1E2-01

多種ノードネットワークのランキング

Ranking Multi-type Nodes in Networks

藤井　通太[*1]　村田　剛志[*2]
Michita Fujii　Tsuyoshi Murata

[*1]東京工業大学 大学院理工学研究科 集積システム専攻
Department of Communications and Integrated Systems, Graduate School of Engineering, Tokyo Institute of Technology

[*2]東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Many methods have been proposed for ranking nodes in networks that are composed of single-type nodes. In the real world, however, there are many networks composed of multi-type nodes, such as the users, resources, and tags. In this research, we propose new methods for the decision of propagation coefficient with unified link analysis and perform experiments of ranking nodes in a multi-node SNS network. We evaluate ranking result by several standards and show that our proposed method outperforms traditional methods.

1. はじめに

近年、インターネットの普及に伴って、SNS(ソーシャル・ネットワーク・サービス) や動画共有サイトといったサービスが多くの人に利用されるようになっている。これらは全て数種類のノード (ユーザー、リソースなど) から成るネットワーク、すなわち多種ノードネットワークとみなすことができる。従来手法の多くはリンク構造を重要な要素としてランキングを行っているが、World Wide Web のような一種類のノードから成るネットワークに対しては上手く機能していても、多種ノードネットワークへの適用には不十分な手法がほとんどである。全てのリンクは同一種類のノード間のリンクである intra-type と異なる種類のノード間のリンクである inter-type の二つのタイプに分類することができる。多種ノードネットワークはこの両方のタイプのリンクで表現されるが、従来手法の多くはどちらか一方のタイプしか上手く扱うことができない。

上記のような状況で多種ノードネットワークに効果的なランキング手法が求められる中、本研究では最も一般的なリンク分析手法である PageRank[4] と HITS[3] の考えを融合させた統合リンク分析を用いて、多種ノードネットワークにおける効果的なノードのランキングを試みる。この統合リンク分析は Wensi ら [5] が提案したもので intra-type と inter-type の両方のリンクを扱うことができるため多種ノードネットワークに有効であると思われる。一方、Wensi らは枠組みを示したものの intra-type を用いた実験は行っておらず、intra-type と inter-type の重みの決め方も示していない。それを踏まえて本研究では重みの決定方法を提案し、提案手法を実際に多種ノードネットワークに適用する実験を行った。得られたランキングについて平均逆順位 (MRR) やスコアの分散等を用いて評価した結果、従来の一般的なランキング手法と比べて有用と思われる結果を得ることができた。

連絡先: 藤井 通太, 東京工業大学 大学院理工学研究科 集積システム専攻 上野研究室, 〒 152-8552 東京都目黒区大岡山 2-12-1 S3-57, fujii@lab.ss.titech.ac.jp

2. 関連研究

2.1 リンク構造を用いたノードの評価手法

リンク構造を用いた代表的なノードの評価手法として、Page らの PageRank と Kleinberg の HITS が挙げられる。PageRank は Web ページの人気を測る中心性として考えられたものである。各 Web ページをノードとして考えると、良質なノードからリンクされているノードは良質なノードであるという考え方から、次の式で表される。

$$pr(p) = \alpha \sum_{q:(q,p)\in E} \frac{pr(q)}{out(q)} + (1-\alpha)\frac{1}{N} \tag{1}$$

$pr(p)$ はノード p の PageRank のスコア、(q,p) は q から p へのリンクを表す。$out(q)$ は q がリンクしている数、N は全ノード数、Web 閲覧者は α の確率でリンクをたどってノードを移動していくが、$(1-\alpha)$ の確率でリンクとは関係なくランダムなノードに移動するというランダムサーファーモデルを表現している。PageRank はこの計算を収束するまで再帰的に繰り返すことによって得たスコアでランキングを行う。

HITS は Web 全体からクエリに関連するノードだけを取り出し、そのリンク構造によって評価をする。各ノードにはハブスコアとオーソリティスコアという 2 つのスコアが割り当てられる。ハブスコアはリンクの質の高さで、オーソリティスコアは内容の質の高さである。ハブスコアは良質なノード (オーソリティスコアの高いノード) にリンクをすることで得られ、オーソリティスコアは質の高いリンクをしているノード (ハブスコアの高いノード) からリンクされることで得られる。

$$\begin{cases} authority(p) = \sum_{q:(q,p)\in E} hub(q) \\ hub(p) = \sum_{q:(q,p)\in E} authority(q) \end{cases} \tag{2}$$

