

情報ネットワークにおける関係の抽出のための減衰流の計算の高速化

野島 裕輔[†] 浅野 泰仁[†] 吉川 正俊[†]

[†] 京都大学大学院情報学研究科 〒 606-8501 京都市左京区吉田本町 36-1

E-mail: [†]nojima@db.soc.i.kyoto-u.ac.jp, ^{††}{asano,yoshikawa}@i.kyoto-u.ac.jp

あらまし Wikipedia や DBLP における二つの概念の関係を、それらのグラフ上の減衰流を用いてモデル化することによって、その強さを求め、さらにその関係の仲立ちをする概念を抽出できるということが、先行研究によって示されている。減衰流は計算コストが大きいので、先行研究では減衰流に含まれるパスの長さを制限する手法を用いて、減衰流を近似的に計算していた。しかし、この手法では、対象となるグラフのスモールワールド性により、パスの長さの上限を大きくすると、急激に計算時間が増えるという問題がある。我々は、辺のサンプリングによる索引と、ヒューリスティックを組み合わせた、減衰流に対する新たな近似的高速化手法を提案する。

キーワード 減衰流, ネットワークフロー, ネットワーク, 関係

Faster Calculation of Generalized Maximum Flow for Relationship Extraction in Information Networks

Yusuke NOJIMA[†], Yasuhito ASANO[†], and Masatoshi YOSHIKAWA[†]

[†] Graduate School of Informatics, Kyoto University

Yoshida-honmachi 36-1, Sakyo-ku, Kyoto 606-8501 Japan

E-mail: [†]nojima@db.soc.i.kyoto-u.ac.jp, ^{††}{asano,yoshikawa}@i.kyoto-u.ac.jp

Abstract Preceding studies show that generalized maximum flow model is useful for extracting relationships and the bridging concepts between two concepts in the information networks such as the Wikipedia and the DBLP. Since the calculation of generalized maximum flow is very expensive, they used the approximative approach where lengths of paths in the flow are restricted to three. However, when a target network has the small world property, short paths can easily cover the entire network. Therefore, we propose the novel approximative method for generalized maximum flow utilizing the edge sampling indexing and heuristics.

Key words Generalized Maximum Flow, Network Flow, Network, Relationship

1. はじめに

関係には、明示的關係と暗黙的關係という二つの種類がある。例えば、ウェブページ間のリンクを明示的關係としたとき、あるウェブページと、そのウェブページからリンクを二回辿って辿りつけるウェブページには、暗黙的關係があるといえる。また、あるウェブページと別のウェブページが、同じウェブページへのリンクを持っていたとすると、この二つのウェブページには暗黙的關係があるといえる。このように、明示的關係が与えられたとき、関係の伝播や共参照などを考えることで、暗黙的關係を発見することができる。

概念間の関係は、一つの頂点が一つ概念を表し、一つの辺が一つの明示的關係を表す情報ネットワークを用いてモデル化することができる。この情報ネットワークに対して、暗黙的關係の強さを測るための手法が数多く提案されてきた。

第一に、距離に基づく手法が挙げられる。この手法では、二つの頂点の距離を暗黙的關係の強さとして採用する。しかし、距離が同じでも暗黙的關係の強さが異なると考えられる場合がある。図 1 において、 u_1-v_1 の距離と、 u_2-v_2 の距離は等しいが、 u_2-v_2 の方がより多くのパスで結ばれているため、より関係が強いとみなすのが自然である。

第二に、連結度に基づく手法が挙げられる。二つの頂点の連結度を暗黙的關係の強さとして採用する。しかし、二つの頂点の連結度は、二つの頂点間の距離を反映していないため、関係の強さを測る指標として問題がある。図 1 において、 u_2-v_2 の連結度と、 u_3-v_3 の連結度は等しいが、 u_2-v_2 のほうが距離が近いので、より関係が強いとみなすのが自然である。

第三に、共参照に基づく手法があげられる。この手法は、二つの頂点の隣接頂点集合の共通部分が大きいとき、その二つの頂点は強く関係しているとみなす。図 1 において、 u_4 と v_4 は

