減衰流を用いた関係の解析

A Generalized Flow based Analysis of Relations

張 信鵬*1 浅野 泰仁*1 吉川 正俊*1
Xinpeng Zhang Yasuhito Asano Masatoshi Yoshikawa

*1京都大学
Kyoto University

Several methods have been proposed for measuring the strength of a relation on a massive information network whose vertices represent objects and whose edges represent relations among the objects. Most of the methods for measuring the strength use one of the following three concepts: distance, connectivity, or co-citation. We explain that using only one of these concepts is inadequate for measuring the strength of a relation. We propose a new method based on all the three concepts using a generalized maximum flow and a newly proposed structure, named a doubled network. A generalized maximum flow is a natural model for measuring distance and connectivity, although it could not be used for measuring cocitation until we propose the doubled network. Furthermore, our method mines objects constituting a relation. Based on the new method, we propose a "ENISHI" Discovery System using Wikipedia, which contains a huge number of relations among objects.

1. はじめに

関係には明示的な関係と暗黙的な関係の２種類が存在している．例えば,「友達」は前者の例であり,「友達の友達」は後者の例である．両方とも現実の世界で重要な役割がある．オブジェクト間の関係は頂点がオブジェクトを表し，枝がオブジェクト間の明示的な関係を表す情報ネットワークにおいて，暗黙的な関係の強さを測る研究が数多く行われてきた．既存の手法の多くは，距離，連結度，共引用のいずれかに基づいているが，それだけでは暗黙的な関係の強さを測るには不十分である．

距離に基づいた手法では，オブジェクト間の最短パスの長さで関係の強さを表現する．最短パスが短ければ短いほど，関係が強いとする．しかし，最短パスだけでは連結度を反映することができないため，関係の強さを測るには不十分である．例えば，図 1 に示すように，u_1 と v_1 間の距離と u_2 と v_2 間の距離は等しい．しかし，連結度が高い u_2 と v_2 間の関係が u_1 と v_1 間の関係より強いと考えるのが一般的である．Hubbell と Katz が提案した "cohesion"[Wasserman 94] は最短パスのみならず，オブジェクト間のすべてのパスを考慮する．長いパスに小さい重みが与えられ，オブジェクト u と v 間の関係の強さは，u と v 間のすべてのパスの重みの合計になる．しかし，"cohesion" には大きな問題が一つ存在する．多くの枝と接続しているような「人気オブジェクト」が u と v 間に存在する場合，u と v 間のパスが非常に多くなる．従って，人気オブジェクトが存在すると，オブジェクト間の関係が非常に強くなる．人気オブジェクトの存在と関係の強さとは本来独立であるはずなので，この性質は関係の強さを測るには不適切だと言える．中山らが提案した LFIBF[中山 07] と Koren らが提案した CFEC[Koren 06] は cohesion に基づいた方法である．基本的に，LFIBF と CFEC では，枝に $(0, 1]$ の重みが与えられ，パスの重みは，そのパス中の枝の重みの積をパス中の各ノードに接続する枝の数の積で割ったものとなる．そのため，"cohesion" と逆に，LFIBF と CFEC は人気オブジェクトを過小評価する性質を持っている．

連結度も関係の強さを測るために使われている．最大流は最小カットの容量に等しく，連結度を計算するに利用できる．しかし，最大流は距離を反映することができないため，関係の強さを測るには不十分である．例えば，図 1 において，u_2 と v_2 間の連結度と u_3 と v_3 間の連結度は等しいが，オブジェクト間の距離が違うため，関係の強さも違うと見なすのが自然である．

関係の強さを測るには，共引用の概念もよく用いられる．オブジェクト u と v がより多くのオブジェクトを共引用していれば，u と v の関係がより強いと見なす．しかし，共引用は向きが異なる二つの枝からなるパスに対応しており，向きが同じ枝からなるパスで表現される関係を測る目的には使えない．例えば，図 1 の u_4 と v_4 間の共引用は 0 であるが，一般的には u_4 と v_4 間の関係の強さが 0 とは見なさない．従って，共引用の概念だけに基づいた方法は不十分である．

