クラスタ係数に基づくHITSアルゴリズムの特性解析と改善

武吉　朋也†　　白石　真一†　　長谷山美紀†　　北島　秀夫†

† 北海道大学大学院情報科学研究科　〒 060–0814 北海道札幌市北区北 14 条西 9 丁目
E-mail: †{tomoya,shinichi,miki,kitajima}@md.ist.hokudai.ac.jp

あらまし　近年，特定の話題に関連するページの集合 (Web コミュニティ) を Web から抽出する手法の研究が盛んに行われている．その代表的手法の一つに，Kleinberg が提案した HITS アルゴリズムがある．しかし，この HITS アルゴリズムは，Kleinberg 自身が定義した Web コミュニティのモデルに反して，互いに密にリンクを張っているページの集合を重要な Web コミュニティとして抽出するという問題がある．そこで本文では，この原因を明らかにするため，ネットワークの集まり具合を表す指標であるクラスタ係数 (clustering coefficient) を用いて HITS アルゴリズムの特性を解析する．さらにこのクラスタ係数を HITS アルゴリズムに応用し，Kleinberg のモデルに適合しないリンクの影響を抑えた Web コミュニティの抽出法を提案する．また，提案法を用いた Web コミュニティ抽出結果と HITS アルゴリズムを用いた場合を比較し，提案法の有効性を示す．

キーワード　Web コミュニティ，HITS アルゴリズム，クラスタ係数

Characteristic Analysis and Improvement of HITS Algorithm Using Clustering Coefficient

Tomoya TAKEYOSHI†, Shin'ichi SHIRAISHI†, Miki HASEYAMA†, and Hideo KITAJIMA†

† Graduate School of Information Science and Technology, Hokkaido University
Kita-ku, Kita-14 Nishi-9, Sapporo 060–0814, Hokkaido, Japan
E-mail: †{tomoya,shinichi,miki,kitajima}@md.ist.hokudai.ac.jp

Abstract　This paper analyzes a problem of HITS algorithm which is one of the extraction method of Web communities that are sets of Web pages with a common interest on a topic. Further, we propose a new Web community extraction method that can solve this problem. The HITS algorithm tends to extract a set of Web pages whose links are dense, against the Web community model defined by Kleinberg, as an important Web community. Therefore, in order to pinpoint the cause of this problem, the characteristics of the HITS algorithm are analyzed by using a cluster coefficient which represents a measure of the density of the links. Moreover, the proposed method applies this coefficient to the HITS algorithm and realizes the accurate Web community extractions which can solve the problem. Finally, we compare the results of the proposed method and that of the traditional HITS algorithm to verify the performance of the proposed method.

Key words　Web community, HITS algorithm, clustering coefficient

1. はじめに

近年，World Wide Web(以降，単に Web と呼ぶ) の普及と拡大に伴い，Web からユーザが必要とする情報を効率良く発見する Web マイニング技術に注目が集まっている．この Web マイニングは，解析する対象に応じて，Web 内容マイニング，Web 利用マイニング，Web 構造マイニングの 3 種類に分類される [1]．本文では，Web ページ間のリンク構造を解析し，関連する Web ページの集合 (以降，単に Web ページ集合) の発見や，重要な Web ページを決定するためのランク付け等に用いられる Web 構造マイニングに着目し，その代表的手法である Kleinberg の HITS アルゴリズム [2] を取り上げ，特性の解析と改善を行う．この HITS アルゴリズムは，Web ページの重要度を決定する手法としてのみならず，Web コミュニティと呼ばれる特定のトピックに関連した Web ページ集合を抽出する手法としても利用されている [3]．HITS アルゴリズムは，検索トピックに関して的確な情報を含む Web ページ (authority) 集合と，複数の authority へのリンクをもつ Web ページ (hub) 集合をそれぞれ抽出する．本文では，文献 [3] と同様に，HITS アルゴリズムを適用して抽出された authority と hub からなる Web ページ集合を

Web コミュニティと定義する．このような特定のトピックに関連した Web コミュニティの発見手法は，Web におけるユーザの情報検索の効率化につながり有用である．

このように HITS アルゴリズムは，Web 上での有効な情報検索手法であるとされている．しかしながら，検索トピックに関連の無い Web ページ集合が密なリンク関係をもつ場合には，これらが重要な authority，hub として抽出されてしまうという問題が指摘されている [4], [5]．これに対し，Kleinberg は良質な authority と hub の特徴として，hub から authority へのリンク関係は密であり，また authority 同士のリンク関係は疎であるとしている．つまり，HITS アルゴリズムをそのまま適用して Web コミュニティを抽出した場合，リンク関係から導かれるモデル上は好ましくない Web ページ集合が重要な Web コミュニティとして抽出されてしまう．

このような背景から，本文ではページ間のリンク関係を考慮して HITS アルゴリズムを変更し，Kleinberg が良質であるとした authority，つまりリンク関係が疎である authority 集合を抽出する手法を提案する．以下，次章では HITS アルゴリズムの処理手順について説明し，これを用いた Web コミュニティ抽出結果を示す．さらに，抽出された authority 間のリンク関係を表す指標としてクラスタ係数 (clustering coefficient) [6] を導入し，HITS アルゴリズムを適用して得られた Web コミュニティを評価することで，密にリンクを張っている Web ページ集合が重要な authority として抽出される問題を確認する．また，authority の決定過程を分析し，この問題が起こる原因について考察を行う． 3. では，クラスタ係数を HITS アルゴリズムの処理手順に導入し，authority 同士のリンク関係が疎である Web コミュニティの抽出手法を提案する．さらに 4. では， 3. で提案する手法を用いた Web コミュニティ抽出結果を従来の HITS アルゴリズム [2] と比較し，提案法の有効性を示す．