オーソリティスコアはリンク元のハブスコアが加算され、ハブスコアはリンク先のオーソリティスコアが加算される。この 2 つのスコアを考慮してノードの評価をする。

2.2 Wensi らのリンク分析手法

Wensi らによるとリンクの種類は intra-type と inter-type に大別できる。前節で触れたように、intra-type は同じ種類の

ノードを結ぶリンク、inter-type は異なる種類のノードを結ぶリンクをそれぞれ表している。図 1 に例を示す。

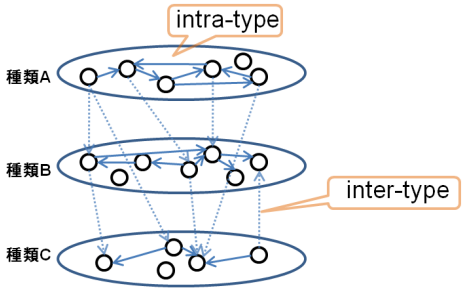

図 1: intra-type と inter-type

Wensi らはこれら両方のリンク分析が可能な手法である Link Fusion を提案した。その計算式は以下のようになっている。

$$\begin{cases} \boldsymbol{w}_M = \alpha_M {L'_M}^T \boldsymbol{w}_M + \beta_{NM} \sum_{\forall N \neq M} {L'_{NM}}^T \boldsymbol{w}_N \\ \text{ただし} \\ \alpha_M + \sum_{\forall N \neq M} \beta_{NM} = 1; \quad \alpha_M > 0 \quad \beta_{NM} > 0; \\ L'_M = \epsilon U + (1-\epsilon) L_M; \quad 0 < \epsilon < 1; \\ L'_{NM} = \delta_N U + (1-\delta_N) L_{NM}; \quad 0 < \delta_N < 1 \end{cases} \tag{3}$$

$\boldsymbol{w}_M$ は空間 M のノードのスコアベクトル (列ベクトル)、L_M, L_{NM} は正規化した隣接行列、α_M は空間 M 内のリンクの重み、β_{NM} は空間 N から空間 M へのリンクの重み、ϵ, δ はランダムサーファーモデルを表現するためのパラメータ、U は一様遷移行列 ($u_{ij} = 1/n$ for all i, j; n は空間 N に含まれる全ノード数) である。式の形から分かるように、これは PageRank と HITS を組み合わせたようなアルゴリズムである。

2.3 多種ノードネットワークに適用される手法

Hotho ら [2] は folksonomy に関する手法を提案している。folksonomy とはインターネット利用者がネット上の情報に対してタグ（分類・検索のためのキーワード）を付けることによって実現される分類方法である。Hotho らの実験では無向三部グラフに対する実験で有効性を確認をしている。

$$\begin{cases} \boldsymbol{w} = \alpha \boldsymbol{w} + \beta A \boldsymbol{w} + \gamma \boldsymbol{p} \\ \text{ただし} \\ \alpha + \beta + \gamma = 1; \quad \alpha, \beta, \gamma \in [0, 1] \end{cases} \tag{4}$$

$\boldsymbol{w}$ はスコアベクトル、A は正規化した隣接行列で、全種類のノード間のリンク構造を一つの行列で表している。$\boldsymbol{p}$ はユーザーの好みを表すベクトルであり、ユーザーの好む Web ページが与えられたとき $\boldsymbol{p}$ はユーザーが好む連結成分に高い重みを与える。

また、Jing ら [6] はソーシャルネットワークにおいてカテゴリーやユーザーの推薦を行うためのノードのスコア計算手法を提案している。Jing らの手法ではランダムサーファーモデルを用いているが、ネットワーク内の全ノードに対してランダムジャンプする確率は等しいと仮定している。

$$\begin{cases} \boldsymbol{s}_Y = \alpha E + (1-\alpha) \sum_{\lambda_{XY} \in \Lambda} \lambda_{XY} M_{XY}^T \boldsymbol{s}_X \\ \text{ただし} \\ E = (1/n, \cdots, 1/n)^T (1, \cdots, 1) \end{cases} \tag{5}$$