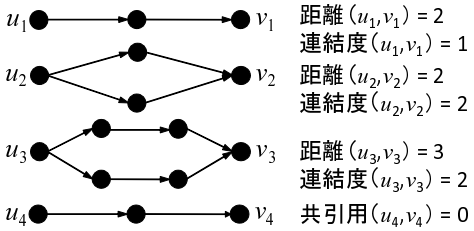

図 1: 関係の強さの例

筆者らは，距離，連結度，共引用の三つの概念全てに基づき，減衰流 (Generalized Flow) を用いて暗黙的な関係の強さを測る手法を提案した [Zhang 09a]．あるデータセットにおいて，始点オブジェクト s から終点オブジェクト t までの関係の強さを測るために，s と t を含む減衰ネットワークを構築する．その後，s から t までフローを送り，t に届いたフローの値で関係の強さを表す．提案手法では，減衰ネットワークのすべての枝の減衰率を 1 以下に設定するので，長いパスに沿って送られるフローが短いパスと比べて比較的小さくなりやすい．従って，短いパスの方がフローに大きく貢献することができ，s から t までの距離を測定できる．また，提案手法は二つのオブジェクト間の連結度を反映することができる．その理由

連絡先: 張 信鵬，京都大学情報学研究科，〒 606-8501 京都市左京区吉田本町 京都大学情報学研究科，075-753-9139，075-753-4970，xinpeng.zhang@db.soc.i.kyoto-u.ac.jp

は，古典的な最大流はネットワークの連結度を測るために用いられており，最大流の拡張である減衰流は，近似的に二つのオブジェクト間の連結度を表すことができるからである．更に，筆者らが提案した二重化ネットワークを利用することにより，互いに向きが異なる枝からなるパスで表現される共引用関係を測ることができる．減衰流では，フローが始点から終点へ送られるため，始点から終点に向かう方向と反対向きの枝がほとんど使われない．共引用関係を測るため，反対向きの枝も利用できるように，元の情報ネットワークに頂点と枝を加えることで「二重化ネットワーク」は構築される．本手法では，流量最大の減衰流に貢献したパス，つまりフローが大量に流れたパスを求めることで，関係に寄与したオブジェクトを発見することもできる．

さらに，本研究では，上で提案した減衰流を用いた関係の強さを測る手法を利用し，「縁」発見システムを提案した [Zhang 09b].「縁」発見システムは関係に寄与したオブジェクトを可視化及び分析することにより，オブジェクト間の関係と，オブジェクト間の関係に基づいたオブジェクトのランキングの理解を支援する．

本研究では，提案した減衰流を用いた関係の強さを測る手法の概要を説明し，「縁」発見システムの概要を紹介する．

2. 関係の解析

本章では，減衰流の概要を紹介する上で，減衰流と二重化ネットワークを用いた関係を解析する手法を説明する．

2.1 減衰流

点集合が V, 辺集合が E, 始点が $s \in V$, 終点が $t \in V$, 各辺 e の容量が $\mu(e) \geq 0$, 各辺の減衰率が $1 > \gamma(e) > 0$ であるような有向ネットワーク $G = (V, E, s, t, \mu, \gamma)$ が与えられたとする．フロー f の各辺 e での流量を $f(e) \geq 0$ で表し，f の値を t に流れ込むフローの総量と定義する．このとき，減衰流最大化問題とは，容量制約 $f(e) \leq \mu(e)$ をすべての辺 e で満たすような，最大の f を求める問題である．

図 2 は，流量最大の減衰流の例である. 始点仏教からアメリカ北部へ 1 単位のフローが流されている，即ち f(仏教, アメリカ北部) $= 1$ であるが，フローが北アメリカに到達した時にはその流量は μ(仏教, アメリカ北部) $= 0.8$ 倍となるため，0.8 単位のみ到達することになる. 同様に，パス (仏教，アメリカ北部，アメリカ) と (仏教，仏教大学院，カリフォルニア州，アメリカ) にそって，終点アメリカには 0.6 単位のみが到達する．

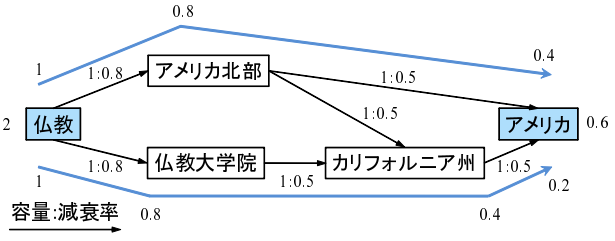

図 2: 流量最大の減衰流

減衰流問題では，長いパスに沿って送られるフローが短いパスと比べて比較的小さくなりやすい．従って，短いパスの方がフローに大きく貢献することができ，始点から終点までの距離を測定できる．また，減衰流は最大流の拡張であるため，近似的に二つのオブジェクト間の連結度を表すこともできる．本手法では，終点 t に届いたフローの値で関係の強さを表す．従って，オブジェクト s と t 間に独立した短いパスが多ければ多いほど，s と t 間の関係が強い．特に，流量最大の減衰流に貢献したパス，つまりフローが大量に流れたパスを求めることで，関係に寄与したオブジェクトを発見することもできる．共引用関係を測るために，二重化ネットワークを利用する．