2. HITS アルゴリズム

本章では，まず HITS アルゴリズム [2] の具体的な処理手順と，これを用いた Web コミュニティ(以降，単にコミュニティと呼ぶ) 抽出手法について説明する．次に，HITS アルゴリズムを用いた具体的な Web コミュニティ抽出結果を示し，さらに抽出された Web コミュニティを構成する authority 間のリンク関係を評価する指標としてクラスタ係数を導入する．また，authority を決定する行列の固有値と固有ベクトルの算出の観点から考察し，密なリンク関係をもつ Web ページ (以降，単にページと呼ぶ) 集合が重要な authority として抽出される原因を明らかにする．

2.1 HITS アルゴリズムの処理手順

HITS アルゴリズムでは，以下の手順に従って，ユーザが指定したトピックに対して関連のあるページを抽出する．

手順 (1)： 検索トピックを本文中に含むページの集合 (root 集合) を得る．

Web 上で利用可能な全文検索型のサーチエンジンに検索トピックをクエリとして入力し，本文中に検索トピックを含むページ (上位 t 件) を収集する．これにより得られたページ集合を root 集合と呼ぶ．

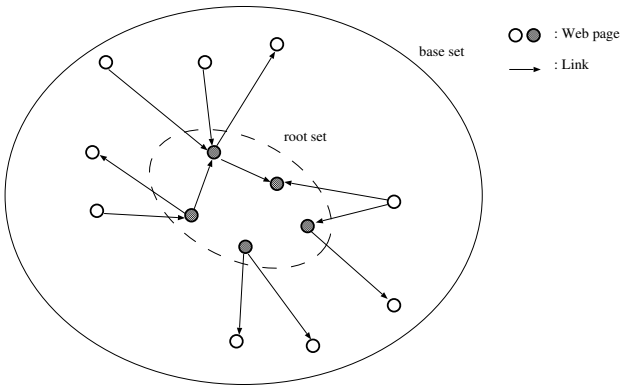

図 1　root 集合から base 集合への拡大
Fig. 1　Expanding the root set into a base set

手順 (2)： root 集合を拡大して検索トピックに関連のあるページを含むページ集合 (base 集合) を得る．

root 集合に含まれるページからリンクされているページを全て収集する．また，root 集合に含まれるページにリンクしているページを，root 集合に属する各ページあたり最大で d 件収集する．収集されたページを root 集合に加え，base 集合を得る (図 1 の実線で囲まれるページ集合)．この base 集合がリンク構造解析の対象となる．

手順 (3)： base 集合内のリンク関係を調べ，隣接行列を作成する．

base 集合に含まれるページ間のリンク関係を全て調べ，隣接行列 $L = [l_{ij}]$ を作成する．ここで隣接行列とは，ページ i からページ j へのリンクが存在する場合には $l_{ij} = 1$ とし，それ以外は 0 としたページ間のリンクによる接続関係を表す行列である．ただし，HITS アルゴリズムでは，同じドメイン下にあるページ間のリンクの存在は無視して隣接行列を作成する．

手順 (4)： 隣接行列を基に base 集合内のリンク構造を解析し，authority と hub を決定する．

隣接行列 L を基に base 集合内のリンク構造を解析し，authority と hub と呼ばれる 2 種類のページを決定する．ここで authority は，検索トピックに関して的確な情報を含むとされるページであり，また hub は，複数の authority にリンクを張っているページである．

このような authority と hub を抽出するため，HITS アルゴリズムではページ i に authority の重み a_i，および hub の重み h_i を与え，これらの値を繰り返し処理によって更新し，その収束値を求める．つまり，ページ i の authority の重みをページ i にリンクしているページの hub の重みの総和によって更新 (図 2 (a)) し，一方，ページ i の hub の重みはページ i からリンクされているページの authority の重みの総和によって更新 (図 2 (b)) される．この処理を a_i，h_i が収束するまで繰り返す．base 集合に含まれるページの総数を n とし，各ページの authority，hub の重みを成分としてもつベクトルをそれぞれ $\boldsymbol{a} = (a_1 \cdots a_n)^T$，$\boldsymbol{h} = (h_1 \cdots h_n)^T$ で表すと，上で説明した authority，hub の重みの更新は，下式で表される．

