リンク構造を用いた類似度計算手法の改良

坂本　剛彦†　　田島　敬史††

† 北陸先端科学技術大学院大学情報科学研究科　〒 923–1292 石川県能美郡辰口町旭台 1-1
†† 京都大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †s-take@jaist.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　本稿では，リンク構造を用いたグラフの類似度計算に関する既存手法の問題点と改善手法について議論する。代表的な既存手法である SimRank では，グラフの全てのノード対に対して，対の各ノードへのリンクを持つノードの間の類似度をノード対の類似度に伝播させることで類似度を計算する。この伝搬を再帰的に行うことにより，グラフ全体にリンク元のノードの対の類似度を伝播させることができる。しかし，この手法では，リンクで隣接しているノードの対は非常に関係が深いと考えられるにも関わらず，そのような関係は類似度に反映されないという問題や，比較するノード対の各々にリンクしているノードの間の類似度しか考慮しないために，比較する一方のノードへリンクしているノードともう一方のノード自身が類似していたとしても，それは反映されないという問題ががあり、そのため類似度が全体的にスパースになってしまうという欠点がある。そこで本稿では，SimRank のアルゴリズムにノード間のリンクの考慮とノード自身に対するセルフリンクを付加することで，上の問題点を改善する手法を提案する。
キーワード　SimRank，類似度計算，リンク解析，反復アルゴリズム

Takehiko SAKAMOTO† and Keishi TAJIMA††

† School of Information Science, Japan Advanced Institute of Science and Technology
Asahidai 1–1, Tatsunokuchi-machi, Nomi-gun, Ishikawa, 923–1292 Japan
†† Department of Social Informatics, Kyoto University
Yoshidahonmachi, Sakyou-ku, Kyoto-shi, 606–8501 Japan
E-mail: †s-take@jaist.ac.jp, ††tajima@i.kyoto-u.ac.jp

1. ま え が き

人と人との繋がりを示すソーシャルネットワークへの関心が高まっている。私たちの住む世界を友人関係や師弟関係などの人と人との間における様々な関係から成るヒューマンネットワークとして表すことができる。このヒューマンネットワークをノードとリンクをもったグラフで表すことができる。グラフのノードは人を表し、グラフのリンクは人と人のつながりを表し、リンクでつながっているノード同士は何らかの関係をもつ。

研究者間における関係の例として、ある研究者 A が書いた論文 A がある研究者 B が書いた論文 B を引用していたなら、研究領域において研究者 A と研究者 B には何らかの関係がある。この関係によって、研究者 A の書いた論文を読んだ研究者 C は研究者 B の書いた論文を読むかもしれない。この場合、ノードは研究者、リンクは論文引用、ノード間の関係は研究者間の関心度合いを表す。

例であげたようなノード間のリンクを基にした関係性の強弱は類似度を用いて表されることが多い。リンク構造を用いて類似度を計算することは有用である。例えば、ウェブページ群に対してリンク構造を基に類似度を求めてあるウェブページに関連するウェブページを探すこともできる。そのため、リンク構造を用いた類似度計算方法はいくつか提案されている。

リンク構造を用いた類似度計算方法の中で代表的なものとして SimRank [1] という計算方法がある。SimRank はノード同士の周辺の類似度を ”類似するオブジェクト同士は類似するオブジェクトに関連付けられる (similar objects are related to similar objects) ”という直観を基に任意の二つのノードに対して共引用を反復的に計算し、類似度を求め、類似度を伝播させるアルゴリズムである。SimRank はリンク構造を用いた類似度計算方法としては良く定義されているが、いくつか問題点がある。

まず一つに、SimRank を同じ関係ドメイン中のノード同士エッジを持たない二部グラフに対して適用した場合には問題は無いが、グラフに対して適用した場合に問題が生じる。例えば、著者と書籍の関係を表す二部グラフへの適応は問題ないが、著者と著者の関係を表すグラフへの適応は問題がある。これは SimRank が同関係ドメイン上にある二つのノードの類似

