ページランク・アルゴリズムの可視化
ーテレポーテーションの秘密ー
Visualization of PageRank Algorithm
- Secret of Teleportation -

安田 友希　　**田嶋 祐衣**　　**向 直人**

Yuki Yasuda　　Yui Tajima　　Naoto Mukai

1. はじめに

今日、22 億を超える人々がインターネットを利用している。また、そのほとんどの利用者が検索サービスを使用している。検索エンジンには、Yahoo!や Google、マイクロソフト製の bing などがあるが、その中でも多くのシェアを誇るのが Google である。その理由は、Google の検索エンジンの中枢である、ページランク・アルゴリズムが利用者の求めるウェブページを高い精度でランク付けして提供するからである。このページランク・アルゴリズムを開発したのは、Google の創始者である、ラリー・ペイジとセルゲイ・ブリンであり、1998 年に論文『The Anatomy of a Large-Scale Hypertextual Web Search Engine』[1]の中で提案した。しかし、我々の多くはその仕組みを知らないまま利用している。そこで本研究では、ページランク・アルゴリズムの仕組みを aritsoc で可視化し、理解を深めることを目的とする。

一般に、検索エンジンを用いてウェブページを検索すると「マッチング」と「ランキング」の段階を得て検索結果が表示される。常時、クローラーは 1 兆ページを越える膨大なウェブページを探索し“インデックス（索引）”を作成している。「マッチング」では、このインデックスを基に、ユーザが入力したキーワードを含むウェブページを絞り込む。しかし、この段階を得てもユーザに提示するには過剰なウェブページが候補に残る（例えば、Google で“ページランク”というキーワードを検索すると 7,200,000 件がヒットする）。そこで、次の「ランキング」では、絞りこまれたウェブページが、ユーザが本当に必要としている情報を含むかどうかを判断し、順序付けを行う。最終的に、この順序に従って検索結果が表示される。本稿で注目するページランクはこのランキングの段階で使用されるアルゴリズムであり、ウェブページ間のリンク構造を基にその順序を決定する。実際のページランクの導出には、行列計算が用いられる。しかし、この行列計算を可視化するのは困難である。そこで、本稿では、ウェブサーファーをエージェントとして実装し、ページランクが導出されるまでのプロセスを視覚的に演出することで、数学者などの専門家以外でも理解できるよう工夫する。

本稿の構成を以下に示す。第 2 章では、数学的な表現を用いてページランク・アルゴリズムの詳細を記述する。第 3 章では、artisoc を利用してページランクの導出プロセスを可視化するためのエージェントの実装方法について述べる。第 4 章では、ページランク・アルゴリズムを名古屋の地下鉄網に適用した結果を報告する。最後に第 5 章でまとめと今後の課題を述べる。

椙山女学園大学 文化情報学部 文化情報学科
〒464-8662
名古屋市千種区星が丘元町 17 番 3 号

2. ページランクのアルゴリズム

ページランクのアルゴリズムは Amy N. Langville 氏らの『PageRank の数理 最強検索エンジンのランキング手法を求めて』[2]や John MacCormick 氏の『世界でもっとも強力な 9 つのアルゴリズム』[3]に詳しく記述されている。本章では、Amy N. Langville 氏らの例を利用しながら、ページランクの計算過程を説明する。

2.1 ページランクの基本公式

ページランクでは、ウェブページ間のリンクをウェブページの推薦と捉え、多くのウェブページからリンクされたウェブページの評価を高く設定する。ここでは、図 1 に示す 2 つのウェブページで構成されたネットワークを考える。

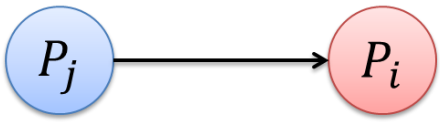

図 1　ウェブページ間のリンク

図中の矢印がリンクを表し、ウェブページP_jからウェブページP_iに遷移するリンクが存在することを意味している。この場合、このリンクを、ウェブページP_iの“入リンク”、ウェブページP_jの“出リンク”と呼ぶことにする。ここで、ウェブページP_iのページランク$r(P_i)$は式(1)で計算される。$|P_j|$はウェブページP_jからの出リンクの総数を表し、B_{P_i}はウェブページP_iに対し入リンクを持つウェブページの集合である。この式から、ウェブページP_jのページランク$r(P_j)$がリンクを経由して他のウェブページに拡散していくことが分かる。この伝播が反復的に実行されることでウェブページ全体のページランクが定まる。