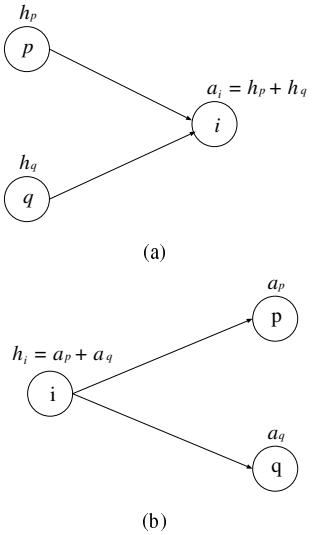

図 2　(a) authority の重み a_i の更新 (b)hub の重み h_i の更新
Fig. 2　(a) Updating of authority weight a_i (b) updating of hub weight h_i

$$\boldsymbol{a} = \alpha L^T \boldsymbol{h}. \tag{1}$$

$$\boldsymbol{h} = \beta L \boldsymbol{a}. \tag{2}$$

ここで α，β は正規化定数であり，L^T は行列 L の転置行列を表す．式 (1), (2) を繰り返し計算することにより $\boldsymbol{a}$ と $\boldsymbol{h}$ の収束値を求め，絶対値が大きな重みに対応するページをそれぞれ authority と hub として抽出する．式 (1) に式 (2) を代入すれば明らかなように，繰り返し計算によって $\boldsymbol{a}$ の収束値を求めることは，行列 $L^T L$ の最大固有値に対応する固有ベクトルを求めることに等しい．同様に，$\boldsymbol{h}$ の収束値は，行列 LL^T の最大固有値に対応する固有ベクトルとして求まる．すなわち，authority の重み a_i，hub の重み h_i は，それぞれ $L^T L$，LL^T の最大固有値に対応する固有ベクトルの第 i 成分として求まる．上記の手順で得られた authority と hub からなるページ集合をコミュニティとする [3]．

文献 [2] において Kleinberg は，「良い authority は互いにリンクを張ろうとはせず，比較的無名なページである hub を通じて接続される」と述べている．したがって，Kleinberg が理想とした authority と hub の間のリンク関係は図 3 のように表される．以降，図 3 で表されるリンク関係をもつ authority と hub，つまりコミュニティを Kleinberg のモデルに適合するコミュニティとする．

また，Kleinberg は文献 [2] において，最大固有値以外の固有値に対応する固有ベクトルからも，取り扱っている主題が異なるコミュニティが抽出され，さらに同じ固有ベクトルに含まれる成分においても，符号によって取り扱う主題の異なるコミュニティが出現する可能性を示唆している．検索トピック “abortion” を設定して Kleinberg が行ったコミュニティ抽出 [2] では，3 番目に大きな固有値に対応する固有ベクトルの成分のうち，正の値と負の値のそれぞれで絶対値が大きなページに関して，値の正負によって取り扱う話題が “pro-choice” と “pro-life” に分離されたと報告されている．本文では，k 番目に大きな固有値に対応する行列 $L^T L$，LL^T の固有ベクトルの成分をそれぞれ

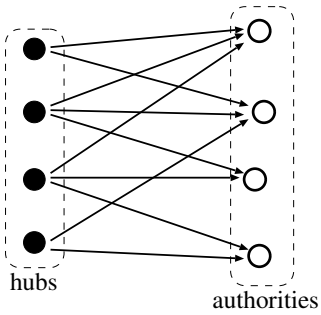

図 3　Kleinberg の Web コミュニティモデル
Fig. 3　Kleinberg's model of a Web community

authority，hub の重みと考える．そして，これらの重みに関して正の値と負の値のそれぞれで絶対値が大きなページに注目し，どちらかに共通のトピックに関連するページ集合が現れた場合には，このページ集合を第 k コミュニティと呼ぶ．また，このようなコミュニティを構成するページを第 k コミュニティの authority，hub とする．さらに本文では，より大きな固有値に対応するコミュニティを上位のコミュニティとする．

2.2　HITS アルゴリズムを用いて抽出されるコミュニティ

本節では，HITS アルゴリズムを用いた具体的なコミュニティ抽出結果を示し，Kleinberg のモデルに適合しないコミュニティが抽出される問題点を確認する．HITS アルゴリズムによるコミュニティ抽出では，検索トピックとしては “Artificial Intelligence(以降，AI と略記)” と “ruby” を設定した．また，手順 (1) では，文献 [2] と同様にサーチエンジン AltaVista [7] を利用して root 集合を作成した ($t = 100$)．手順 (2) の root 集合に含まれるページにリンクしているページの検索にも AltaVista を利用し，base 集合を作成した ($d = 50$)．上記の設定でコミュニティ抽出を行った結果，“AI” に関しては良好なコミュニティ抽出結果が得られたが，“ruby” については文献 [4], [5] で指摘されている問題が発生した．以下で，個々の検索トピック毎に，コミュニティ抽出結果の詳細を説明する．

a）　トピック “AI”

トピック “AI” を設定して得られた第 1 コミュニティを構成するページのうち，authority の重みの絶対値の大きいものから上位 5 件の URL を表 1 に示す．これらの authority の内容を確認した所，第 2 位と第 5 位の authority が**人工知能**に関するページであったが，これら以外のページは検索トピックと直接関連の無いページであった．また，第 2 コミュニティを構成するページのうち，authority の重みの大きいものから上位 5 件の URL を表 2 に示す．第 2 コミュニティの第 1 位と第 2 位の authority はそれぞれ AAAI(American Association for **Artificial Intelligence**) と JAIR(Journal of **Artiricial Intelligence** Research) のページであり，これら以外の authority も全て検索トピックに関連のあるページであった．紙面の都合上，表には示していないが，これら第 1，第 2 コミュニティの authority を上位 10 件まで比較すると，第 1 コミュニティでは人工知能に関する倫理や思想，定義に関するページが中心となり，また第 2 コミュニティでは人工知能の研究に関するページが中心となった．また，表 1,

表 1　トピック “AI”：第 1 コミュニティの authority
Table 1　Topic “AI”: authorities of 1st community.