X, Y はノードの種類、$\boldsymbol{s}_X, \boldsymbol{s}_Y$ は種類 X, Y のノードのスコアベクトル、α はランダムサーファーモデルを表現するためのパラメータ、λ_{XY} は種類 X のノードから種類 Y のノードへの遷移確率 (重み)、Λ は λ_{XY} の集合、M_{XY} は X と Y の間のリンクに対応する遷移確率行列、n はネットワーク内の全てのノード数をそれぞれ表している。

これらの手法は数種類のノードを考慮してはいるが、Hotho らの手法では有向グラフに対して有効性が確認されておらず、Jing らの手法ではランダムジャンプの確率がノードの種類に関係なく等しく設定されてしまっている。一方、本研究でベースに用いる統合リンク分析ではこれらの問題点も解決できるため、より様々なネットワークに対して有効だと考えられる。

3. 提案手法

3.1 スコア計算式

多種ノードネットワークにおいて各種類のノードの集まりを空間と捉えることで、その空間ごとにノードのスコア計算を行えるようになり、ノードの種類を考慮したランキングが可能となる。このとき、同じ空間内のノードを結ぶリンクは intra-type、異なる空間のノードを結ぶリンクは inter-type とみなすことができる。intra-type と inter-type のリンクを扱う手法として Wensi らの Link Fusion を元にした統合リンク分析を考え、多種ノードネットワークに適用する。大規模ネットワークへの適用を考え、スコアベクトルを正規化することで Link Fusion と比べてメモリ使用量を削減を実現している。

$$\begin{aligned} \boldsymbol{w}_M(1-\epsilon)U &= (1-\epsilon)\boldsymbol{w}_M U \\ &= (1-\epsilon)(1/n, 1/n, \cdots, 1/n) \\ &= (1-\epsilon)\boldsymbol{u} \end{aligned} \tag{6}$$

ここで $\boldsymbol{w}_M$ は各要素の合計が 1 のスコアベクトル (行ベクトル)、ϵ は定数、U は一様遷移行列、$\boldsymbol{u}$ は 1 行 n 列の行ベクトルである。式 (6) の考えを用いて Link Fusion を改良し、本研究における統合リンク分析としてノードのスコア計算式を以下のように定める。

$$\begin{cases} \boldsymbol{w}_M = \alpha_{MM}\{\epsilon_{MM}\boldsymbol{w}_M L_{MM} + (1-\epsilon_{MM})\boldsymbol{u}\} \quad + \\ \qquad \sum_{\forall N \neq M} \alpha_{NM}\{\epsilon_{NM}\boldsymbol{w}_N L_{NM} + (1-\epsilon_{NM})\boldsymbol{u}\} \\ \text{ただし} \\ \alpha_{MM} + \sum_{\forall N \neq M} \alpha_{NM} = 1; \quad \alpha_{NM} \geq 0 \\ \boldsymbol{u} = (1/n, 1/n, \cdots, 1/n) \end{cases} \tag{7}$$

$\boldsymbol{w}_M$ は空間 M のノードのスコアベクトル (行ベクトル)、α_{NM} は空間 N から M への支持の伝搬係数、ϵ_{NM} はランダムサーファーモデルを表すためのパラメータ、L_{NM} は空間 N から M への正規化隣接行列、$\boldsymbol{u}$ は一様遷移を表す行ベクトルである。本研究では用いないが、隣接行列を用いるとスコアが正の無限大に発散してしまう場合でもこの式 (7) なら問題なく計算できる。提案手法では式 (7) を収束するまで反復計算しスコアを求めることでランキングを行う。

3.2 伝播係数の決定方法

intra-type と inter-type の伝播係数 α_{NM} の決定方法として本研究では以下の二つの方法を提案する。

(i) 空間 M への出リンクを持っているノードを含む空間の数で決定する

(ii) 各空間における、空間 M への出リンク数の比で決定する

(i) は各空間から空間 M へのリンクの重要度はそれぞれ等しいと考えた場合で、ノードの全ての種類の重要度が等しいと考えてランキングを行いたい時に効果的な手法である。これを提案手法 1 とする (式 (8))。

$$\alpha_{NM} = \begin{cases} 1/S_{toM} & (\text{N から M へのリンクが存在する}) \\ 0 & (\text{それ以外}) \end{cases} \tag{8}$$