2.2 二重化ネットワーク

減衰流では，フローが始点から終点へ送られるため，始点から終点に向かう方向と反対向きの枝がほとんど使われない．しかし，互いに向きが異なる枝からなるパスで表現される共引用関係を測るためには，反対向きの枝も利用する必要がある．反対向きの枝を利用するために，簡単な方法としては，元の減衰流問題のネットワークの任意の辺を反転させてできるすべての「部分反転ネットワーク」において，流量最大の減衰流を求めることである．反転された辺に，もともとその辺に与えられた減衰率と異なる減衰率 rev が与えられ，反転された辺の減衰率を「反転辺減衰率」と呼ぶ．ネットワーク G に対し，E の任意の部分集合 E_r 中の辺を反転させてできる部分反転ネットワーク $G_{E_r,rev}$ が $2^{|E|}$ 個存在している．図 3(b) は，図 3(a) のネットワーク G に対する二つの部分反転ネットワーク $G_{\{(s,u)\},rev}$ と $G_{\{(t,u)\},rev}$ を描いている．従って，G と rev に対するすべての部分反転ネットワークにおいて流量最大の減衰流を求めるには指数時間かかってしまう．この問題を多項式時間で求めるために，二重化ネットワークを提案する．図 3(c) は図 3(a) のネットワーク G に対する二重化ネットワークを描いている．

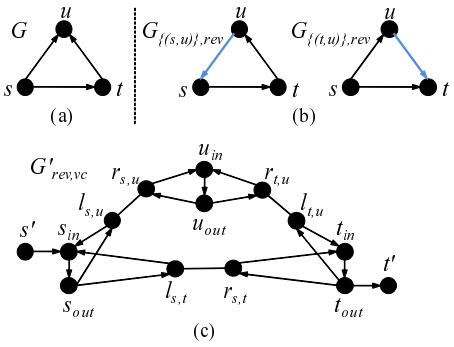

図 3: 二重化ネットワーク

定義 1 $G = (V, E, s, t, \mu, \gamma)$ を有向ネットワーク, $rev : E \to (0, 1]$ を G の各辺の反転辺減衰率とする.G の rev に対する二重化ネットワーク $G'_{rev} = (V', E', s', t', \mu', \gamma')$ を以下のように定義する．V' は G の各点に対応する「超頂点対」と各辺に対応する「制御点対」,「仮想ソース,「仮想デスティネーション」で構成される．

(1) G の点 $v \in V$ に対応する超頂点対は G'_{rev} の *2* つの点 v_{in}, v_{out} で表す．

(2) 辺 $(u, v) \in E$ に対応する制御点対は G'_{rev} の *2* つの点 $\ell_{u,v}, r_{u,v}$ で表す．

(3) 仮想始点を s' で表す．

(4) 仮想終点を t' で表す．

E' は以下の *5* 種類の辺で構成される．

(1) *2* 本の有向辺 $e_1 = (s', s_{in})$, $e_2 = (t_{out}, t')$. 容量は $\mu'(e_1) = \mu'(e_2) = \infty$ とし，減衰率は $\gamma'(e_1) = \gamma'(e_2) = 1$ とする．

(2) G の各点 $v \in V$ につき *1* 本の有向辺 $e = (v_{in}, v_{out})$. 容量は ∞ とし，減衰率は $\gamma'(e) = 1$ とする．

(3) G の各辺 $(u, v) \in E$ につき *1* 本の無向辺 $e = \{\ell_{u,v}, r_{u,v}\}$. 容量は $\mu'(e) = \mu(u, v)$ とし，$\gamma'(e) = 1$ とする．

(4) G の各辺 $(u,v) \in E$ につき 2 本の有向辺 $e_1 = (u_{out}, \ell_{u,v})$, $e_2 = (r_{u,v}, v_{in})$. 容量は $\mu'(e_1) = \mu'(e_2) = \mu(u,v)$ とし，減衰率は $\gamma'(e_1) = \gamma'(e_2) = \sqrt{\gamma(u,v)}$ とする．