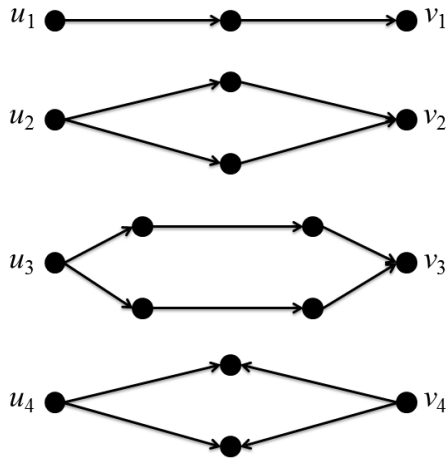


図1 関係の例

同じ二つの頂点を指しているため、共参照は2である考えられる。しかし、共参照では、距離や連結度を考慮できないため、問題がある。

Zhang [1] らは、これら三つの手法全ての性質を併せ持つ、減衰流を用いた手法を提案した。減衰流とは、古典的な最大流の拡張であり、各辺が減衰率と呼ばれる値を持つ。減衰流において、長いパスは大きく減衰し、逆に短いパスはあまり減衰しない。よって、減衰流を使うことで、二つの頂点間の距離を、関係の強さに反映することができる。また、減衰流は最大流と似た性質を持っているため、二つの頂点間の連結度を、関係の強さに反映することができる。さらに、Zhang 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、減衰流を用いて共参照関係を測定することができる。

しかし、減衰流は、多くの望ましい性質を持つ一方で、計算に非常に時間がかかるという欠点がある。頂点数 n 、辺数 m のネットワークに対し、現在知られているアルゴリズムでは、最適解を計算するのに $O(m^2(m+n))$ の時間がかかる。したがって、Wikipedia や DBLP のような巨大な情報ネットワークに対して、減衰流問題の最適解を計算することは現実的ではない。

Zhang らは、減衰流に含まれるパスの長さを3以下に制限することによって、Wikipedia に対して減衰流を近似的に計算していた。しかし、パスの長さを3以下に制限した場合、距離が4以上あるような頂点間の関係を計算することができない。また、この近似手法では、時間と精度のトレードオフを微調整することができない。なぜなら、パスの長さの上限を少しでも緩和すると、爆発的に利用可能な頂点数が増加し、計算時間が急激に増大するためである。

そこで我々は、辺のサンプリングによる索引と、いくつかのヒューリスティックを組み合わせた、減衰流に対する新たな近似的高速化手法を提案する。

我々の手法では、まず、減衰流の値をできるだけ保存したまま、与えられたネットワークの圧縮表現を作成する。圧縮されたネットワークに対して、減衰流を計算するアルゴリズムを適

用することによって、高速に減衰流の近似値を得る。

2. 関連研究

Aggarwal [2] らは、辺のサンプリングと頂点のサンプリングを行うことで、巨大なグラフに対して高速に連結度を計算する手法を提案した。また彼らは、連結度の小さい頂点对に対して、辺のサンプリングによる近似が非常に良い結果を返すことを理論と実験の両面から実証した。さらに、連結度が10を超えるような場合でも、辺のサンプリングが比較的良好な性能を示すことを実験的に確かめた。我々の手法は、Aggarwal らの手法を減衰流に対して拡張したものである。

また、減衰流を計算するアルゴリズムに関して、様々な研究がなされている。Onaga [3] は、減衰流が最適であるための必要十分条件は、増加パスと一般化増加パス(減衰閉路からの終点への増加パス)が残余グラフに存在しないことであることを示した。この必要十分条件は、減衰流を計算する上での基礎となっている。さらに Onaga は、減衰流を計算する最初のアルゴリズムを提案した。Goldberg [4] らは、減衰流問題を解く最初の多項式時間アルゴリズムを構築した。Tardos と Wayne [5], [6] は gain-scaling と呼ばれるテクニックを使い Goldberg らのアルゴリズムを改良した。Tardos らのアルゴリズムの時間計算量は現在知られているアルゴリズムの中で最良である。Wayne [7] らは、 m を辺数としたとき $O(m^2)$ の時間で減衰流の最大値を近似するアルゴリズムを提案した。Fleischer [8] らは、 $O(\varepsilon^{-2}m^2)$ の時間で減衰流を ε -近似するアルゴリズムを提案した。

3. 減衰流問題の定義

この節では、減衰流問題を定義する。

[定義 3.1] 減衰ネットワークとは、有向グラフ $G = (V, E)$ 、容量関数 $c: E \rightarrow \mathbb{R}_+$ 、減衰率関数 $\gamma: E \rightarrow \mathbb{R}_+$ 、始点 $s \in V$ 、終点 $t \in V$ の組である。ただし、任意の辺 $e \in E$ に対して、 $0 < \gamma(e) \leq 1$ である。