URL
http://psyche.cs.monash.edu.au/
http://www.kosara.net/thoughts/ai.html
http://dmoz.org/about.html
http://artsci.wustl.edu/~philos/MindDict/index.html
http://www.inm.de/kip/

表 2　トピック “AI”：第 2 コミュニティの authority
Table 2　Topic “AI”: authorities of 2nd community.

URL
http://www.aaai.org/
http://www.cs.washington.edu/research/jair/home.html
http://www.cs.berkeley.edu/~russell/ai.html
http://www.loebner.net/Prizef/loebner-prize.html
http://www.generation5.org/

表 2 に示した authority 間のリンク関係は疎であり，これらの authority から構成されたコミュニティは Kleinberg のモデルに適合している．

b）トピック “ruby”

トピック “ruby” を設定して得られた第 1 コミュニティを構成するページのうち，authority の重みの絶対値の大きいものから上位 5 件の URL を表 3 に示す．これらの authority の内容を確認した所，全て O’REILLY 社が管理しているページであった．さらに，トピック “ruby” に関連する情報を含んでおらず，適切な authority ではなかった．また，表 4 に示す第 2 コミュニティの authority 上位 5 件に関しても，第 2 位と第 3 位の authority 以外は O’REILLY 社が管理するページであり，第 1 位の authority として抽出されたページにはプログラミング言語 **Ruby** に関する記述が存在したが，それ以外の authority には “ruby” に関する情報は存在しなかった．このように，検索トピック “ruby” に関連の無いコミュニティが複数上位に抽出されたため，検索トピックに関して有益な情報を含むコミュニティを効率的に発見できていない．ここで，検索トピックに関連が無いにも関わらず，重要なコミュニティとして抽出されたページ集合のリンク構造を明らかにするため，表 3 に示された authority と対応する hub の各上位 3 件のリンク関係を図 4 に示す．図 4 から，この O’REILLY 社が管理するページで構成されたコミュニティに含まれる authority 同士，hub 同士には密なリンク関係が存在し，文献 [4], [5] で指摘された問題点，つまり HITS アルゴリズムは互いに密にリンクを張っているページ集合を重要な authority，hub として抽出するという問題点を確認できる．

トピック “ruby” に関する適切なコミュニティは第 3 コミュニティとして現れたので，表 5 に上位 5 件の authority の URL を示す．表に示された各 authority はプログラミング言語 **Ruby** に関するページであった．特に，第 1 位の authority(http://www.ruby-lang.org/en/) はプログラミング言語 **Ruby**

表 3　トピック “ruby”：第 1 コミュニティの authority
Table 3　Topic “ruby”: authorities of 1st community.

URI
http://safari.oreilly.com/
http://www.xml.com/
http://webservices.xml.com/
http://digitalmedia.oreilly.com/
http://www.onjava.com/

表 4　トピック “ruby”：第 2 コミュニティの authority
Table 4　Topic “ruby”: authorities of 2nd community.

URI
http://www.codezoo.com/
http://servlets.com/
http://www.linuxquestions.org/
http://www.osdir.com/
http://www.openp2p.com/

表 5　トピック “ruby”:第 3 コミュニティの authority
Table 5　Topic “ruby”:authorities of 3rd community.

URL
http://www.ruby-lang.org/en/
http://www.tmtm.org/en/mysql/ruby/
http://www.ruby-lang.org/
http://jruby.sourceforge.net/
http://www2s.biglobe.ne.jp/~Nori/ruby/

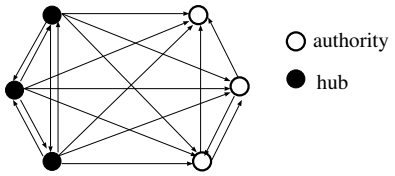

図 4　O’REILLY 社のコミュニティのリンク構造
Fig. 4　Link structure of the O’REILLY’s community.

のソースコードを配布しているページである．これらの authority 間のリンク関係は，O’REILLY 社が管理するページで構成されたコミュニティよりも疎であり，より Kleinberg のモデルに近いコミュニティとなる事が確認された．

2.3　クラスタ係数を用いた Web コミュニティの評価

2. 2 節で述べたように HITS アルゴリズム用いると，Kleinberg のモデルに反して，企業等が作成した互いに密にリンクを張っているページ集合が，重要なコミュニティとして抽出されてしまう．そこで，抽出されたコミュニティを構成する authority 間のリンク関係と Kleinberg のモデルとの適合度を表す指標として，クラスタ係数 (clustering coefficient) [6] を導入する．

クラスタ係数は，無向グラフにおいてネットワークの集まり具合を表すために用いられる指標である．図 5 のようなネットワークを考えると，ノード i のクラスタ係数 c_i は下式で定義される．

$$c_i = \frac{2E_i}{|e_i|(|e_i|-1)}. \tag{3}$$

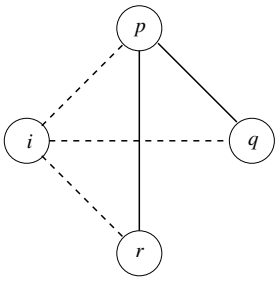

図 5　無向グラフの例 ($c_i = 0.66$)
Fig. 5　Example of an undirecited graph.