(5) G の各辺 $(u,v) \in E$ につき 2 本の有向辺 $e_1 = (v_{out}, r_{u,v})$, $e_2 = (\ell_{u,v}, u_{in})$. 容量は $\mu'(e_1) = \mu'(e_2) = \mu(u,v)$ とし，減衰率は $\gamma'(e_1) = \gamma'(e_2) = \sqrt{rev(u,v)}$ とする．

元のネットワークの辺 (u,v) は，(3)–(5) の 5 本の辺に対応している．これらの辺の集合を (u,v) の「二重化辺集合」と呼び，$D(u,v)$ で表す．

図 4 は有向ネットワークの辺と，その二重化辺集合を描いている．なお，(u_{in}, u_{out}) と (v_{in}, v_{out}) の 2 本の辺も一緒に描いている．

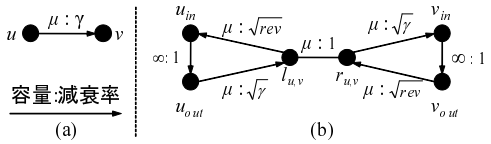

図 4: 辺 (u,v) とその二重化辺集合

次に述べるのが，本章の主たる定理である．

定理 1 E のすべての部分集合の族を $\{E_1, E_2, ..., E_{2^m}\}$ で表す．与えられた反転辺減衰率 rev と各部分集合 E_i $(1 \leq i \leq 2^m)$ に対する G の部分反転ネットワークを $G_{E_i,rev}$ とする．有向ネットワーク G に対する減衰流最大化問題の解 (流量最大の減衰流の流量) を $genmax(G)$ で表す．このとき，以下の等式が成立する．

$$\max_{1 \leq i \leq 2^m} genmax(G_{E_i,rev}) = genmax(G'_{rev}).$$

定理 1 を証明するために，二重化辺集合について「原初正則」と「反転正則」という二つの概念を導入する．

定義 2 $G'_{rev,vc}$ を二重化ネットワークとし，その上での任意の減衰流を f' とする．ある二重化辺集合 $D(u,v)$ において，その制御点対 $\ell_{u,v}$, $r_{u,v}$ で超過も不足もないときに，

- $f'(u_{out}, \ell_{u,v}) > 0 \wedge f'(\ell_{u,v}, r_{u,v}) > 0 \wedge f'(r_{u,v}, v_{in}) > 0 \wedge f'(v_{out}, r_{u,v}) = 0 \wedge f'(\ell_{u,v}, u_{in}) = 0$ であるならば，f' はその二重化辺集合の上で原初正則であるという．
- $f'(v_{out}, r_{u,v}) > 0 \wedge f'(r_{u,v}, \ell_{u,v}) > 0 \wedge f'(\ell_{u,v}, u_{in}) > 0 \wedge f'(u_{out}, \ell_{u,v}) = 0 \wedge f'(r_{u,v}, v_{in}) = 0$ であるならば，f' はその二重化辺集合の上で反転正則であるという．

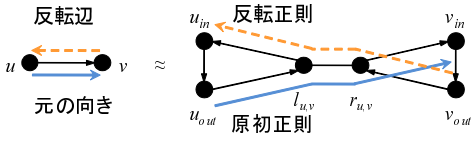

図 5: 原初正則なフローと反転正則なフロー

G の辺 (u,v) と，それに対応する $G'_{rev,vc}$ の二重化辺集合 $D(u,v)$ について，原初正則なフローを $D(u,v)$ に流すことは，(u,v) に (元の向きで) フローを流すことに対応する．反転正則なフローを $D(u,v)$ に流すことは，反転辺 (v,u) にフローを流すことに対応する．これらの例を図 5 に示す．以下に定理 1 の証明を省略する．

3. 「縁」発見システム

減衰流を用いた関係の強さを測る手法を利用し，オブジェクト間の関係と，オブジェクト間の関係に基づいたオブジェクトのランキングの理解を支援するための「縁」発見システム（図 6）を提案する．本システムでは，Wikipedia を利用し，関係の解析を行っている．Wikipedia において，各ページをオブジェクトと見なし，ページ間のリンクをオブジェクト間の明示的な関係と見なすことができる．オブジェクト s とオブジェクト t の関係を解析するために，Wikipedia のリンク情報を利用し，s と t を含める情報ネットワークを構築し，s から t までの流量最大の減衰流を求める．t に到達するフローの量が大きければ大きいほど，s と t 間の関係がより強い．フローに貢献したパス，つまりフローが大量に流れたパス中のノードが表すオブジェクトは，関係に寄与したオブジェクトとする．関係 r に寄与したオブジェクト o が関係 r または関係 r の終点オブジェクトに影響すると言う．