度を計算する際に二つのノード間のリンクを考慮しないため、二つのノード間にリンクが存在するという，明らかにノード間の関連の高さを表している情報を見逃してしまう。二つめに、SimRank が類似度を計算する二つのノードに共引用となりうるインリンクで隣接するノードの対の類似度のみしか類似度計算に反映させないところに問題がある。例えば、ある二つのノードの SimRank のスコアを求める際に、あるノードにインリンクで隣接するノードがもう片方のノードに類似していたとしても、SimRank はインリンクで隣接するノードの対でしか類似度を考慮しないために類似度計算に前者の類似度を含めることはできない。最悪の場合、片方のノードにインリンクで隣接している一つのノードがもう片方のノードに類似していたとしても、すべての隣接ノードの対の類似度が 0 ならば、SimRank の類似度が 0 になってしまう。

一旦，あるノード対の類似度が 0 になるとその箇所から類似度を伝播させることができずに、全体的に類似度をスパースにさせてしまうといった問題を引き起こす。スパース問題は ”類似するオブジェクト同士は類似するオブジェクトに関連付けられる ”という SimRank の趣旨の機能実現を妨げる。そこで、我々は SimRank の ”類似するオブジェクト同士は類似するオブジェクト同士に関連付けられる ”という趣旨により沿った ”類似するオブジェクト同士は類似するオブジェクト、または、オブジェクト同士により関連付けられる ”という趣旨を基に SimRank を拡張、改善するための類似度計算方法を提案する。類似度計算に対して新しい解釈や類似度を計算するグラフに対して新しいリンクを付加することにより改善を行う。

本稿では著者-著者関係を持つ同一ドメイン上のグラフにおける類似度について言及する。我々の提案する類似度計算の仕組みが現実のデータセットに対していかに作用するかを述べる。実験に用いた現実のデータセットは DBLP の書誌データからなる。

本稿の主な貢献は以下の通りである。

- 我々は SimRank アルゴリズムを拡張した、任意のノード間にエッジが存在しうる一般のグラフに適した類似度計算方法を提案し、類似度計算方法の形式的定義を示す。(第 3 章)
- 我々は SimRank と我々の提案する類似度計算方法の違いを実験結果を元に評価し報告する。(第 4 章)

第 2 章では SimRank について述べる。

2. SimRank

SimRank は任意の二つのノードに対して 0 から 1 の範囲内で類似度を算出する。SimRank の二つのノードの類似度の計算方法の形式的定義は式 (1) となる。

$$Sim_{l+1}(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} Sim_l(I_i(a), I_j(b)) \quad (1)$$

SimRank のアルゴリズムは ”類似するオブジェクト同士は類似するオブジェクト同士に関連付けられる ”という直観を基に類似度の伝播を繰り返す再帰的な不動点アルゴリズムである。

二つのノード a とノード b の類似度を求める際に二つのノードが同じ (a=b) ならば類似度は 1 になり、二つのノードが同じでないなら式 (1) を用いる。C は $0 < C < 1$ の定数で、反復計算の減衰因子として用いられる。

式 (1) に用いられている、$I(a)$、$I(b)$ はインリンクで隣接するノードの集合を表し、$I(a)$ はノード a にインリンクで隣接するノードの集合を意味する。$I_i(a)$、$I_j(b)$ は個々のインリンクで隣接するノードを表し、$|I(a)|$、$|I(b)|$ はインリンクで隣接するノードの数を表し、$|I(a)||I(b)|$ はインリンクで隣接するノードの対の総数を表す。i と j に関しては $1 \leqq i \leqq |I(a)|$、$1 \leqq j \leqq |I(b)|$ となる。式 (1) より、あるノード a とノード b の類似度は $I(a)$ と $I(b)$ のすべてのインリンクで隣接するノードの対、$I_i(a)$ と $I_j(b)$、に対する類似度の合計として計算される。l は反復の回数を表す。$l = 0$ のときの SimRank の最初の反復は $I_i(a)$ と $I_j(b)$ が等しくなる $Sim_l(I_i(a), I_j(b)) = 1$ の場合、すなわち共引用のみを考慮に入れるが、$l > 1$ のときに SimRank の反復は共引用と $I_i(a)$ と $I_j(b)$ が等しくなくかつ $Sim_l(I_i(a), I_j(b)) > 0$ となる類似度の伝播も考慮に入れる。つまり、SimRank は共引用のスコアを基に反復間で類似度を伝播させていくのである。反復計算によって合計計算の中に含まれる二つのノードの類似度 $Sim_l(I_i(a), I_j(b))$ が上昇すると、$Sim_{l+1}(a,b)$ の類似度も上昇するという類似度の伝播が SimRank の特徴である。SimRank はこの反復計算によって ”類似するオブジェクト同士は類似するオブジェクトに関連付けられる ”という趣旨を満たす。