[定義 3.2] 減衰流とは、関数 $f: E \rightarrow \mathbb{R}_+$ であって、以下の2条件を満たすものである。

- (容量制約) 全ての辺 $e \in E$ に対して、 $f(e) \leq c(e)$ 。
- (フロー制約) 始点、終点以外の全ての頂点 $v \in V \setminus \{s, t\}$ に対して、 $\sum_{e \in \delta^-(v)} \gamma(e)f(e) - \sum_{e \in \delta^+(v)} f(e) = 0$ 。ただし、 $\delta^-(v)$ は、終点が v である辺の集合であり、 $\delta^+(v)$ は、始点が v である辺の集合である。

[定義 3.3] 減衰流 f の値とは、 f による終点 t の正味流入量、すなわち $\sum_{e \in \delta^-(t)} \gamma(e)f(e) - \sum_{e \in \delta^+(t)} f(e)$ である。

[定義 3.4] 減衰流問題とは、与えられた減衰ネットワークに対して、最大の値を持つ減衰流を求める問題である。

4. 近似的高速化手法

提案手法では、減衰流を高速に計算するために、辺のサンプリングを用いてグラフを圧縮する。まず、図2のように、グラフの各辺を一定の確率 p でサンプルする。次に、サンプルされた辺以外の辺を無視し、グラフを強連結成分分解または辺の向きを無視した連結成分分解を行う。最後に、図3のように、各

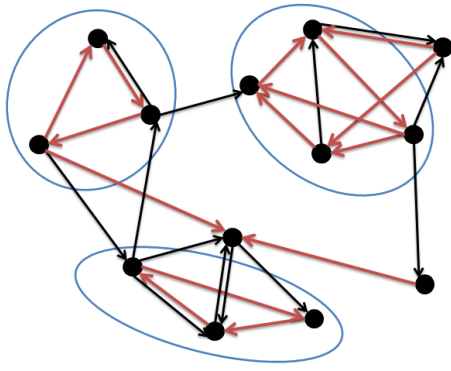


図2 サンプリングの例．色の付いた辺がサンプルされた辺

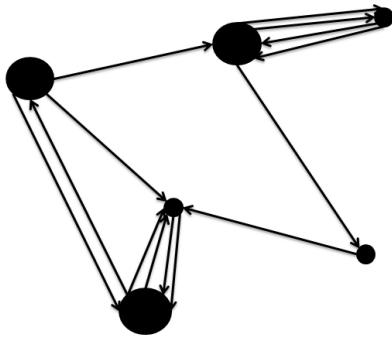


図3 図2の強連結成分を縮約し，圧縮グラフを作成した図

連結成分を1点に縮約する．この圧縮グラフを使って，始点 s から終点 t への減衰流の値を，始点の縮約先 s' から終点の縮約先 t' への減衰流の値で近似する．

この手法では，サンプリング確率 p を上手く設定することで，圧縮グラフの頂点数を柔軟に調整することができる．傾向として，サンプリング確率 p を大きくすると，より多くの頂点が1つの頂点に縮約されやすくなるため，圧縮グラフの頂点数は少なくなり，サンプリング確率 p を小さくすると，圧縮グラフの頂点数は多くなる．よって，サンプリング確率 p で二分探索をすることによって，望む頂点数に近い頂点数を持つ圧縮グラフを作ることができる．

この方法でグラフ G を圧縮して作った任意の圧縮グラフ G_C に対して， $(G$ における減衰流の最大値) $\leq (G_C$ における減衰流の最大値) が成り立つ．頂点を縮約することによって，圧縮前では辺が存在しなかった頂点同士が連結される可能性がある．このため，圧縮により，減衰流が増加する可能性がある．逆に，もともと連結されていた頂点が圧縮により非連結になるということはない．よって，圧縮により，減衰流が減少することはない．

より詳しく圧縮グラフにおける減衰流の値を評価するために，以下の定理を示す．

[定理 4.1] 減衰流 f が最大であるとき，グラフから $c(e) - f(e) = 0$ である辺 $e \in E$ (飽和辺と呼ぶことにする) を全て取り除くと，

始点 s と終点 t が分離される．(始点から終点へのパスが存在しなくなる)

飽和辺でない辺のみを用いた始点 s から終点 t までのパスが存在すると仮定すると，そのパスにそって水を流し，減衰流の値を増加させることができる．これは， f が最大であることに反する．よって，そのようなパスは存在しない．□