ここで $|e_i|$ はノード i と隣接関係にあるノード (図 5 における ノード p，q，r) の数を表し，E_i はこれら $|e_i|$ 個のノード間に 存在する辺 (図 5 における実線) の数を表す．式 (3) では，実際 に存在する辺の数 E_i を，$|e_i|$ 個のノード間で考えられる辺の 総数で割っている．このため，ノード i のクラスタ係数 c_i は，ノード i と隣接関係にある任意の二つのノードを選んだ場合に，これらのノード間に辺が存在する確率となる．このように，クラスタ係数は本来，無向グラフに対して定義されるものであるが，一方，本文で対象とする Web は有向グラフである．つまり，ページ i からページ j へのリンクとページ j からページ i へのリンクとは異なる有向辺とみなす．したがって，コミュニティ評価に用いるクラスタ係数は式 (3) に代えて，下式で定義されるものを用いる必要がある．

$$c_i = \frac{E_i}{o_i(o_i - 1)}. \tag{4}$$

ここで o_i はページ i のリンク先のページ (以降，out-link ページと呼ぶ) 数であり，E_i はページ i の out-link ページ間に存在するリンクの数を表す．ただし，$o_i \leqq 1$ の場合にはクラスタ係数を算出できないため，$c_i = 0$ とする．Kleinberg の定義によれば，hub の out-link ページは authority であり，良質なコミュニティでは，これらの authority は互いにリンクを張らない．すなわち，hub に対してクラスタ係数を算出すれば，それは小さな値になるはずである．

次に，上で求められた hub に対するクラスタ係数を用いてコミュニティ全体に対するクラスタ係数を算出する．HITS アルゴリズムの手順 (4) では，各ページの hub の重みを成分としてもつ正規化されたベクトル $\boldsymbol{h}$ を求めている事から，$\sum |h_i|^2 = 1$ となる事を利用し，$|h_i|^2$ をコミュニティのクラスタ係数算出における重みとして利用する．つまりコミュニティのクラスタ係数を次式で定義する．

$$\sum_{i=1}^{n} c_i |h_i|^2 = \boldsymbol{h}^T C \boldsymbol{h}. \tag{5}$$

ここで，行列 C は base 集合に含まれるページのクラスタ係数を対角要素にもつ対角行列を表し，$C = \mathrm{diag}(c_1, \ldots, c_n)$ である．式 (5) で定義されるコミュニティのクラスタ係数は，重要な hub として抽出されたページが平均してどの程度のクラスタ係数をもつかを表す．したがって，Kleinberg のモデルに適合するコミュニティに対しては，コミュニティのクラスタ係数は小さくなる．

表 6　コミュニティのクラスタ係数
Table 6　Clustering coefficient of each communities.

#	“AI”	“ruby”
1	0.0150	0.6938
2	0.0206	0.6030
3	0.3624	0.0094

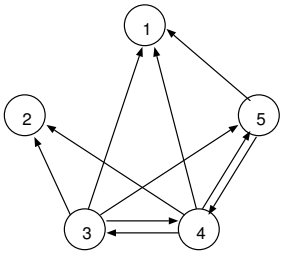

図 6　Jupitermedia コミュニティの authority 間のリンク構造
Fig. 6　Link structure of the Jupitermedia’s authorities

ここで，O’REILLY 社のコミュニティのように，企業等が管理する互いに密にリンクを張っているページ集合からなるコミュニティに対しては，式 (5) で定義したコミュニティのクラスタ係数が高い値を示す事を確認する．2.2 節で抽出したコミュニティのクラスタ係数を，トピック毎に第 3 コミュニティまで表 6 に示す．トピック “AI” の第 3 コミュニティは高いクラスタ係数をもつが，このコミュニティの authority 上位 5 件は全て Jupitermedia という企業が管理するページであり，これらの authority 間でのリンク関係は密であった (図 6)．また，トピック “ruby” では第 1，第 2 コミュニティ，つまり O’REILLY 社により管理されているページからなるコミュニティのクラスタ係数が高く，前述したように Kleinberg のモデルに適合していない．このように，コミュニティのクラスタ係数を用いて，抽出されたコミュニティと Kleinberg のモデルとの適合度を測る事が可能である事が確認された．

2.4　コミュニティ抽出過程の分析

本節では，高いクラスタ係数をもつコミュニティが上位に抽出される原因を，authority を決定する行列 $L^T L$ の固有値と固有ベクトルの算出の観点から考察する．式 (1) と式 (2) を用いて第 k コミュニティの authority を求める，すなわち，行列 $L^T L$ の k 番目に大きな固有値 λ_k に対応する固有ベクトル $\boldsymbol{a}_k$ を求めるということは，$\boldsymbol{a}_k^T L^T L \boldsymbol{a}_k$ を最大にする $\boldsymbol{a}_k$ を，以下の条件 1，条件 2 の下で求めることに等しい [8].