二つのノードの SimRank のスコアは対称となるので、$Sim_l(a,b) = Sim_l(b,a)$ である。SimRank のスコアが $Sim_l(a,a) = 1$ となっても、$Sim_l(a,b) = 1$ となることは決してない。なぜなら、SimRank のスコアは式 (1) より右辺の減衰因子 C の掛け算により必ず 1 より低いスコアに減衰するようになっている。これは同じもの同士の類似度より違うもの同士の類似度の方が信頼性が低いことを述べている。

すべてのノードのペアについて類似度を計算するために SimRank の空間計算量は $O(n^2)$、時間計算量は $O(Ln^2d_2)$ となる。空間計算量は Sim_l の計算結果をストアするために単純に $O(n^2)$ となる。時間計算量の中の L は反復回数を示し、d_2 は $|I(a)||I(b)|$ のインリンクで隣接するノードの対の平均個数を示す。時間計算量はすべてのノードペア n^2 のスコアをインリンクで隣接するノードの対 d_2 からの類似度で計算するために各反復での時間計算量は $O(n^2d_2)$ となり、すべての反復での時間計算量は $O(Ln^2d_2)$ となる。ノード a とノード b の SimRank のスコアは $\lim_{l\to\infty} Sim_l(a,b)$ で定められるが、SimRank の全体のスコアは反復回数 5 以内で収束するので不動点アルゴリズムとしてはとても安定している。

本稿では SimRank で使われるこの表記方法と同じものを使う。

3. SimRank の改良

SimRank のアルゴリズムは共引用のスコアを基に類似度を伝播させるものである。これは ”類似するオブジェクト同士は類似するオブジェクト同士に関連付けられる ”という方針を実現するためのものであった。しかし、SimRank のアルゴリズムを

用いた場合に類似度を伝播しづらいあるいは伝播しないケースがいくつかある。類似度を伝播させられないと SimRank のスコアが全体的に低く抑えられる。その場合、特に ”類似するオブジェクト同士は類似するオブジェクト同士に関連付けられる ” という方針を満たすことができない。

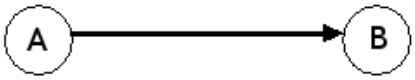

図 1

例えば、図 1 のような片方のノード A がもう片方へのノード B へのリンクだけを持つグラフに対して SimRank を計算する。ノード A にインリンクがないために SimRank のスコアは 0 になる。人間の目で図 1 を直視すれば二つのノード A とノード B に関連性がありそうだが、SimRank は二つのオブジェクトの類似度を計算する際にその二つのオブジェクト間のリンクを考慮に入れない。ちなみに、このような問題は二部グラフでは起きない。なぜなら、著者-著者のような同じ関係ドメイン中のノード同士でリンクを持たないからである。しかし、一般のグラフの場合はすべてのノード同士の間でリンクを持つことができるためにこのような問題が生じる。

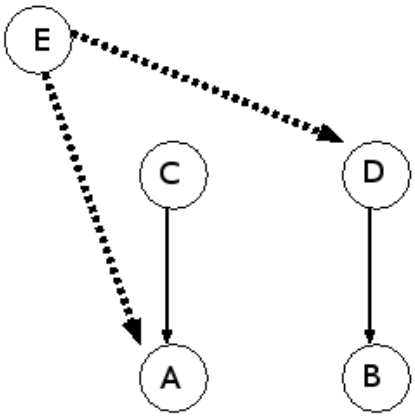

図 2

また、図 2 のようなノード B の隣接するノード D がノード A に対してノード E を基に類似性をもっている元で、ノード A とノード B に対して SimRank を計算する。ノード A とノード B に隣接するインリンクの対はノード C とノード D となり、ノード A とノード B の SimRank のスコアはノード C とノード D の類似度に左右される。ノード C とノード D の類似度が 0 ならば、ノード B の隣接するノード D がノード A に対してノード E を基に類似性を持っていたとしても、ノード A とノード B の SimRank のスコアは 0 となる。SimRank は ”類似するオブジェクト同士は類似するオブジェクト同士に関連付けられる ”という直観には従っているが、図 1、図 2 のような直観的に類似していると判断できるケースを考慮しない。なぜなら、SimRank は ”オブジェクト同士 ”という趣旨の枠内でインリンクで隣接するノードの対のみの類似度を合計するアルゴリズムだからである。

