HITS 手法を応用したイントラネットにおけるページランキング

森谷 浩貴† 岡部 正幸†† 梅村 恭司‡

†‡豊橋技術科学大学 情報工学系 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

††豊橋技術科学大学 情報メディア基盤センター 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: †moriya@ss.cs.tut.ac.jp, ††okabe@imc.tut.ac.jp, ‡umemura@tut.jp

あらまし ある企業内検索システムで利用する，リンク解析モデルを用いたランキング手法を提案する．リンク解析モデルの代表的なものとして，google で利用される PageRank や Ask.com で利用される HITS 手法がある．しかし，これらの手法は Web 全体を対象としたものであり，リンク構造が異なるイントラネットに向いていないと考えられる．そこで，それらの構造の違いを考慮するため，類似した複数のイントラネットを利用した重みつき HITS 手法によるランキング手法を提案する．提案手法では，それらのイントラネットに共通して出現するページは重要であるとし，権威得点を計算する際に重み付けを行うことで，通常の HITS 手法では重要でも低い得点になりがちなぶら下がりノードのランキングを上昇させることが可能であることを示す．

キーワード HITS 手法，イントラネット，ページランキング，リンク解析モデル

Modified HITS for Page Ranking on intranet

Hirotaka MORIYA† Masayuki OKABE†† Kyoji UMEMURA‡

†‡Department of Information and Computer Sciences, Toyohashi University of Technology 1-1 Hibarigaoka, Tenpaku-cho,Toyohashi, Aichi, 441-8580 Japan

††Information and Media Center, Toyohashi University of Technology 1-1 Hibarigaoka, Tenpaku-cho,Toyohashi, Aichi, 441-8580 Japan

E-mail: †moriya@ss.cs.tut.ac.jp, ††okabe@imc.tut.ac.jp, ‡umemura@tut.jp

Abstract This research proposes the ranking technique that uses the link analysis model used by search engine in certain enterprise. There are HITS used by Ask.com and PageRank used by Google, which are typical link analysis models. However, these techniques are not suitable for Intranet where link structure is different because these techniques are the one intended for the entire Web. Therefore, we propose a ranking technique at HITS with weight using two or more similar Intranet to consider the difference of those structures. In the proposal technique, it is shown that the ranking of the hanging node that tends to become a low score even if it is important is raised by assuming that a similar page that appears in those Intranet is important and weighting when the authority score is calculated.

Keyword HITS, intranet, pageranking, link analysis model

1. はじめに

本研究室はある企業内検索システムの開発にかかわっている．現在は，検索結果を表示する際に，検索単語と抽出されたキーワードの類似度をスコアとし，表示しているが，Web のリンク解析モデルは利用できない．

そこで， はじめに「Google の PageRank の数理」[1] を参考に，ページの重要性の取得に関する調査を行った．内容は Google で利用される PageRank の他に Ask.com で利用される HITS 手法や SALSA などのリンク解析モデルについて調査を行った．

その結果，PageRank や HITS 手法はハイパーリンク構造が網目状に張り巡らされている Web 全体を対象としてつくられており，そのため，リンク構造が異なる社内 Web（イントラネット）では適用できないという問題が明らかになった．

そこで，それを解決するために重みつき HITS 手法を提案し，その振る舞いを確認することが本研究の目的となる．

2. HITS 手法

2.1. 原理

PageRank は各ページに対して人気得点を 1 つ作成するが, HITS 手法は各ページに対して入リンクと出リンクとの両方を使って，2 つの人気得点を作成する．さらに，クエリー従属であることも PageRank との大きな違いである．HITS 手法は，Web ページを権威とハブとして考える．多くの入リンクを持っていると権威と呼ばれ，多くの出リンクを持っているとハブと呼ばれる．権威とハブは次の巡回的な主張が成り立つとき良いといわれる．つまり,「良い権威たちは良いハブたちによって指されており，良いハブたちは良い権威たちを指している」. そして，各ページは権威としての指標とハブとしての指標を持つ．