条件 1　$||\boldsymbol{a}_k|| = 1$

条件 2　$\boldsymbol{a}_k^T \boldsymbol{a}_{k-1} = \boldsymbol{a}_k^T \boldsymbol{a}_{k-2} = \cdots = \boldsymbol{a}_k^T \boldsymbol{a}_1 = 0$

さらに，このときの $\boldsymbol{a}_k^T L^T L \boldsymbol{a}_k$ の最大値は固有値 λ_k と一致する．ここで，ベクトル $\boldsymbol{a}_k$ の i 番目の成分を $a_i^{(k)}$ とし，隣接行列 L の i 番目の列を列ベクトル $\boldsymbol{l}_i = (l_{1i} \cdots l_{ni})^T$ で表すと，$\boldsymbol{a}_k^T L^T L \boldsymbol{a}_k$ は，

$$\sum_i \sum_j a_i^{(k)} \boldsymbol{l}_i^T \boldsymbol{l}_j a_j^{(k)} \tag{6}$$

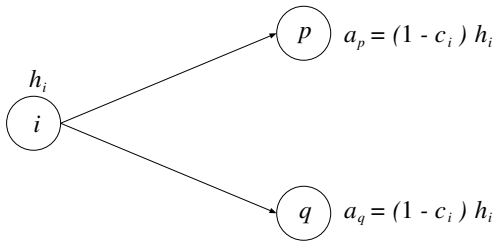

図 7　提案法における authority の重みの更新
Fig. 7　Updating authority weight using the proposed method.

と表される．さらに，式 (6) は以下のように書き改められる．

$$\sum_i |a_i^{(k)}|^2 b_i + \sum_i \sum_j a_i^{(k)} s_{ij} a_j^{(k)}. \qquad (7)$$

ここで，b_i はページ i の入次数，つまりページ i にリンクを張っているページの数を表し，s_{ij} はページ i，ページ j の両方にリンクを張っているページの数とする．ただし，$s_{ii} = 0$ とする．ここで，式 (7) の値を最大にするためには，まず多くのページからリンクを張られているページの authority の重みについては絶対値が大きな値を与える．さらに，ページ i とページ j の両方にリンクを張っているページの数が多い場合には，絶対値が大きく，かつ同符号の authority の重みをページ i とページ j に与える．これらから，HITS アルゴリズムを適用した場合には，b_i が大きなページと s_{ij} が大きなページが authority として抽出されやすくなる．したがって，図 4 で表されるような互いに密にリンクを張っているページ集合は，各ページが高い authority の重みを得やすくなる．

3. クラスタ係数を用いた HITS アルゴリズムの変更

本章では，各ページのクラスタ係数を利用して HITS アルゴリズム (以降，従来法と呼ぶ) における authority の重みの更新を変更し，Kleinberg のモデルに適合しない図 4 のようなコミュニティの出現を抑制する手法を提案する．Kleinberg のモデルに適合するコミュニティでは，authority 間のリンク関係は疎であるため，hub のクラスタ係数は小さくなる．このような hub が抽出されるように，各ページのクラスタ係数 c_i を対角要素にもつ対角行列 C を用いて，式 (1) の更新処理を

$$\boldsymbol{a} = \gamma L^T (I - C)\boldsymbol{h} \quad (\gamma : \text{正規化定数}) \qquad (8)$$

と変更する．ここで，I は単位行列を表す．式 (8) では，図 7 で表されるように，authority の重みの更新において，ページ i の hub の重み h_i をそのページのクラスタ係数に応じて減じ $((1-c_i)h_i)$，これを伝播させる．これにより，高いクラスタ係数をもつページからリンクを張られているページの authority の重みが抑えられ，互いに密にリンクを張っているページ集合に属するページが authority として抽出される事を抑制する．その結果として，クラスタ係数が低く，Kleinberg のモデルに適合するコミュニティの抽出が期待できる．なお，式 (8) に式 (2) を代入すれば明らかなように，上で提案した手法では行列 $L^T L - L^T C L$ の最大固有値に対応する固有ベクトルとして $\boldsymbol{a}$ が求まる．以降，式 (8) を用いるコミュニティ抽出手法を提案法と呼ぶ．

表 7　トピック “AI”:提案法により抽出された第 1 コミュニティの authority
Table 7　Topic “AI”: authority of 1st community.

http://psyche.cs.monash.edu.au/
http://www.kosara.net/thoughts/ai.html
http://artsci.wustl.edu/~philos/MindDict/index.html
http://inm.de/kip/
http://dmoz.org/about.html

4. 実　　験

本章では，提案法を用いたコミュニティ抽出結果と 2.2 節で示した従来法の結果を比較し，提案法の有効性を確認する．また，2.4 節と同様に authority を決定する行列 $L^T L - L^T C L$ の固有値と固有ベクトルの算出の観点から，提案法を用いて得られるコミュニティ抽出結果について考察する．

4.1　提案法を用いたコミュニティ抽出結果

本節では，提案法を用いたコミュニティ抽出結果を示す．検索トピックとして，2.2 節と同様に “ruby”，“AI” を設定した．また，root 集合と base 集合の作成にも 2.2 節と同様の設定でコミュニティ抽出を行った．以下で，個々の検索トピック毎に，提案法を用いたコミュニティ抽出結果の詳細を説明し，従来法との比較を行う．

a)　トピック “AI”

提案法を用いて得られた第 1 コミュニティを構成するページのうち，authority の重みの絶対値の大きいものから上位 5 件の URL を表 7 に示す．従来法で得られた第 3 位の authority(`http://dmoz.org/about.html`) が，提案法では順位を下げ，第 5 位の authority として抽出された．このように，authority の順位に変動はあったが，第 1，第 2 コミュニティに関しては従来法を用いた場合とほぼ同様のコミュニティ抽出結果が得られた．大きな変化が確認されたのは第 3 コミュニティであり，このコミュニティを構成するページのうち authority の重みの大きいものから上位 5 件の URL を表 8 に示す．従来法では Jupitermedia 社が管理するページからなるコミュニティが抽出されたが，提案法では人工知能に関するページからなるコミュニティが抽出された．式 (5) を用いてコミュニティのクラスタ係数を算出した結果，提案法による第 3 コミュニティのクラスタ係数は 0.0231 となった．また，表 8 に示された authority 間のリンク関係を図 8 (注1)に示す．図 6 と比較し，提案法により抽出された第 3 コミュニティの authority 間のリンク関係は疎である．これは，提案法を用いる事でクラスタ係数の高いコミュニティの出現，つまり authority 間のリンク関係が密であるコミュニティの出現が抑制された事を表している．