そこで我々はインリンクで隣接するノードの対のみならず、図 1 や図 2 のようなリンクを考慮できるように SimRank を拡張する。この拡張により ”類似するオブジェクト同士は類似するオブジェクト同士に関連付けられる ”というよりは ”類似するオブジェクト同士は類似するオブジェクト、または、オブジェクト同士により関連付けられる ”という直観に変わる。

図 1 の二つのノード類似度を計算する際に二つのノードの間のリンクを考慮に入れるために SimRank の形式的定義を以下のように式を変更する。

$$Sim_{l+1}(a,b) = \frac{C}{|I(a)||I(b)|+1}\Big(\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|} Sim_l(I_i(a), I_j(b)) + \frac{1}{2}In(b \to a) + \frac{1}{2}In(a \to b)\Big) \quad (2)$$

また、二つのノード間にリンクがあるかないかを以下のように定義する。

$$In(b \to a) = \begin{cases} 1, & \text{b から a にインリンクがあれば} \\ 0, & \text{b から a にインリンクがなければ} \end{cases}$$

式 (2) では、従来の SimRank に二つのオブジェクト間にリンクがあれば SimRank のスコアに 0.5 を足し、二つのオブジェクト間にリンクがなければ SimRank のスコアに変更を加えず、さらに、新しい類似度加算の基準を一つ加えた分、分母に 1 を足してスコアを割っている。

式 (2) はノード間のリンクを考慮するのみならず、インリンクで隣接するノードの対を形成できなかったとしても、ノード間にリンクがあれば類似度が 0 以上になる。このため、従来の SimRank よりも類似度の伝播をより強く行うことができる。

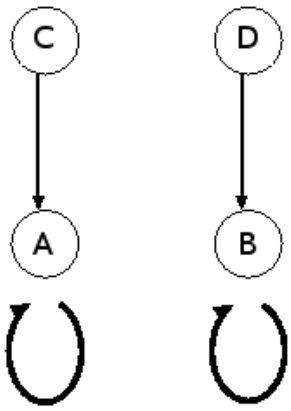

図 3

次に図 2 の二つのノードの類似度を計算する際に片方のノードにインリンクで隣接するノードが、もう片方のノードに関連しているケースを考慮に入れるために図 3 のようにグラフを SimRank の計算時に変化させる。図 3 では類似度を計算する二つの各ノードに自身へのセルフリンクを加えている。セルフリンクを加えることにより、セルフリンクのあるノードともう片方のノードにインリンクで隣接するノードとの関連性を考慮

に入れることができる。図 3 のグラフの変更に対応するために SimRank の形式的定義を以下のように式変形させる。

$$Sim_{l+1}(a,b) = \frac{C}{|I(a)||I(b)|+|I(a)|+|I(b)|}\Bigg(\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|} Sim_l(I_i(a),I_j(b)) + \sum_{j=1}^{|I(b)|} Sim_l(a,I_j(b)) + \sum_{i=1}^{|I(a)|} Sim_l(b,I_i(a))\Bigg) \quad (3)$$

式 (3) では、セルフリンクと他方のインリンクで隣接するノードの対の類似度を加算している。また、加算した分だけ分母に足して割っている。但し、従来の SimRank の計算の中にはこのセルフインリンクを加味しない。セルフリンクのあるノードと他方のノードにインリンクで隣接するノードの対を考慮に入れる直観は ”類似するオブジェクト同士は類似するオブジェクトに関連付けられる ”である。これは SimRank の定める対にならなくても図 2 のようにあるノードにインリンクで隣接するノードが他方のノードに類似していれば関連付けられるというのが我々の方針である。式 (3) はインリンクで隣接するノードと SimRank のスコアを求めるもう片方のノードとの類似度を考慮するのみならず、対となるノードの類似度がたとえ 0 だったとしても、インリンクで隣接する単体のノードと他方のノードとの類似度が 0 以上になれば SimRank のスコアは 0 以上になる。このため、従来の SimRank よりも類似度の伝播をより強く行うことができる。