[定義 4.1] グラフ $G = (V, E)$ から，最大減衰流 f による飽和辺をすべて取り除き，さらに全ての辺の向きを逆にしたグラフを $G'_f = (V, E')$ とする． G'_f において，終点 t から辿れる頂点の集合を V_B ，そうでない頂点の集合を V_A とする． V_A と V_B にまたがる飽和辺の集合 $C_f = \{e = (u, v) \in E \mid u \in V_A, v \in V_B, c(e) - f(e) = 0\}$ を G の f による飽和カットと呼ぶことにする．

ある最大減衰流 f による飽和カット C_f によってグラフ G が2つのグラフ G_A と G_B に分割されたとする．ただし， G_A は始点 s を含み， G_B は終点 t を含むものとする．このグラフ G を辺サンプリングによって圧縮したとき，減衰流がどう変化するか考える．

まず，飽和カット C_f に含まれる辺がいずれも辺サンプリングによってサンプリングされなかった場合を考える．このとき， G_A に含まれた辺がどのようにサンプリングされたとしても， C_f を流れる減衰流の値は変化しない．なぜなら，辺サンプリングによる圧縮は水の流量を増加させる方向に作用するが， C_f が既に飽和しているため， G_A から G_B にこれ以上水が流れることはないからである．一方， G_B の辺サンプリングは減衰流の値に影響を及ぼす．辺サンプリングによる圧縮により，減衰流に含まれるパスの長さが減少した場合，減衰流の値が増加する可能性がある．また，飽和カット C_f に含まれる辺がサンプリングされた場合， G_A から G_B に流入する水の量が増加し，減衰流の値が増加する可能性がある．

まとめると，圧縮による減衰流の値の変化は以下の二つの要因で引き起こされる．

(1) 飽和カット C_f に含まれる辺がサンプリングされ，減衰流のボトルネックが解消される．

(2) G_B に含まれる辺がサンプリングされ，減衰流内のパスの長さが減少する．

(1) に関しては，一つのグラフに対して，圧縮グラフを複数個作成し，各圧縮グラフに対して減衰流を計算し，それらの最小値を取ることで対処する．なぜなら，飽和カットに含まれる辺が一つでもサンプリングされる確率は $1 - (1 - p)^{|C_f|}$ であり， L 個の圧縮グラフ全てで飽和カットがサンプリングされる確率は $(1 - (1 - p)^{|C_f|})^L$ であるため，(1) が起こる確率は， L を大きくすれば指数関数的に0に近づくからである．

(2) に関しては，縮約をする際に，始点 s から同じ距離にある頂点同士でのみ縮約をすることで対処する．これにより，圧縮により始点からの距離が小さくなるということを防ぐことができる．よって，このヒューリスティックにより，縮約によって減衰流内のパスの長さが短くなることをある程度防ぐことができる．

表1 実験結果

アルゴリズム	サンプル率	計算時間	近似解 ベースライン解
ε -近似 (ベースライン)	-	約 30 時間	1.00
強連結成分圧縮	0.5	約 8 分	2.37
強連結成分圧縮	0.9	約 8 分	2.37
連結成分圧縮	0.05	約 19 分	1.78
連結成分圧縮	0.1	約 3 分	2.37
強連結成分距離別圧縮	1.0	約 3 分	1.78
連結成分距離別圧縮	0.05	約 4 分	1.78
連結成分距離別圧縮	0.5	23 秒	1.78

5. 実験

日本語 Wikipedia を入力グラフとして使用した．このグラフは 2012 年 1 月 21 日の時点における日本語 Wikipedia のダンプファイルから得たものである．グラフは，Wikipedia の記事を頂点とし，記事間のリンクを辺とした．頂点数は 2,087,085 であり，辺数は 63,230,008 であった．全ての辺の容量は 1 とし，減衰率は 0.75 とした．また，始点と終点はランダムに 1 対を選択し，それを全てのアルゴリズムで使用した．実験には Core i5-2500K, メモリ 16GB を搭載したコンピュータを用いた．

最大減衰流を計算するアルゴリズムとしては，Fleisher-Wayne のアルゴリズムを利用した．このアルゴリズムは，最大減衰流問題に対する ε -近似アルゴリズムであり， $O(\varepsilon^{-2}m^2)$ の時間で実行可能である．今回の実験では $\varepsilon = 0.2$ とした．

圧縮グラフは 4 つ作成し，4 つの圧縮グラフに対して並列に最大減衰流を計算し，その最小値を近似解として採用した．計算は全て 1 台の計算機で行った．

実験結果は表 1 のようになった．表における $\frac{\text{近似解}}{\text{ベースライン解}}$ の値は，ベースラインの出した解と，各アルゴリズムが出した解の比であり，1 にできるだけ近い値であることが望ましい．ベースラインが ε -近似アルゴリズムを使用しているため，ベースライン解は厳密解でないことに注意されたい．