(注1)：実験に用いた隣接行列においては，これらの authority 間にはリンク関係は存在しなかったが，実際に各 authority を調べたところ図 8 中の破線矢印で表したリンクが存在した．

表 8　トピック “AI”:提案法により抽出された第 3 コミュニティの authority
Table 8　Topic “AI”: authorities of 3rd community.

URL
http://www.generation5.org/
http://www.cs.berkeley.edu/~russell/ai.html
http://www.tau.ac.il/humanities/philos/ai/links.html
http://ai-depot.com/
http://liinwww.ira.uka.de/bibliography/Ai/

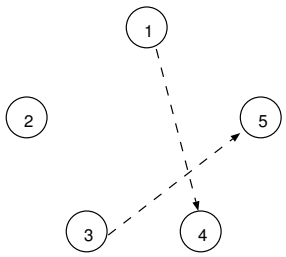

図 8　提案法を用いた第 3 コミュニティの authority 間のリンク関係
Fig. 8　Link structures of authorities using the proposed method

b)　トピック “ruby”

提案法により抽出されたトピック “ruby” の第 1 コミュニティに関しては，authority の順位は変動していたが，従来法を用いて得られた第 1 コミュニティと同じく，O’REILLY 社が管理するページからなるコミュニティであった．ここで，比較のために従来法，提案法を用いて得られた第 1 コミュニティを構成するページのうち，authority の重みが大きいものから上位 5 件のページ間のリンク関係をそれぞれ図 9，図 10 に示す．図より，提案法を用いて得られた authority 間のリンク関係が，従来法よりも密となっている．この原因としては，O’REILLY 社の管理するページは，O’REILLY 社以外の多数のページからリンクを張られていることが挙げられる．これにより，O’REILLY 社が管理するページ集合内におけるクラスタ係数に基づく重みの伝播の抑制量よりも，他のページからの重みの流入量が大きくなった結果，図 10 に示されるリンク関係をもつ authority が抽出されたと考えられる．

また，第 2 コミュニティを構成するページのうち，authority の重みが大きいものから上位 5 件の URL を表 9 に示す．従来法では O’REILLY 社が管理するページからなるコミュニティであった第 2 コミュニティは，提案法ではプログラミング言語 **Ruby** に関するページ，つまり従来法で第 3 コミュニティの authority として抽出されたページで構成されていた．ここで，各コミュニティのクラスタ係数を表 10 に示す．表から，トピック “ruby” においても “AI” と同様に，提案法によりクラスタ係数の高いコミュニティの出現が抑制されている事が確認された．また，表 9 に示した authority 間のリンク関係を図 11 に示す．第 1 位と第 2 位の authority 間には，実際にはリンクが存在するが，隣接行列作成の際に同一ドメイン間のリンクは削除しているため，図 11 においては破線矢印でリンクを表している．

表 9　トピック “ruby”:提案法により抽出された第 2 コミュニティの authority
Table 9　Topic “ruby”: authorities of 2nd community.

URL
http://www.ruby-lang.org/en/
http://www.tmtm.org/en/mysql/ruby/
http://jruby.sourceforge.net/
http://www2s.biglobe.ne.jp/~Nori/ruby/
http://www.ruby-lang.org/

表 10　提案法を用いて抽出されたコミュニティのクラスタ係数
Table 10　Clustering coefficient of each communities.

#	“AI”	“ruby”
1	0.0144	0.6844
2	0.0186	0.0079
3	0.0231	0.0213

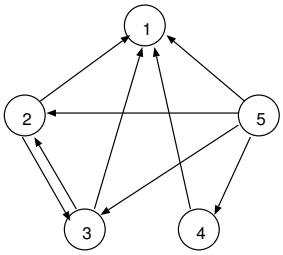

図 9　HITS を用いた第 1 コミュニティの authority 間のリンク関係
Fig. 9　Link structures of authorities using HITS

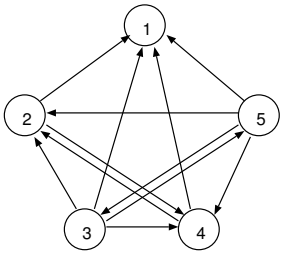

図 10　提案法を用いた第 1 コミュニティの authority 間のリンク関係
Fig. 10　Link structures of authorities using the proposed method

図 11 から，authority 間のリンク関係が疎であることが確認できる．したがって，提案法を用いた場合には，リンク関係が疎である authority からなるコミュニティを従来法よりも上位に抽出可能であるといえる．一方で，提案法を適用した場合においても，従来法と同様に O’REILLY 社が管理するページからなるコミュニティが重要なコミュニティとして抽出された．次節では，この原因について考察する．