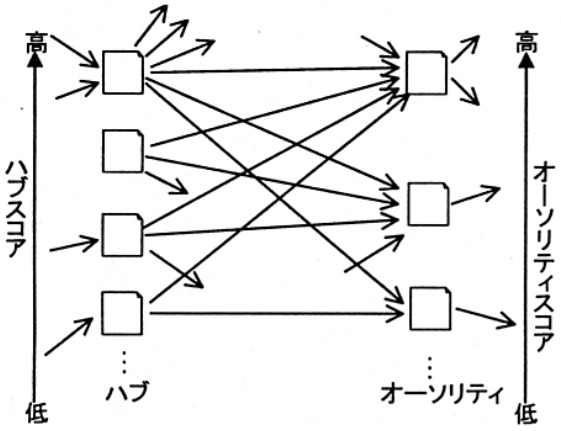

図 1　ハブと権威（オーソリティ）([2]より引用)

HITS 手法の実装には 2 つの主なステップがある. 最初は，検索後に関連する近傍グラフ N を作ることである．2 番目は，N の各ページに対して権威得点とハブ得点（x と y）を計算し，もっとも権威のあるページともっとも“ハブっぽい”ページをユーザーに提示することである．

次に，HITS が行っていることを式で示す．

$$x_i^{(k)} = \sum_{j:e_{ji}\in E} y_j^{(k-1)} \qquad y_i^{(k)} = \sum_{j:e_{ij}\in E} x_j^{(k)} \qquad (2.1)$$

x_i　:ページiの権威得点
y_i　:ページiのハブ得点
E　:Webグラフのすべての有向辺の集合
e_{ij}　:ノードiからノードjへの有向辺

式（2.1）は，Web の有向グラフの隣接行列 L の助けを借りて行列の形に書くことができる．

$$x^{(k)} = L^T y^{(k-1)} \qquad y^{(k)} = Lx^{(k)} \qquad (2.2)$$

L　:隣接行列

さらに，式（2.2）は次のように簡単にできる．

$$x^{(k)} = L^T Lx^{(k-1)} \qquad y^{(k)} = LL^T y^{(k-1)} \qquad (2.3)$$

L^TL　:権威行列
LL^T　:ハブ行列

式(2.3)は行列 L^TL と LL^T の主固有ベクトルを計算するための繰返しベキ乗法を定義している．

しかし，これには権威とハブとのベクトルが一意に収束しないという問題がある．つまり，権威（そしてハブ）ベクトルの極限が，初期ベクトルの選択を変えると，異なる値になることを意味する．この収束が唯一ではないという問題の本質は可約性にある．

行列が可約であるとは，いくつかの状態があり，その状態に入ることは可能だがいったん入るとそこから出るのは不可能であることを意味する．既約であるとは，すべての状態が他のすべての状態から到達可能であることを意味する．ペロン - フロベニウスの定理によると，既約非負行列がペロンベクトルと呼ばれる唯一の正規化された正の主固有ベクトルを持つことが保障されている．よって, HITS アルゴリズムが一意でない解に収束するのは行列の可約性によるものである．

任意の正方行列が既約である必要十分条件はその有向グラフが強連結であることである．強連結とは，どの 2 つのノード（N_i, N_j）に対しても から に向かう道が存在するということである．また，この条件は A が任意の置換行列 P に対して式(2.4)を満たすことと等価である．

$$P^T AP \neq \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix} \qquad (2.4)$$