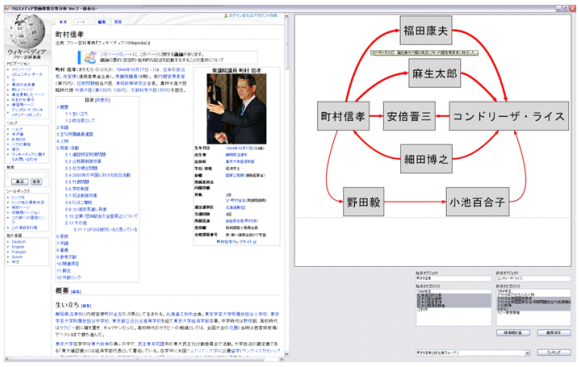

図 6: 「縁」発見システム

「縁」発見システムは三つの特徴を持つ．(1) 関係理解のために関係に寄与した重要なオブジェクトを可視化する．(2) 各関係に寄与したオブジェクトを分析することにより，オブジェクト間の関係に基づいたオブジェクトのランキングの分類を可能にしている．例えば，石油への関係に基づいた国家のランキングに対し，本手法で分類された二つのグループはそれぞれ「石油生産国」と「石油消費国」に対応していることが容易に理解できる．(3) 各関係に寄与したオブジェクトの差分を抽出することにより，あるオブジェクトが別のオブジェクトより特定のオブジェクトと強い関係を持つ理由を可視化する．

この三つの特徴について，それぞれ例を挙げて説明する．

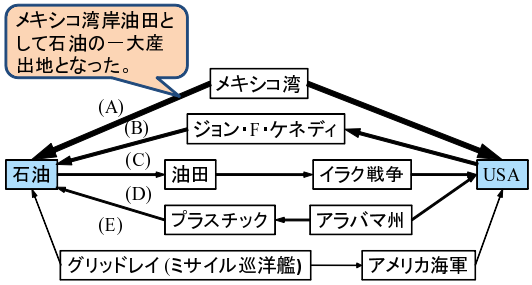

図 7: 関係の可視化

図 7 は特徴 1 の例であって，“石油” と “アメリカ” の関係を説明している．この関係の理解を支援するために，(1) Wikipedia のリンク情報から “石油” と “アメリカ” を含める情報ネットワークを構築する，(2) その情報ネットワークのダブルネットワークにおいて，“石油” から “アメリカ” へフローを送る，(3) “石油” から “アメリカ” まで流れる流量最大の減衰流に貢献し

たパスを可視化する．この例では，五つのパス (A)–(E) が描画されている．パスの太さがパスの重要性を表している．パスに沿ってより多くのフローが送られれば，パスがより重要であるとする。枝 $e(u,v)$ に対して，e の意味を説明するために u の Wikipeida ページから v を含める文が抽出される．そして，ポインタを枝の上に置けば，枝の意味を表す文が吹き出しで表示される．例えば，図 7 に示しているように，枝 (メキシコ湾 , 石油) の意味が吹き出しに表示されている．本システムを使えば，ユーザがたくさんの Wikipedia ページを読まなくても，関係を理解することが容易にできる．

「縁」発見システムでは，関係の理解の支援だけではなく，オブジェクト間の関係に基づいたオブジェクトのランキングの理解も支援している．オブジェクト間の関係に基づいたオブジェクトのランキングが様々な分野で使われている．例えば，複数の人物の活動を理解するために，特定の人との関係の強さにより複数の人物をランキングすることは有用である．図 8 は特徴 2 の例である．この例では，石油と各国の関係の強さによる国家のランキングを示している．さらに，石油と各国の関係に寄与したオブジェクトを分析することで，国家を分類している．k ($k \in Z, k > 0$) 個以上の共通な関係に寄与したオブジェクトに影響される終点オブジェクトを一つのグループ g_i に分類する．そして，任意のグループ g_i に属している終点オブジェクトに影響する共通な関係に寄与したオブジェクトの全てに影響されない終点オブジェクトを一つのグループ g_i' に分類する．この例では，ランキングされた十カ国が二つのグループ g_1 と g_1' に分類されている．図 8 に示されているテーブルの青いセルに書かれているグループ g_1 の国家は「石油価格」,「産油国」,「油田」,「石油資本」,「OPEC」,「OAPEC」に影響される．逆に，白いセルに書かれているグループ g_1' の国家はこれらのオブジェクトに影響されない．これらの六つのオブジェクトにより，ユーザはグループ g_1 の国家が産油国，グループ g_1' の国家が石油消費国だと理解できるだろう．従って，この特徴はランキングの理解の支援に有効だと言える．

順位	終点
1	日本
2	クウェート
3	アメリカ
4	ロシア
5	イギリス
6	サウジアラビア
7	中国
8	アラブ首長国連邦
9	カタール
10	イラン

始点: 石油
終点: 国家
オブジェクト:
"石油価格",
"産油国",
"油田",
"石油資本",
"OPEC",
"OAPEC".

