アルゴリズムでは抽出することが難しい authority から 2 つ以上のリンクを介してつながっているページを抽出するアルゴリズムを提案した．その結果，中間ノードとしての性格を持つページを抽出することができた．authority と medium のページを良好に分離できるような手法の検討が今後の課題である．

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

(a)

| 順位 | authority | hub |
|---|---|---|
| 1 | 東北大学付属図書館 | caos web リンク集 |
| 2 | 東北大学 Top | 複雑系 リンク集 |
| 3 | 東北大学 日本語 HP | 枝松研究室 リンク集 |
| 4 | 電気通信研究所 | 艮陵 リンク集 |
| 5 | NACSIS Webcat | 物性理論 リンク集 |

(b)

| 順位 | authority | medium |
|---|---|---|
| 1 | 東北大学付属図書館 | 東北大学付属図書館 |
| 2 | 東北大学 Top | 東北大学 Top |
| 3 | 東北大学 日本語 HP | 電気通信研究所 |
| 4 | 電気通信研究所 | 東北大学 日本語 HP |
| 5 | NACSIS Webcat | 東北大学加齢医学研 |

| 順位 | hub |
|---|---|
| 1 | 東北大学 伊藤研究室 |
| 2 | 東北大学 組織一覧 |
| 3 | caos web リンク集 |
| 4 | 矢野研究室 リンク集 |
| 5 | 艮陵 リンク集 |

## 参考文献

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

[2] 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.

[3] 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.