ここで S_{toM} は空間 M への出リンクを持っているノードを含む空間数である。一方、(ii) はリンクの重要度は各空間におけ

る出リンク数に比例すると考えた場合で、多くの出リンクを持っているほどそのノードの種類が重要と考えてランキングを行いたい時に効果的な手法である。これを提案手法 2 とする (式 (9))。

$$\alpha_{NM} = \begin{cases} l_{NtoM}/l_{ALLtoM} & (\text{N から M へのリンクが存在する}) \\ 0 & (\text{それ以外}) \end{cases} \quad (9)$$

ここで l_{NtoM} は空間 N から空間 M へのリンク数、l_{ALLtoM} は全ての空間から空間 M へのリンク数である。伝播係数決定の例を図 2 に示す。今回は二つの伝播係数の決定方法を提案したが、伝播係数を意図的に変えることで他にも目的に応じた様々なランキングを得ることが可能である。

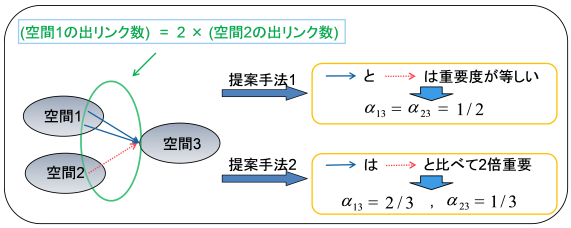

図 2: 伝播係数 α の決定方法

4. 実験

4.1 実験データ及び実験環境

本研究では提案手法の有用性を確認するためにソーシャルニュースサイトである Digg[1] のデータを用いて実験を行った。Digg のデータは、ユーザー、ストーリー、トピックの 3 種類のノードを含む多種ノードネットワークである。Digg の仕組みとしては主に以下のようになっている。

- ユーザーが興味のある記事 (ストーリー) にタグ (トピック) をつけて推薦 (submit) する
- どのトピックにも分類されないで submit されるストーリーも存在する。
- ユーザーはすでに submit されたストーリーで気に入ったものはさらに支持 (digg) することができる。
- 気に入ったユーザーがいれば、そのユーザーのファンになることもできる。

2010 年 10 月 8 日から 10 月 15 日の 1 週間に submit されたストーリーを対象としてクロールし、それに付けられたトピック、submit もしくは digg したユーザーを含めて、67,816 のユーザー、568,050 のストーリー、11(“ トピック無し ”を含む) のトピック情報を取得した。ちなみに Digg においてトピックの種類は“ トピック無し ”を含めて全部で 11 種類である。本実験ではノードの種類ごとに空間を作成し、ハイパーリンクを各空間のノード間のリンクに分解する。その際、ストーリーおよびトピックがユーザーを評価するという考えは適していないと思われるため、ストーリー、トピックからユーザーへのリンクは考慮しないこととした (図 3)。

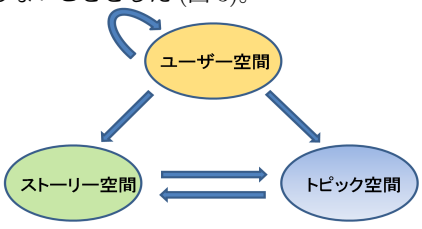

図 3: Digg 空間の関係

集められたデータから無作為に選んだ 10,000 のユーザー、10,000 のストーリー、それに 11 のトピックを加えた 20,011 のノードから成る 3 種ノードネットワークに対して実験を行った。表 1 は各種類のノード間のリンク数を示したものである。20,011 のノードに含まれていないノードからのリンクは全て無視し、種類に関係なくノード間のリンクは 1 本のみを考えている。ユーザー間のリンクは intra-type、それ以外のリンクは全て inter-type である。実験には Core2Duo E8400 3.00GHz, RAM 4GB, MATLAB Version7.11.0.584 を用いた。