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|} \qquad \text{式(1)}$$

2.2 ハイパーリンク行列

ページランクの計算は前節で述べた公式が基本となるが、ウェブページ全体を対象として反復的に計算するために、行列計算が用いられる（世界最大規模の行列計算として知

られる）。ここでは、図 2 の 6 つのウェブページで構成されるネットワークを考える。

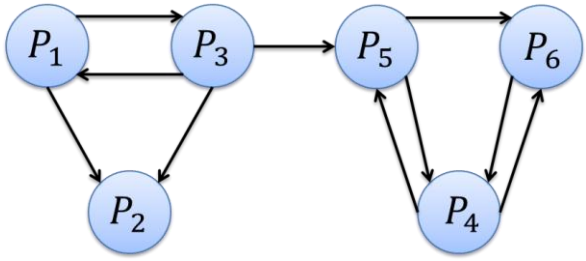

図 2　6 つのウェブページで構成されたネットワーク

まず、ウェブページのページランクの初期値を決定する。一般に、ネットワークに存在するページの総数をnとすると初期値は$1/n$で与えられる。式(2)は、「ページランク・ベクトル$\boldsymbol{\pi}^{\mathrm{T}}$」と呼ばれ、図 2 に含まれる 6 つのウェブページのページランクの初期値をベクトルで表現している（左の要素から順にウェブページP_1~P_6のページランクを表している）。

$$\boldsymbol{\pi}^{T} = (1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6) \qquad \text{式(2)}$$

次に、ウェブページ間のリンク関係を表す「ハイパーリンク行列$\mathbf{H}$」を式(3)で定義する。ハイパーリンク行列は$n \times n$の隣接行列であり、行i・列jの値が非ゼロ要素の場合、ウェブページP_iからウェブページP_jへのリンクが存在することを表す。また、各要素の値は“1”を出リンクの総数で割って求める。例えば、図 2 のP_1はP_2とP_3への出リンクを持っているため、行1・列2と行1・列3の値は“1/2”となる。一方、P_2は出リンクを持たないため行 2 の全ての要素は“0”となる。

$$\mathbf{H} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \qquad \text{式(3)}$$

最終的にページランクを導出するには、ページランク・ベクトル$\boldsymbol{\pi}^{\mathrm{T}}$とハイパーリンク行列$\mathbf{H}$の積を式(4)に従って反復的に計算する。ここで、$\boldsymbol{\pi}^{(K)T}$はK回目の繰り返しにおけるページランク・ベクトルを表している。

$$\boldsymbol{\pi}^{(K+1)T} = \boldsymbol{\pi}^{(K)T}\boldsymbol{H} \qquad \text{式(4)}$$

例えば、図 2 のネットワークにおいて、上記の式を反復的に適用すると表 1 に示す結果となる。全てのページランクの初期値は$1/6$であるが、繰り返しと共に、リンクを介して他のウェブページにページランクが伝播することが分かる。K = 2では、P_4のページランクが最も高く、P_1とP_3のページランクが最も低い値となった。

K = 0	K = 1	K = 2
$r(P_1) = 1/6$	$r(P_1) = 1/18$	$r(P_1) = 1/36$
$r(P_2) = 1/6$	$r(P_2) = 5/36$	$r(P_2) = 1/18$
$r(P_3) = 1/6$	$r(P_3) = 1/12$	$r(P_3) = 1/36$
$r(P_4) = 1/6$	$r(P_4) = 1/4$	$r(P_4) = 17/72$
$r(P_5) = 1/6$	$r(P_5) = 5/36$	$r(P_5) = 11/72$
$r(P_6) = 1/6$	$r(P_6) = 1/6$	$r(P_6) = 14/72$