図 1 と図 2 の両方の問題を扱うために、式 (2) と式 (3) を組み合わせて式 (4) を生成する。

$$Sim_{l+1}(a,b) = \frac{C}{(|I(a)|+1)(|I(b)|+1)}\Bigg(\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|} Sim_l(I_i(a),I_j(b)) + \sum_{j=1}^{|I(b)|} Sim_l(a,I_j(b)) + \sum_{i=1}^{|I(a)|} Sim_l(b,I_i(a)) + \frac{1}{2}In(b\to a)+\frac{1}{2}In(a\to b)\Bigg) \quad (4)$$

式 (4) は式 (2) と式 (3) の個々の SimRank の変形よりも類似度の伝播をより強く行うことができる。これにより、”類似するオブジェクト同士は類似するオブジェクト又はオブジェクト同士により関連付けられる ”という直観を実現できる。また、スコアは常に $0 \leqq Sim_l \leqq 1$ となる。

4. 実　　験

我々は我々の提案手法と SimRank の性能を実際の論文の引用関係データを用いて比較した。実験では本提案手法と SimRank を実際の著者間の論文の引用グラフに適用した。我々は著者間の引用グラフを計算機科学分野の論文目録サイト DBLP で公開されている文献データを用いて作成した。グラフのノードが論文の著者、リンクが引用となる。引用グラフは ER と EDBT のデータベースの国際会議で発表された論文の著者間の引用関係から成る。本実験はこの引用グラフの第一著者間の類似度を求めるために提案手法と SimRank を適用した。第一著者の総数は 1074、引用総数は 3199 である。

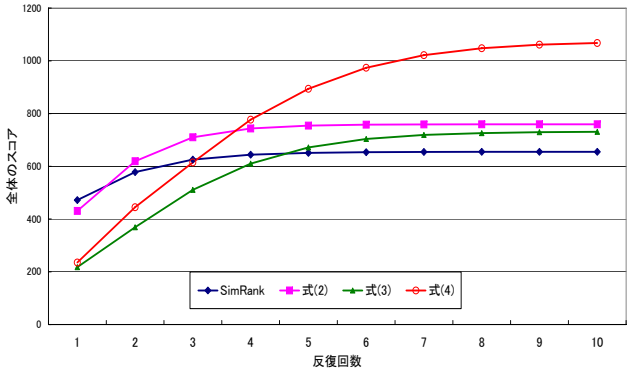

図 4　反復毎の全体のスコア遷移

本提案手法はノード間のリンクを考慮する式 (2) とセルフリンクを持つノードと他方のノードにインリンクで隣接するノードの対の類似度を加算する式 (3) と式 (2) と式 (3) を組み合わせた式 (4) の 3 つから成る。本実験ではこの 3 つの提案手法と SimRank の性能比較を行うものとする。

本提案手法と SimRank の性能を比較する際に、我々は次の質問に答えることに照準を絞った。

（1） 反復毎に全体のスコアがどのように変化したか？

（2） 全体のスコアがどのように変化したか？

本実験の反復回数は 10 である。同著者間の類似度は常に 1 になることは自明なので、同著者間の類似度は実験結果には含めない。著者間で対称となる類似度の重複も含めない。

4.1 反復毎の全体のスコアの変化

図 4 は本提案手法と SimRank の全体のスコアの反復毎の推移を表す。スコアの推移の検証は最も重要なポイントとなる。なぜなら、SimRank は不動点アルゴリズムで各スコアは必ずある値に収束するので、全体のスコアは必ずある値に収束しなければならない。もし、全体のスコアが収束せず発散すれば、そのようなアルゴリズムは類似度計算の手法としては使用できない。図 4 を見る限り、本提案手法の全体のスコアはすべて収束に向かっているので、本提案手法は SimRank の性質を保った機能拡張ということができる。本提案手法は反復回数が 1 のときにどれも SimRank の全体のスコアより低いが、これは SimRank の形式定義に含まれる分母の数より本提案手法の形式的定義に含まれる分母の数が大きいためにこのようになる。反復回数が