表 1: 空間 N から空間 M へのリンク数

N \ M	ユーザー	ストーリー	トピック
ユーザー	25327	2617	1604
ストーリー	0	0	1374
トピック	0	1374	0

4.2 結果の評価

提案手法におけるランキング結果を PageRank、HITS と比較することで評価した。提案手法 1 と 2 においてユーザーのランキングのみ同じ結果になっている。

4.2.1 ランキングの有用性

上位のストーリーとトピックのランキング間の相関によって、有用性を評価した。平均逆順位 (MRR)、上位 k 位以内のストーリーにおけるトピックの人気比率 (@k) の 2 つの基準を用いた。平均逆順位は順位を考慮でき、人気比率はスコアの値を考慮できるので評価基準として適していると思われる。どちらも値が大きいほどストーリーのランキングとトピックのランキングに相関があり有用なランキングだと考えられる。

平均逆順位

それぞれの手法におけるトピックのランキング結果で上位のトピックを質問とみなし、ストーリーのランキング中に質問と関連するストーリーが含まれている場合、そのストーリーを正解とすることで以下の式 (10) より平均逆順位を求める。

$$MRR = \frac{1}{|q|} \sum_{q \in Q} \frac{1}{r_q} \quad (10)$$

$|q|$ は正解の数 (= 質問の数)、Q は上位 $|q|$ 位以内のトピックの集合、q は質問とみなされるトピック、r_q は質問 q に対する正解の中で最も良い順位を表す。結果を図 4 に示す。

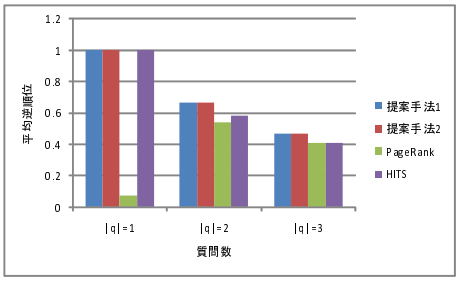

図 4: 平均逆順位

上位のストーリーの内容と上位のトピックが一致するほど平均逆順位は大きくなり、ストーリーとトピックのランキング間に相関があると言うことができる。上位 1 のトピックのみを質問とみなした時 PageRank だけ著しく低いのは、PageRank におけるトピック 1 位は technology であったが、ストーリーのランキングで technology に関する最上位のストーリーが 13 位であったためである。図 4 から分かるように、提案手法 1,2 共に PageRank、HITS と比べて全体的に上回っている。

上位のストーリーにおけるトピックの人気比率

限られたトピックに関するストーリーが上位を独占するよりも、様々なトピックに関するストーリーが上位に来る方がランキング間の相関があり、上手く機能していると言える。そこでもう一つランキング間の相関を測る評価基準として、上位 k 位

以内のストーリーに関するトピックのスコアがトピック全体に占める比率 @k を定義した。

$$@k = \frac{1}{s_{all}} \sum_{t \in T_k} s_t \tag{11}$$

s_{all} は全てのトピックのスコア合計、t はトピック、T_k は k 位以内のストーリーに関するトピックの集合、s_t はトピック t のスコアを表す。結果を図 5 に示す。

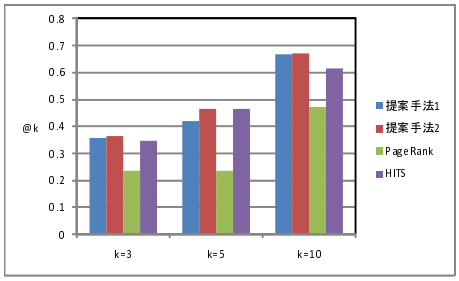

図 5: @k

@k は上位のトピックに関するストーリーが上位に来ることで大きくなるだけでなく、上位のストーリーのトピックが多様になることでも大きくなる。@k の値が大きいほど、それぞれのトピックにおいて人気のあるストーリーを特定できていると考えられる。図 5 より、PageRank や HITS と比べて提案手法の方が様々な内容に関する人気のあるストーリーを上位にランクインさせていることが分かる。