4.2　提案法によるコミュニティ抽出過程の分析

2.4 節と同様に authority を決定する行列の固有値と固有ベクトルの算出の観点から説明する．式 (8) を用いて第 k コミュニティの authority を求める，すなわち，行列 $L^TL - L^TCL$ の k 番目に大きな固有値 λ_k に対応する固有ベクトル $\boldsymbol{a}_k$ を求めることは，2.4 節で述べた条件 1，条件 2 を満たした上で下式を最大にする $\boldsymbol{a}_k$ を求めることに等しい．

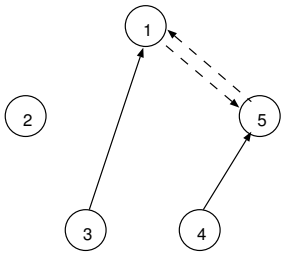

図 11　提案法を用いた第 2 コミュニティの authority 間のリンク関係
Fig. 11　Link structures of authorities using the proposed method

表 11　提案法を用いた場合の固有値
Table 11　Eigen values using the proposed method

	“AI”	“ruby”
λ_1	575	1890
λ_2	522	884
λ_3	369	732

$$\boldsymbol{a}_k^T(L^TL - L^TCL)\boldsymbol{a}_k. \tag{9}$$

さらに，このときの最大値は固有値 λ_k と一致する．また，ここで式 (2) を上式に代入すると次式が得られる．

$$||L\boldsymbol{a}_k||^2(1 - \boldsymbol{h}_k^TC\boldsymbol{h}_k). \tag{10}$$

ここで，上式中の $\boldsymbol{h}_k^TC\boldsymbol{h}_k$ は，式 (5) で定義したコミュニティのクラスタ係数を表している．つまり，コミュニティのクラスタ係数が高くなるほど，式 (10) の値は小さくなる．これにより，提案法を用いた場合，トピック “AI” では Jupitermedia 社が管理するページで構成されたコミュニティが第 3 コミュニティとして抽出されず，またトピック “ruby” では O’REILLY 社が管理するページで構成されたコミュニティが第 2 コミュニティとして抽出されない結果となった．ここで，表 11 に authority を決定する行列 $L^TL - L^TCL$ の固有値のうち，3 番目に大きい固有値までを示す．トピック “ruby” の λ_1，λ_2 のように，authority を決定する行列の固有値間の差が大きくなる場合には，提案法の効果が十分に現れなくなる．その結果，トピック “ruby” では第 1，第 2 コミュニティの順位が入れ替わること無く，O’REILLY 社が管理するページで構成されたコミュニティが従来法を用いたコミュニティ抽出と同様に第 1 コミュニティとして抽出された．これは提案法の問題点であり，新たな改善策が必要となる．

5. ま と め

本文では，HITS アルゴリズムを用いて抽出された Web コミュニティと，Kleinberg のモデルとの適合度を表す評価値としてクラスタ係数を導入した．その結果，Kleinberg のモデルに適合せず，互いに密にリンクを張っているページ集合からなるコミュニティを，クラスタ係数により判別できる事を確認した．また，この HITS アルゴリズムによって，互いに密にリンクを張っている Web ページ集合が重要な Web コミュニティとして抽出される原因について，authority を算出する行列の固有値と固有ベクトルの算出過程に着目する事で考察を行った．さらに，クラスタ係数を用いて HITS アルゴリズムにおける authority の重みの更新処理を変更し，Kleinberg のモデルに適合しないコミュニティの出現を抑制する手法を提案した．最後に，提案法を用いたコミュニティ抽出結果と，HITS アルゴリズムによる抽出結果を比較し，提案法の有効性を確認した．しかしながら，提案法においても HITS アルゴリズムにおける問題点の改善が十分でない事が確認された．したがって，今後は提案法を用いて発生する問題の原因について詳細な検討を行う．さらに，文献 [9] で述べられているリンク解析手法 out-link normalized rank 等を用いて，提案法の改良を検討する予定である．

文　献

[1] R. Kosala and H. Blockeel: “Web mining research: A survey”, SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, **2**, (2000).
[2] J. M. Kleinberg: “Authoritative sources in a hyperlinked environment”, Journal of the ACM, **46**, 5, pp. 604–632 (1999).
[3] D. Gibson, J. Kleinberg and P. Raghavan: “Inferring web communities from link topology”, Proc. 9th ACM Conference on Hypertext and Hypermedia (HyperText 98), pp. 225–234 (1998).
[4] A. Borodin, G. O. Roberts, J. S. Rosenthal and P. Tsaparas: “Finding authorities and hubs from link structures on the world wide web”, World Wide Web, pp. 415–429 (2001).
[5] K. Bharat and M. R. Henzinger: “Improved algorithms for topic distillation in a hyperlinked environment”, Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, Melbourne, AU, pp. 104–111 (1998).
[6] D. J. Watts and S. H. Strogatz: “Collective dynamics of ‘small-world’ networks”, Nature, **393**, pp. 440–442 (1998).
[7] AltaVista: `http://www.altavista.com/`.
[8] 津田, 仙田, 美濃, 池田：“共起行列の固有ベクトルを用いる単語クラスタリング法〜文書データベースの概要を表す単語クラスタの抽出〜”, 情処研報, NL94–103, pp. 41–48 (1994).
[9] C. Ding, X. He, P. Husbands, H. Zha and H. Simon: “PageRank, HITS and a unified framework for link analysis”, Technical Report 49372, LBNL (2002).