表 1　反復によるページランクの導出

2.3 推移確率行列

前節で述べたページランクの導出方法では、図 3 に示す「閉じ込め」や、図 4 に示す「閉路」を有するネットワークにおいて、正しくページランクを計算することができない。「閉じ込め」では、出リンクを持たないウェブページ（図中の“3”）がページランクを独占することになり、他のウェブページへの伝播が断ち切られてしまう。また、「閉路」では、ウェブページ間（図中の“1”と“2”）で反復毎にページランクを交換することになり、ページランクの値が収束しない。

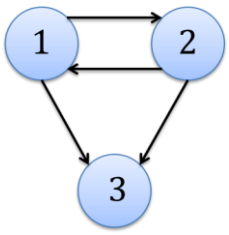

図 3　閉じ込め

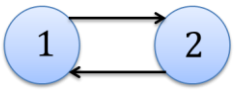

図 4　閉路

そこで、上記の問題を解決するために、ハイパーリンク行列$\mathbf{H}$を式(5)のように修正した。この行列は「推移確率行列$\mathbf{S}$」と呼ばれ、出リンクを持たないウェブページ（図 2 におけるP_2）の行要素を$1/n$に変更したものである。これは、出リンクを持たないウェブページから他の全てのウェブページへの仮想的なリンクを強制的に設定したとみなすことができる。これにより、「閉じ込め」のあるネットワークにおいても、ページランクが他のウェブページに平等に伝播されることになる。

$$\mathbf{S} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \qquad \text{式(5)}$$

2.4 テレポーテーション行列

推移確率行列**S**を導入しても、「閉路」がある場合、ページランクが閉じたネットワーク内に溜まってしまう。そこで、登場するのが“テレポーテーション（瞬間移動）”という概念である。ページランク・アルゴリズムにおけるテレポーテーションとは、一定の確率でリンクの有無とは無関係に、他のウェブページにページランクを伝播することを意味する。つまり、各ウェブページは、自身を含む全てのウェブページに対し、出リンクを仮想的に持つことになる。この挙動を表すための行列は「テレポーテーション行列**E**」と呼ばれ式(6)で与えられる。また、ページランクの伝播方法を確率的に決定するために閾値$\alpha(0 \le \alpha \le 1)$を導入する。例えば$\alpha = 0.85$であれば、85%の確率で「推移確率行列**S**」を選択し、15%の確率で「テレポーテーション行列**E**」を選択する。

$$\mathbf{E} = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix} \quad \text{式(6)}$$

3. Artisoc によるページランクの視覚化

本章では、2 章で述べたページランク・アルゴリズムを artisoc で実装する方法について述べる。本研究の目的は、ページランクの伝播の様子を可視化することで、学習者の理解を助けることにある。そこで、行列計算を用いてページランクを導出するのではなく、エージェント（ウェブサーファー）によって伝播させることで実現する。

エージェントによるページランクの伝播の様子を図 6 に示す。今、P_1に 10 体のエージェントが存在する。これらのエージェントはP_1の出リンクを辿り、他のウェブページに移動する。このとき、エージェントはP_1のページランク$r(P_1)$を、10 体で分け合い、移動先のウェブページに運ぶ。例えば、$r(P_1) = 1$とすると、1 体のエージェントが持つランクは$1/10$となる。P_1の出リンクはP_2とP_3の 2 つであるため、エージェントは 50%の確率でP_2またはP_3に移動する。この結果、P_2とP_3にそれぞれ 5 体のエージェントが移動したとすると、P_2とP_3のページランクは$5/10 = 1/2$となる。このプロセスを繰り返すことで、行列計算と同じ結果を導き出すことが可能となる。