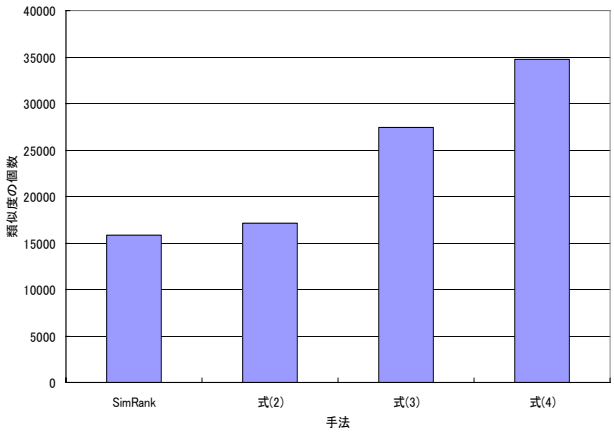

図 5　0 より大きい類似度の個数

1 のときに本提案手法の全体のスコアは SimRank の全体のスコアより低いが、ある一定回数の反復を行ったときに、本提案手法が SimRank の全体のスコアを上回っているため、従来の SimRank よりも類似度の伝播をより強く行っているということがわかる。

4.2　全体のスコアの変化

図 5 は本提案手法と SimRank においてすべての二つのノードのスコアが 0 より大きい類似度の個数を示している。図 5 のスコアは反復回数 10 のときのスコアを対象としている。二つのノードのスコアが 0 より大きくなる数は SimRank で 15828 個, 式 (2) で 17113 個, 式 (3) で 27442 個, 式 (4) で 34726 個となる。SimRank の問題点として類似度がうまく伝播しないスパース問題があることを第 3 章ですでに述べた。図 5 の式 (3) はインリンクで隣接するノードとスコアを求めるもう片方のノードの類似度を考慮に入れることにより、スパース問題を顕著に改善できることを示している。反復毎の全体のスコアの推移と同様に、図 5 から本提案手法が従来の SimRank よりも類似度の伝播をより強く行っているということがわかる。

本提案手法と SimRank の二つの比較より、本提案手法が類似度の伝搬という点において SimRank を改善したものであることがわかる。

5.　関 連 研 究

本稿はリンク構造を用いた類似度計算方法として SimRank について言及してきた。リンク構造を用いた類似度計算方法の中で最も有名な方法として co-citation [2] がある。co-citation は科学分野の論文の類似度を計算する際によく使われる手法である。co-citation の計算方法において、二つの論文 a と論文 b の間の類似度は論文 a と論文 b の両方を引用している論文の数を基に決められる。co-citation と SimRank の二つのノードに対する類似度の計算方法の違いは co-citation が隣接するノード、親ノード、によって類似度を計算するのに対して、SimRank は反復計算によってグラフ全体の類似度を計算する点にある。”類似するオブジェクト同士は類似するオブジェクトに関連付けられる ”という趣旨の基、類似しているノードが隣接していればいるほど類似度がより高くなる点で類似度の計算方法として SimRank は co-citation より優れている。Glen Jeh ら [1] は co-citation と SimRank の性能評価を現実のデータセットを基に行い、SimRank の方が類似度の計算方法として co-citation より性能が高いことが期待できる。本提案手法も類似しているノードが隣接していればいるほど類似度がより高くなるので、本提案もまた類似度の計算方法として co-citation より性能が高いといえる。

Zhenjiang Lin らは類似度を計算する際にリンク構造を用いたノードの重要度を考慮する PageSim [3] を提案した。PageSim はノードの重要度を PageRank [4] を用いて計算し、ノード間で PageRank を伝播させ、より重要度の高いノードと類似するノードを計算する仕組みである。Glen Jeh らは SimRank の計算でノードの重要度を考慮できる以下の方法を提案した。

$$Sim_P(a,b) = Sim(a,b) \times |I(a)|^P \qquad (5)$$

P はノードの重要度となり、インリンクで隣接するノードの重要度を SimRank の計算に反映させる。本提案もまたそれに従ってノードの重要度を加味できる。しかし、ノードの重要度のパラメーター P の設定は人間の手によって決められるため、PageSim のように重要度と類似度を統一的に扱うことができない。重要度と類似度を統一的に扱えるよう SimRank を改善することが今後の課題となる。また、引用解析における重要度と類似度を統一的に扱えるノイマンカーネル [5] がある。ノイマンカーネルは引用解析の際に co-citation と HITS [6] を組み合わせて計算することで重要度と類似度の計算を統一的に扱う。