P^TAP　:Aの対称置換
X,Y　:正方行列

よって, HITS アルゴリズムを既約な状態にするために次式で示すような「原始性調整」を行う．

$$\zeta L^T L + (1-\xi)/nee^T \qquad (2.5)$$

$$\zeta LL^T + (1-\xi)/nee^T \qquad (2.6)$$

ξ　:0と1の間のスケーリングパラメータ
e　:すべての要素が1の列ベクトル

式(2.5)を既約行列に修正した権威行列 L^TL，式(2.6)をハブ行列 LL^T として扱う．

これらの修正によって，2 つの行列は既約であり，ペロン - フロベニウスの定理によって，どちらも正規化された正の主固有ベクトルを持つことになる．さらに，このベキ乗法は有限回のステップで収束することが保障されている．

2.2. 定義

良い権威たちは良いハブたちによって指されており，良いハブたちは良い権威たちを指している．

3. イントラネットのリンク構造における HITS 手法の欠点

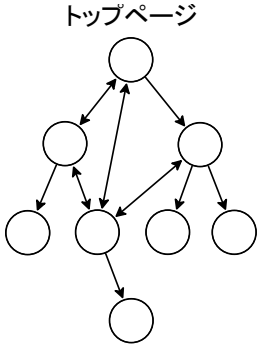

図 2　イントラネットにおけるリンク構造

図 2 に示すように，イントラネットは top ページを最上位にツリー構造のようになりがちであり，そのため top ページへリンク数が増え得点が集中し，必然的に上層部が高い得点になってしまう．つまり，イントラネットでは下層部(ぶら下がりノード)に PDF や jpeg などの重要なページがあったとしても，そのページの得点が低くなってしまう傾向がある．

よって，下層部でも重要なページは上位にランキングされるようなイントラネットの構造を考慮した新しいランキング手法が必要である．

4. 重みつき HITS 手法

イントラネットで望ましい手法を用いた場合，top は出リンクが多くなるためハブの性質が高くなり，リーフとなるぶら下がりノードは入りリンクのみとなるためオーソリティの性質が高くなるのが妥当である．

そのため，リーフであるぶら下がりノードはハブにはなりえないため，入りリンクに重みをつけ，オーソリティスコアを求めることが妥当であると考えられる．

よって，本研究では重要なぶら下がりノードのランキングを上げるため，隣接行列 L を上記のような重みつき隣接行列とすることを提案する．

重みは次のようなものからとることができる．

- ページ内部のリンク場所
- アンカーテキストからの類推
- リンク先のページの内容

そして，これらの方法によりリンクに重みつけを行った隣接行列を L_S とし式(4.1)を定義する．

L_S の値は，クローリングの結果より決定されるべきものであるが，現時点ではある方法で求まるものとしている．

$$x^{(k)} = L_S^T L_S x^{(k-1)} \quad y^{(k)} = L_S L_S^T y^{(k-1)} \qquad (4.1)$$

$L_n^T L_S$　:既約行列に修正された重みつき権威行列
$L_S L_n^T$　:既約行列に修正された重みつきハブ行列

尚，式(4.1)における重みつき権威行列とハブ行列はこれまでの HITS 手法同様に既約行列に修正されている．

5. 実験と考察

本実験では図 3 の test1, test2, test3 を類似したイントラネットとし，これらのイントラネットを用いて重みなし HITS 手法と重みつき HITS 手法の振る舞いの変化の確認を行った．図 3 における，a～m の英字はそのページのアドレスを示し，矢印はそのページがハイパーリンクしている先を示している．また，リンクに付けられるラベルあ～けは，Web ページの内容が推定されたリンクの種別であり，その重みが表 1 のように決まっているとした．尚，重みを計算するために 3 つのイントラネットを使用したが，実際にランキングを行うのは test1 のイントラネットとする．