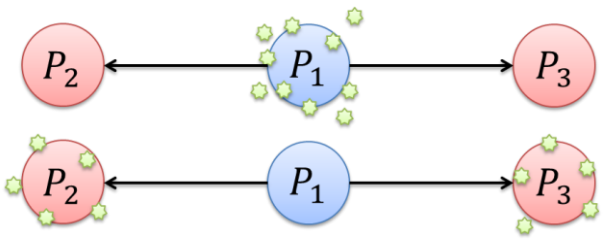

図 6　エージェントによるページランクの伝播

ページランク・アルゴリズムの効果を段階的に示すために、3 種類のエージェント（「ハイパーリンク・エージェント」、「推移確率エージェント」、「テレポーテーション・エージェント」）を定義する。次節では、これらのエージェントのフローチャートを示し図 2 のネットワークに適用した結果を示す。

3.1 ハイパーリンク・エージェント

ハイパーリンク・エージェントは「ハイパーリンク行列**H**」に従ってページランクを伝播する。ハイパーリンク・エージェントのフローチャートを図 7 に示す。まず、エージェントは、現在訪れているウェブページから、他のウェブページへ運搬するページランクの量を計算する（ページランクの値÷エージェント数）。次に、ウェブページが出リンクを有しているかどうかを調べる。出リンクを有している場合は、リンクを辿り、移動先のウェブページに運び出したページランクを加算する。もし、出リンクがない場合は、ページランクの伝播は行わず、その場に留まる。

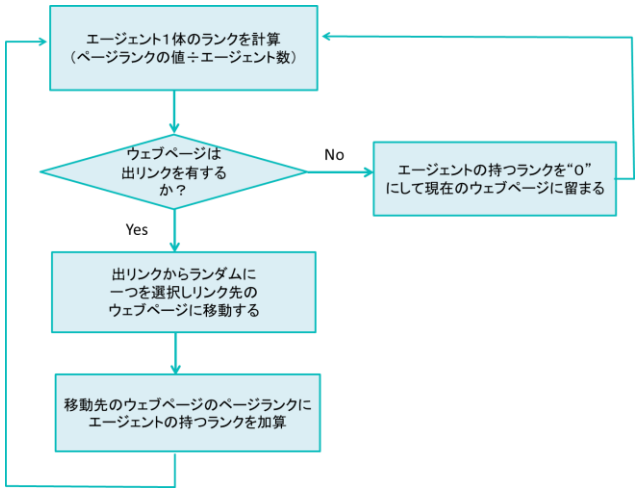

図 7　フローチャート
（ハイパーリンク・エージェント）

ハイパーリンク・エージェントを図 2 のネットワークに適用した結果が図 8 と図 9 である。図 8 はウェブページに滞在しているエージェントの様子を示している（ウェブページ周辺の小さな丸がエージェント）。P_1とP_3は入リンクを持たないため、エージェントは存在しなくなることが分かる。一方、入リンクを持つP_2・P_4・P_5・P_6には一定量のエージェントが滞在していることが分かる。ここで注意すべきは、P_2が出リンクを持たないため、P_2を訪れたエージェントは行き場を失い永遠にその場に留まってしまうことである。これは 2 章で述べた「閉じ込め」の現象を表している。図 9 は各ウェブページのページランクの値を表している。P_1とP_3が入リンクを持たないこと、また、P_2において「閉じ込め」が発生することにより、P_1・P_2・P_3のページランクは“0”になってしまうことが分かる。