SimRank の別の問題として計算量の問題があげられる。SimRank はすべてのノードの対の類似度を計算するために空間計算量が正方になり $O(n^2)$ となってしまう。SimRank を拡張している本提案もまた空間計算量が $O(n^2)$ となる。そこで Glen Jeh らは枝刈りによって遠方のノードを考慮せずに近隣のノードだけに類似度の伝播を絞ることで計算量を減らす方法を提案した。Glen Jeh らは SimRank のスキームを変えることで空間計算量の問題に取り組んだ。これに対して Daniel Fogaras らは最初から空間計算量を軽減したスキームで設計されたリンク構造を用いた類似度計算方法 PsimRank [7] を提案した。二つのランダムウォークが互いにグラフ上を同じゴールに向かっているかどうかを基に類似度計算する。二つの各ノードからノードのリンクを逆向きに辿るランダムウォークを放ち、二つのランダムウォークが最初に出会ったときに類似度を計算する。一度グラフを散策してしまえば、すべてのノードの対の類似度を計算できるために空間計算量は線形となり $O(n)$ となる。

Jimeng Sun ら [8] は SimRank で用いられているアイデア ”類似するオブジェクト同士は類似するオブジェクトに関連付けられる ”を二部グラフ上で別の方法で実現した。グラフのリンク構造を隣接行列に変換し、ランダムウォークを二部グラフ上で

再帰的に走らせて算出した隣接行列の固有値をノード間の類似度として計算した。

David Liben-Nowell ら [9] は隣接するリンク構造を用いた類似度計算使って、過去のグラフデータの類似度から現在のグラフデータのリンク予測の検証を試みた。本稿ではすべての時間の類似度を計算しており、過去、現在、未来等のスナップショットの類似度計算を対象とはしていない。そのため、本提案手法と SimRank がリンク予測を求める際にどれ程性能が変わってくるのかを検証することを次に試みたい。

6. 結　　論

我々は SimRank の類似度の伝播の改善を図るためにノード間のリンクの考慮とノード自身に対するセルフリンクを付加し隣接する単体のノードとの類似度の考慮を含む SimRank アルゴリズム提案した。さらに本提案と SimRank の類似度の反復によるスコアの推移の違いや全体のスコアの違いを実験により示した。その実験から本提案が SimRank の類似度の伝播をより高められることが確認できた。

7. 謝　　辞

本研究の一部は，文部科学省科学研究費補助金萌芽研究「AND-OR グラフを用いるデータモデルとその操作系，制約記述系に関する研究」（研究代表者：田島敬史，課題番号：18650021）によるものです。ここに記して謝意を表すものとします。

文　　献

[1] Glen Jeh and Jennifer Widom. SimRank: A measure of structural-context similarity. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 2002.

[2] Henry Small. Co-citation n the scientifc literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24: 265-269, 1973.

[3] Zhenjiang Lin, Michael R. Lyu, and Irwin King. PageSim: a novel link-based measure of web page aimilarity. WWW 2006: 1019-1020.

[4] Lawrence Page, Serge Brin, Rajeev Motwani, and Terry Winograd. The Pagerank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Libraries, 1998.

[5] Jaz Kandola, John Shawe-Taylor, and Nello Cristianini. Learning semantic similarity. In Advances in Neural Information Processing Systems (NIPS) 15, pages 657-664, 2002.

[6] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Proceeding of the nineth annual ACM-SIAM symposium on Discrete algorithms, 668-677, 1998.

[7] Daniel Fogaras and Balazs Racz. Scaling link-based similarity search. WWW 2005: 641-650.

[8] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. Neighborhood Formation and Anomaly Detection in Bipartite Graphs. ICDM 2005: 418-425.

[9] David Liben-Nowell and Jon M. Kleinberg. The link prediction problem for social networks. CIKM 2003: 556-559.