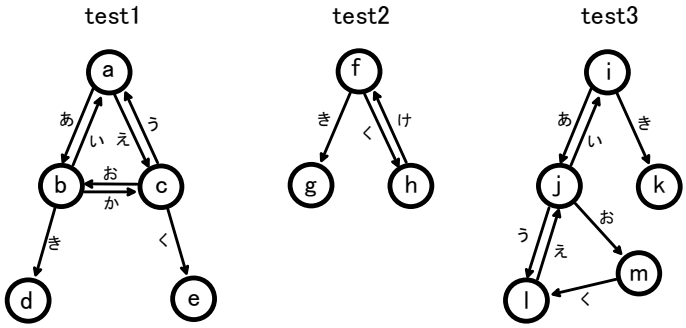

図 3　類似したイントラネットの例

ラベル = か	score = 1.5
ラベル = け	score = 1.5
ラベル = い	score = 5.0
ラベル = え	score = 5.0
ラベル = あ	score = 5.0
ラベル = お	score = 5.0
ラベル = う	score = 5.0
ラベル = き	score = 8.5
ラベル = く	score = 8.5

表 1　リンクの重み

5.1. オーソリティスコアの比較

まず，オーソリティスコアの比較を行う．図 4 には重みなし HITS 手法による結果，図 5 には重みつき HITS 手法による結果を示す．

尚，HITS の計算回数は 15 回，スケーリングパラメータを 0.95 とて計算を行った．

```
url = a scor = 0.26727
url = b scor = 0.23163
url = c scor = 0.23163
url = d scor = 0.13473
url = e scor = 0.13473
```

図 4　重みなし HITS 手法（オーソリティスコア）

```
url = d scor = 0.27762
url = a scor = 0.26139
url = c scor = 0.21404
url = e scor = 0.16676
url = b scor = 0.08019
```

図 5　重みつき HITS 手法（オーソリティスコア）

結果，重みつけを行った場合，重みなし HITS 手法と比べて，ぶら下がりノードである（d, e）のランキングが多少上昇していることが確認できた．これにより，重みつき HITS によってぶら下がりノードに対する重要度を制御出来ることを確認した．

5.2. ハブスコアの比較

次に，ハブスコアの比較を行う．図 6 には重みなし HITS 手法による結果，図 7 には重みつき HITS 手法による結果を示す．また，5.1.同様に HITS の計算回数は 15 回，スケーリングパラメータを 0.95 とした．

```
url = b scor = 0.36411
url = c scor = 0.36411
url = a scor = 0.26736
url = d scor = 0.00221
url = e scor = 0.00221
```

図 6　重みなし HITS 手法（ハブスコア）

```
url = b scor = 0.52317
url = c scor = 0.31418
url = a scor = 0.16251
url = d scor = 0.00001
url = e scor = 0.00001
```

図 7　重みつき HITS 手法（ハブスコア）

結果，重みつけを行った方がノード(d, e)のスコアが低くなってしまった．ハブスコアを得るためにはそのノードから出リンクが出ていなければならないため，出リンクが存在しないノード(d, e)のスコアが一番低いスコアとなることは明白である．また，ノード(a)のハブスコアも大幅に下がっていることが分かる．これは，ノード(b, c)の方がよりスコアの高い出リンク（き，く）を多く持っていたからだと考えられる．

この実験結果から，ハブスコアの計算ではぶら下がりノードをすくい上げることができず，重みつき HITS 手法は権威スコアを計算する際にのみ有効であることがわかる．これは，イントラネットにおける top のハブ性質と，リーフのオーソリティ性質により，ぶら下がりノードはハブになりえないため所望の結果といえる．

6. 今後の課題

今回は HITS 手法の重みの有無による振る舞いの変化を確認するために，単純で小さな仮想イントラネットを使用した．そして今回の実験により，重みつき HITS 手法のぶら下がりノードに対する振る舞いが確認できたので，実際のイントラネットのクローリングを行い，それに対する重みの方法を決め，ぶら下がりノードに対しても重要性が判定出来る重みつき HITS 手法の実験を行いたいと考えている．

参　考　文　献

[1] Amy N.Langville , Carl D.Meyer , “Goolge PageRank の数理”, 共立出版.

[2] 特許庁標準技術集「リンク構造解析」, http://www.jpo.go.jp/shiryou/s_sonota/hyoujun_gijutsu/search_engine/b/b54.htm