実験結果から，約 30 時間かかっていた計算が 15 秒から 19 分程度に短縮されており，グラフの圧縮により減衰流の計算が高速化されることが分かった．しかし，単純に強連結成分分解や連結成分分解をした場合，解の精度が非常に悪く，上手く減衰流を近似することができなかった．始点からの距離が等しい頂点のみを縮約するようにアルゴリズムを修正した場合，修正前よりも近似の精度が向上した．しかし，サンプル率を小さくしてもベースライン解の 1.75 倍よりも良い解が出力されなかった．

距離別の縮約は始点からの最短距離を圧縮後も保存することによって近似の精度を上げるヒューリスティックだが，実験結果は，最短パスだけではなく，その他のパスの長さを保つことが，精度向上のために必要であることを示唆している．パスの長さを保つためのヒューリスティックとして，例えば次のようなものが挙げられる．

- 各連結成分 C について， C に含まれない頂点との間に辺を持つような頂点を縮約の対象から外す．

- 始点や終点の近くの頂点を縮約の対象から外す．

今後，このようなヒューリスティックについて実験を行い，精度と速度を評価していきたい．また，より多くの始点-終点の対に対して精度を評価していきたい．

6. まとめと今後の課題

グラフを圧縮することで，巨大なグラフに対して減衰流を高速に計算する手法を提案し，またその性能を実験により評価した．グラフの圧縮は，まず辺のサンプリングを行い，サンプルされた辺のみを使用してグラフの頂点を強連結成分分解または連結成分分解し，各連結成分を 1 つの頂点に縮約することで行う．また，圧縮による減衰流の値の変化の原因は 2 種類に分類されることを議論した．

実験結果から，減衰流の計算は大幅に高速化されたが，精度の面ではまだ課題があることが分かった．特に，圧縮によるパスの長さの減少に対処するためのさらなる工夫が必要であることが分かった．よって，始点からの距離が等しい頂点同士のみを縮約するという現行のヒューリスティックの他にも，縮約する頂点を上手く選ぶためのヒューリスティックを考案・評価していきたい．

また，今回の実験による近似精度の測定は，1 対の始点-終点について行ったものであり，精度上の問題を発見するためには役にたったが，近似精度を定量的に評価するためには，試行回数が少なすぎるという問題があった．Wikipedia においてベースラインを十分な回数実行するのは，計算時間的に現実的でない．そのため，Wikipedia からサンプリング等を行うことで小さな部分グラフを作り，そのグラフに対して精度の測定を行うなどの方法で実験を行いたい．

また，提案手法では，減衰流の値を知ることは可能だが，減衰流そのものの形を直接知ることはできない．なぜなら，提案手法で得られるのは圧縮グラフにおける減衰流であるからだ．今後は，圧縮グラフ上の減衰流から元のグラフにおける減衰流を高速に求めるための手法についても検討していきたい．

文献

- [1] Xinpeng Zhang, Yasuhiro Asano, and Masatoshi Yoshikawa. A generalized flow based method for analysis of implicit relationships on wikipedia. *IEEE Transactions on Knowledge and Data Engineering*, 10 Nov. 2011.
- [2] Charu Aggarwal, Yan Xie, and Philip S Yu. Gconnect : A connectivity index for massive disk-resident graphs. *VLDB Journal*, Vol. 2, No. 1, pp. 862–873, 2009.
- [3] Kenji Onaga. Dynamic programming of optimum flows in lossy communication nets. *IEEE Trans. Circuit Theory*, pp. 308–327, 1966.
- [4] Andrew V. Goldberg, Serge A. Plotkin, and Éva Tardos. Combinatorial algorithms for the generalized circulation problem. In *FOCS*, pp. 432–443, 1988.
- [5] Éva Tardos and Kevin D. Wayne. Simple generalized maximum flow algorithms. In *IPCO*, pp. 310–324, 1998.
- [6] Kevin D. Wayne. *Generalized Maximum Flow Algorithms*. PhD thesis, Graduate School of Cornell University, 1999.
- [7] Kevin D. Wayne and Lisa Fleischer. Faster approximation algorithms for generalized flow. In *SODA*, pp. 981–982, 1999.
- [8] Lisa K Fleischer and Kevin D Wayne. Fast and simple approximation schemes for generalized flow. *Mathematical Programming*, Vol. 91, No. 2, pp. 215–238, 2002.