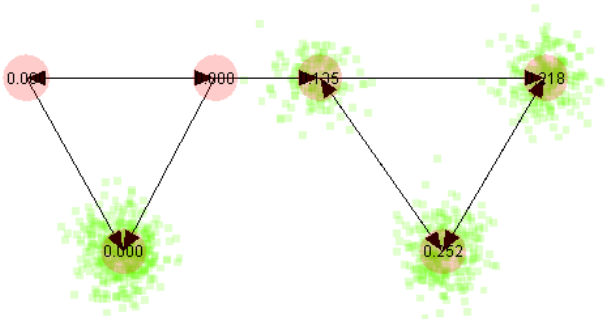

図 8　エージェントの状態
（ハイパーリンク・エージェント）

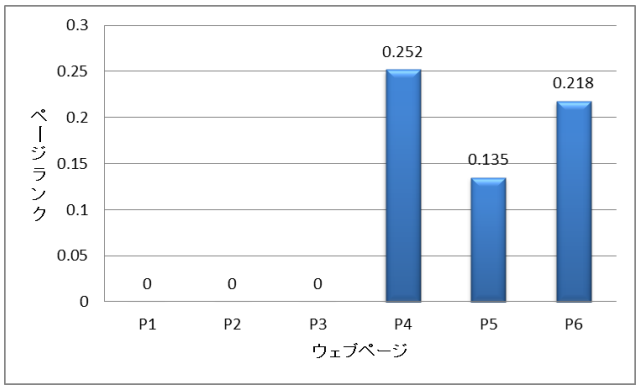

図 9　ページランクの結果
（ハイパーリンク・エージェント）

3.2 推移確率・エージェント

推移確率・エージェントは「推移確率行列**S**」に従ってページランクを伝播する。推移確率・エージェントのフローチャートを図 10 に示す。ハイパーリンク・エージェントとの振る舞いの違いは、ウェブページが出リンクを有していない場合に、全ウェブページからランダムに一つを選択し移動することである。これにより「閉じ込め」による弊害を防ぐことができる。

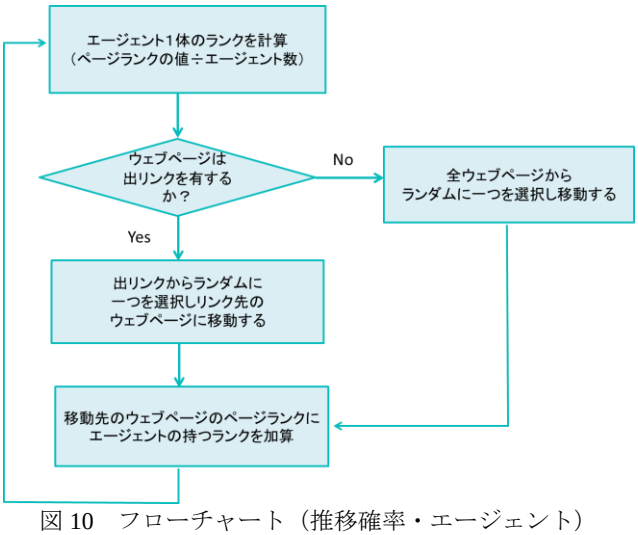

図 10　フローチャート（推移確率・エージェント）

推移確率・エージェントを図 2 のネットワークに適用した結果が図 11 と図 12 である。図 11 から、出リンクを持たないP_2であっても、エージェントが留まることなく脱出できていることが分かる。しかし、P_4・P_5・P_6が「閉路」となっているため、P_1・P_2・P_3にエージェントが流れこむことが出来ない。この結果、図 12 から分かるように、P_1・P_2・P_3のページランクはやはり“0”になってしまう。