4.2.2 ランキングの安定性

ランキングの安定性をノードスコアの分散によって評価した。ネットワーク内にスコアがほぼ等しいノードがたくさんあることは、リンク構造が少し変化するだけで大幅にランキングが変わる可能性があることを意味する。それは上位のノードについても例外ではなく、スコアがある程度大きくても有力なノードからのリンクを失えば大幅にランキングを下げることに成りかねない。逆に、上位のノードのスコアと下位のノードのスコアの差が大きいほど、多少リンク構造が変わったとしても、ランキングに大きな変化は生じないため安定しているとみなすことができる。以上を踏まえ、各種類のノードにおけるスコアの散らばり具合からランキングの安定性について評価する。スコアの散らばり具合の評価基準として不偏分散 u^2 を用いる。

$$u^2 = \frac{1}{n_M - 1} \sum_{i=1}^{n_M} (\bar{s}_M - s_i)^2 \tag{12}$$

u^2 は空間 M 内の全ノードスコアの分散、n_M は空間 M 内のノード数、$\bar{s}_M$ は空間 M 内の全ノードスコアの期待値、s_i は空間 M 内のノード i のスコアを表している。結果を表 2 に示す。

表 2: 全ノードスコアの分散

手法 / ノードの種類	提案手法 1	提案手法 2	PageRank	HITS
ユーザー	3.45e-07	3.45e-07	7.47e-09	3.81e-07
ストーリー	1.63e-07	2.76e-07	1.63e-08	6.02e-09
トピック	1.12e-03	1.27e-03	3.84e-04	2.44e-06

表 2 を見ると、ユーザーにおいては提案手法と比べて HITS の結果の値が僅かに大きくなっているが、他 2 種類については提案手法の方が大きい。トピックに関しては HITS の結果における分散の値はとても小さく、人気のあるトピックと人気のないトピックの差別化があまりできていないので、全体的には提案手法の方がランキングが安定していると思われる。

実際には上位のノードのみが重視されることが多いので、上位 10 に入っているノードのスコアのみを考慮した分散も考え評価した。結果を表 3 に示す。

表 3: 上位 10 のノードスコアの分散. トピックに関しては上位 10 のトピックと全トピックはほぼ同じ意味を持つので割愛した.

手法 / ノードの種類	提案手法 1	提案手法 2	PageRank	HITS
ユーザー	1.08e-04	1.08e-04	2.14e-06	9.79e-07
ストーリー	3.68e-05	7.95e-05	4.24e-07	1.16e-07

いずれの種類も提案手法が PageRank、HITS を上回っており、ストーリーにおいては提案手法 2 が提案手法 1 を上回っている。この結果から、各空間に入り込むリンク数を考慮した方が上位のノードスコアの分散が大きくなり、安定したランキングを得ることができると思われる。

5. まとめと今後の課題

多種ノードネットワークに対して、統合リンク分析を用いてランキングを行う手法を提案した。Digg のデータを用いた実験で、各種類のノードスコアの伝搬係数を適切に決定し、PageRank、HITS の適用結果と比べた結果、平均逆順位やトピックの人気比率、スコアの分散の観点から、より有用と思われるランキングを得ることができた。

今後の課題としては、伝搬係数の決定方法の多様化が挙げられる。伝搬係数の決定方法として二つの方法を提案したが、数多くのネットワークに対応するためにも新たな決定方法を模索する必要がある。

参考文献

[1] Digg http://digg.com/news.

[2] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies:search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, Vol. 4011 of *LNAI*, pp. 411–426. Springer, Heidelberg, Jun 2006.

[3] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, Vol. 46(5), pp. 604–632, Sep 1999.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. *Stanford Digital Library working paper SIDLWP-1999-0120*, Nov 1999.

[5] Wensi Xi, Benyu Zhang, Zheng Chen, Yizhou Lu, Shuicheng Yan, Wei-Ying Ma, and Edward A. Fox. Link fusion: A unified link analysis framework for multi-type interrelated data objects. In *Proc. 13th International World Wide Web Conference*, New York, May 2004.

[6] Jing Zhang, Jie Tang, Bangyong Liang, Zi Yang, Sijie Wang, Jingjing Zuo, and Juanzi Li. Recommendation over a heterogeneous social network. In *Proc. of the 9th International Conference on Web-Age Information Management(WAIM)*, ZhangJiaJie, China, Jul 2008.