以上のオブジェクトに影響される
以上のオブジェクトに影響されない

図 8: 関係の分類

特徴 3 として,「縁」発見システムでは関係に基づいたランキングの理由を可視化することで説明している．従来の関係の強さを測る手法または可視化する手法では，あるオブジェクトが他のオブジェクトより特定オブジェクトと強い関係を持つ理由を説明できない．そのため，既存の手法を利用したランキングでは，ユーザーがランキングを十分に理解することは困難である.「縁」発見システムでは，減衰流を用いた関係の強さを測る手法を用いることにより，ランキングをより理解しやすくする．図 9 は図 8 に示されているランキングについて,「クウェート」が「カタール」と「インドネシア」より高くランキングされた理由を説明している．この理由を説明するために，まずは「石油」とこの三か国のそれぞれの関係に寄与したオブジェクトから，以下のオブジェクトが抽出される： 「クウェート」だけに影響するオブジェクト， 「クウェート」と「カタール」に影響するオブジェクト， 「クウェート」と「カタール」と「インドネシア」に影響するオブジェクト．この三種類のオブジェクトが図 9 に示されたように異なる色の長方形に囲まれている．例えば，オレンジ色の長方形に囲まれている「アラビア石油株式会社」が「クウェート」だけに影響するオブジェクトである．それから，この三種類のオブジェクトを通して「石油」とこの三か国を連結しているパスが描かれる．減衰流を用いた関係の強さを測る手法では，独立した短いパスが多ければ多いほど，関係が強い．「石油」と「クウェート」の間に短いパスがこの三か国の中で一番多いため，ユーザ「石油」と「クウェート」の関係が強い理由が理解できると思われる．

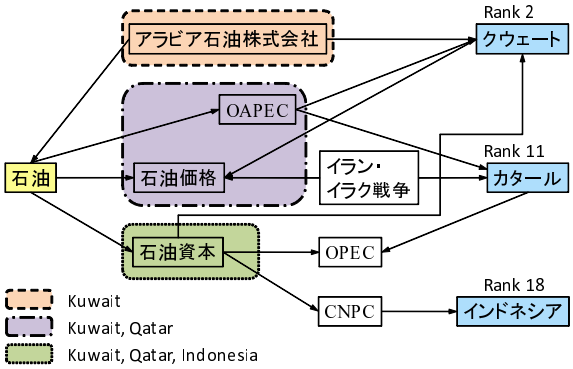

図 9: ランキングの解釈

4. まとめ

本研究では，減衰流を用いた関係の強さを測る手法を提案した．特に，本手法が関係に寄与したオブジェクトを発見することができる．それに，提案手法を利用し，関係の理解と，関係に基づいたランキングの理解の支援のための「縁」発見システムを提案した．最後に,「縁」発見システムの三つの特徴の事例を提示することで，それぞれの有用性を確認した．

5. 謝辞

本研究の一部は，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです．ここに記して謝意を表します．

参考文献

[Koren 06] Koren, Y., North, S. C., and Volinsky, C.: Measuring and extracting proximity in networks, in *Proc. of 12th ACM SIGKDD Conference* (2006)

[Wasserman 94] Wasserman, S. and Faust, K.: *Social Network Analysis: Methods and Application*, Cambridge University Press (1994)

[Zhang 09a] Zhang, X., Asano, Y., and Yoshikawa, M.: A Generalized Flow based Analysis of Implicit Relations: Measuring Strength through Mining Elucidatory Objects, in *Submitted* (2009)

[Zhang 09b] Zhang, X., Asano, Y., and Yoshikawa, M.: Visualized elucidations of ranking by exploiting object relations, in *Proc. of 3th DBRank* (2009)

[中山 07] 中山 浩太郎, 原 隆浩, 西尾 章次郎：Web 事典からのシソーラス辞書構築手法, 情報処理学会論文誌, Vol. 40, No. SIG 11, pp. 27–37 (2007)