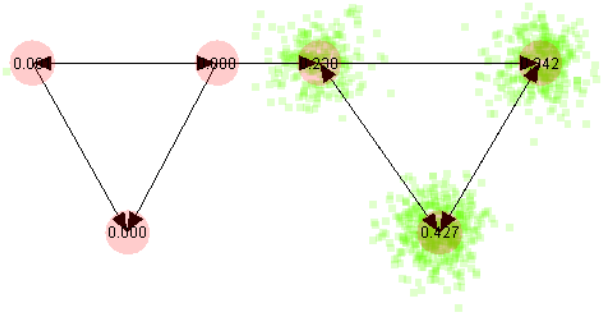

図 11　エージェントの状態
（推移確率・エージェント）

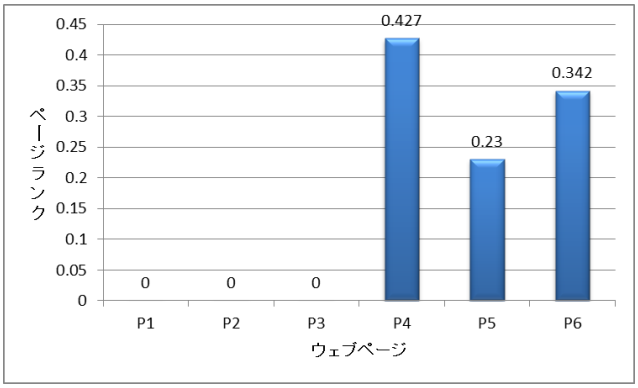

図 12　ページランクの結果
（推移確率・エージェント）

3.3 テレポーテーション・エージェント

テレポーテーション・エージェントは閾値αに従い確率的に、推移確率行列**S**とテレポーテーション行列**E**を使い分ける。本実験では、閾値αの値は Google が実際に用いている値と同じ 0.85 を用いた。一定確率でテレポーテーションすることにより「閉路」による弊害を防ぐことができる。

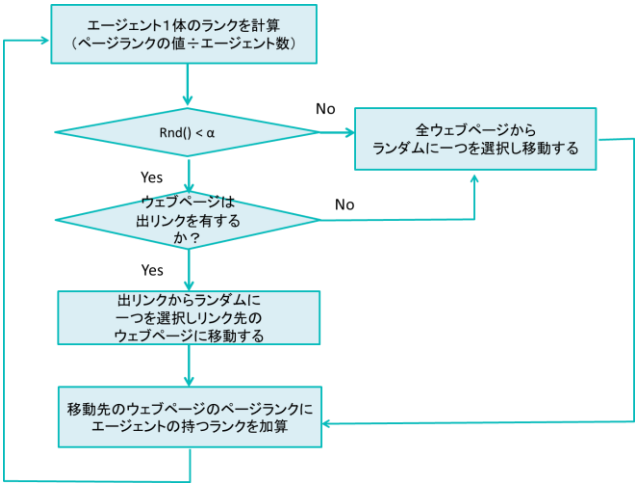

図 13　フローチャート
（テレポーテーション・エージェント）

テレポーテーション・エージェントを図 2 のネットワークに適用した結果が図 14 と図 15 である。図 14 から、テレポーテーションにより、偏りはあるものの、全てのウェブページにエージェントが分散して存在していることが分かる。また、図 15 から、全てのウェブページのページランクの値が“0”になることなく、一定の値に収束していることが分かる。この結果を用いてP_4・P_6・P_5・P_2・P_3・P_1の順にウェブページをランク付けることができる（これは行列計算で求めた順位と一致する）。

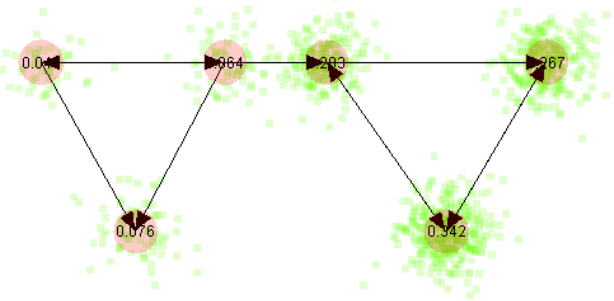

図 14　エージェントの状態
（テレポーテーション・エージェント）

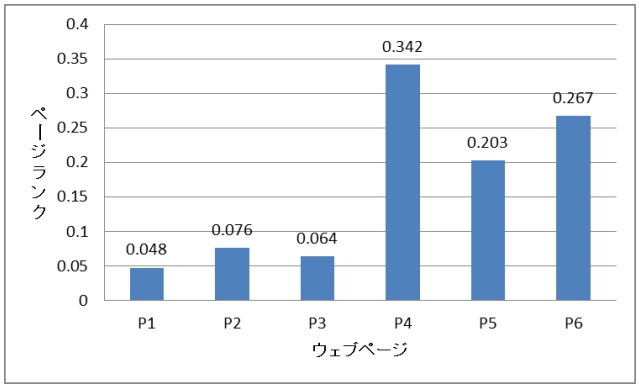

図 15　ページランクの結果
（テレポーテーション・エージェント）

4. 名古屋の地下鉄網への応用

ページランク・アルゴリズムはリンク構造という極めてシンプルなルールに基づき計算される。このため、ウェブページだけではなく、その他の一般的なネットワークにも汎用的に適用することが可能である。そこで、名古屋の地下鉄網にページランクを適用した結果を報告する。

名古屋市交通局のホームページ[4]に掲載されている地下鉄路線図を基に図 16 のようにネットワークを構成した。“鶴舞線”と“東山線”が利用可能な“名古屋駅”などの「2 路線以上が交差する駅」や、“東山線”の終点である“高畑駅”などの「ターミナル駅」をネットワークのノードとした。これにテレポーテーション・エージェントを適用した結果が図 17 である。図 2 の小さなネットワークと比べるとページランクの値の収束が難しく、駅間の明確な優位差を得ることができなかったが、やはり“名古屋駅”や“伏見駅”などの主要な駅のページランクが高くなる結果が得られた。

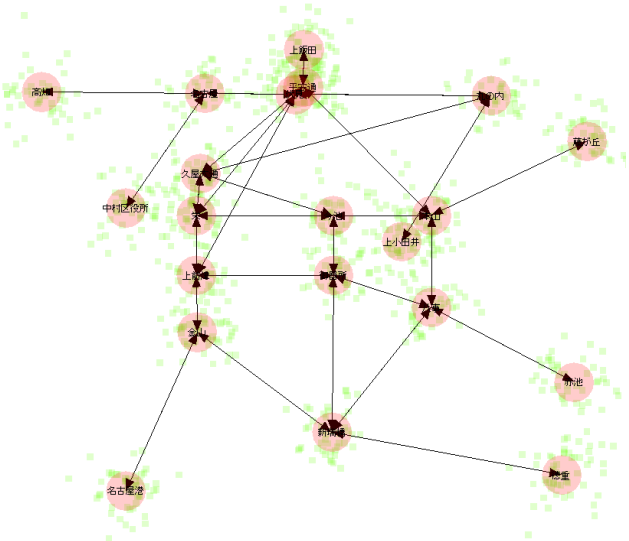

図 16　名古屋の地下鉄網

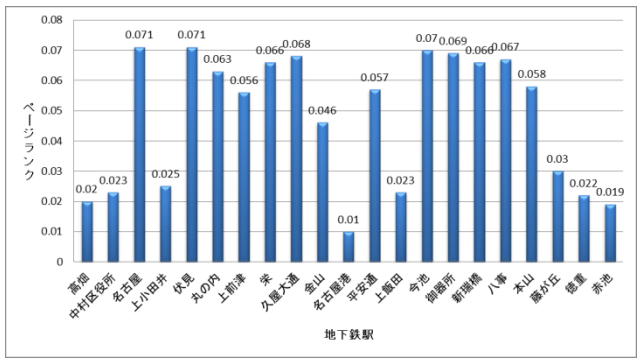

図 17　ページランクの結果（名古屋の地下鉄網）

5. まとめ

本研究では、誰もが利用する検索エンジンの裏側に潜むページランクというアルゴリズムに焦点を当て、本来は行列計算が必要な導出手続きを、エージェントモデルに基づき実装することで、可視化することに成功した。また、ページランク・アルゴリズムを段階的に理解するため、「ハイパーリンク・エージェント」、「推移確率・エージェント」、「テレポーテーション・エージェント」の３種類のエージェントを定義した。これにより、数学者などの専門家だけでなく、情報工学を学ぶ学生はもちろん、インターネットを利用する全ての人々に、ページランクの仕組みを直感的に伝えることが可能になった。今後は、ウェブページだけでなく、その他の一般的なネットワークにもページランクを適用することで、社会にとって有用な知識を発見したいと考えている。

参考文献

[1] S. Brin, L.Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Seventh International World-Wide Web Conference (1998)

[2] Amy N. Langville, Carl D. Meyer 著, 岩野和生・黒川利明, 黒川洋訳, “PageRank の数理 最強検索エンジンのランキング手法を求めて”, 共立出版 (2009).

[3] John MacCormic,著, 長尾高弘訳, “世界でもっとも強力な９つのアルゴリズム”, 日経 BP 社 (2012).

[4] 名古屋市交通局, http://www.kotsu.city.nagoya.jp/