2009－MPS－73 (47)
2009／3／6

大規模ネットワークに対する効率的な信頼度算出法について

亀山周明[†] 平野未来[†] 白山晋[††]

大規模なネットワーク型データに対する情報信頼度を，できるだけ少ない人的コストで定量的に算出する手法を提案する．また，提案手法をブログデータに適用する．

An Efficient Approach for Estimating Trust of Information on the Large-scale Network

SHUMEI KAMEYAMA[†] MIKU HIRANO[†] SUSUMU SHIRAYAMA[††]

We propose an efficient approach for quantitatively estimating trust of information on the large-scale network with a little human effort. The proposed method is applied to construct a trust map on a weblog network.

1. はじめに

近年，インターネットは，Web2.0 という形態に移行し，Web 上での能動的な情報発信が急増し，発信手段としてのブログや SNS が急速に発展している．また，このような仕組みは会社組織の中でも情報共有や組織の活性化のために利用されている[1]．しかしながら，そうした情報には粗悪なものが混在し，利用者の情報選別に対する負荷を大きくしている．このため，効率的，かつ効果的な情報選別の手法が求められている．

情報選別の一つの方法として信頼度算出にもとづくものがある．その方法は，自然言語処理を主体とする意味内容を用いるものと，ネットワーク構造を利用したリンク解析にもとづくものに大別される．また，近年の信頼度算出の研究分野ではスパム発見が重要なテーマであり，信頼度算出の鍵にもなっている．スパムの発見は人手と機械処理に分けられる．さらに，機械処理によるものは，自然言語処理を用いるもの[2,3]とネットワーク構造を用いるもの[4,5]に大別される．

前者には，他言語に対応できない，新語の扱いが難しい，無意味な単語の羅列である WordSalad を検出するのが難しい，といった問題点がある．

後者の主なアプローチは２つである．１つは，２部グラフ構造などの特徴的なスパムのパターンを抽出するものである[4]．この方法には，２部グラフ内の個々のページの信頼度を測定できないことや，特徴的な構造を有さないスパムを発見できないという問題がある．もう１つは，スパムの評価値を定め，ネットワークを介してその評価値を他のページへ伝播させ，スパムか否かを判断するものである．この方法は，Web 上の信頼度の算出そのものに利用されている．代表的なものに Gyongyi らによって提案された TrustRank がある[5]．TrustRank は，有識者などが Web ページの一部を精査し（seed：シードと呼ぶ），信頼度を数値（score：スコアと呼ぶ）として与え，信頼度を Web 全体に伝播させるというものである．この方法には，シードの与え方と，シードから距離が遠いページの信頼度算出の精度が低下するという問題がある．ネットワークが大規模になると，この問題のために多くのページで信頼度の算出ができなくなる．

このように，意味的内容にもとづくものには自然言語の解釈にともなう困難が内在し，ネットワーク構造を利用したものには推定精度の問題がある．また，両者ともにデータの大規模性にともなう問題がある．

本研究では，これらの問題を解決し，Web 上の情報の信頼度をできるだけ少ない人的コストで定量的に評価することを目的とする．

2. 提案手法

ノード集合をV，エッジ集合をEとし，ネットワークを$G=(V,E)$で表わされる集合とする．要素であるノードをv，あるいは識別子を付けて示す（例えば，v_i）．場合によっては識別子のみでノードを示す（例えば，ノードi）．エッジはeで表す．ノードの総数をn，エッジの総数をmとする．また，隣接行列を$\mathbf{A}$，その成分をa_{ij}とし，

$$a_{ij}=\begin{cases}1 & i\text{から}j\text{へのエッジがあるとき}\\ 0 & i\text{と}j\text{にエッジがないとき}\end{cases} \quad (1)$$

とする．

提案手法は，TrustRank[5]と BadRank[6]を改良し，併用することによって信頼度の算出を行うものである．

はじめに，TrustRank と BadRank に，コミュニティグラフ[7]を導入し，それらの問題点を解決する．

2.1 コミュニティ抽出

リンク密度の観点によるものと，構造の類似性によるものによってコミュニティを抽出する．両者には有向グラフに対するものが提案されているが，計算負荷が大きいなどの問題点も多い．そこで，本稿では無向化したネットワークからコミュニティを抽出し，抽出後に有向化する．

また，前者として，Clauset ら[8]と Blondel ら[9]の方法を利用する．これらの方法で抽出されたコミュニティ内部のノードは密に結合しており，それらのノードは関連性が高く，信頼度においても似通っていると考えられる．

後者としては類似度行列を用いた方法（SMC 法）[10]を用いる．この方法では，他のノードとの関係性が類似するものをグルーピングするもので，ネットワークに含まれる特徴的な構造が抽出できる．

ここで，コミュニティの識別子をαとする．あるコミュニティをG_αとすると，$G_\alpha \subseteq G$である．コミュニティの総数をNとすると，$V=\bigcup_{\alpha=1}^{N} V_\alpha$となる．また，２つのコミュニティ$\alpha$と$\beta$に対して，$V_\alpha \cap V_\beta = \emptyset$である．

*† 東京大学大学院工学系研究科環境海洋工学専攻
Department of Environmental and Ocean Engineering, School of Engineering, The University of Tokyo
†† 東京大学人工物工学センター
Research into Artifacts, Center for Engineering , The University of Tokyo

2.2 コミュニティグラフの作成

コミュニティをノードとし，コミュニティ間に存在するリンク数を重みとしたリンクをもつコミュニティグラフ[7]を作成する．

コミュニティαを，v_αで示す．コミュニティグラフに対する重み付き隣接行列を$\mathbf{A_C}$とし，その成分を$a^{\mathrm{C}}_{\alpha\beta}$とし，

$$a^{\mathrm{C}}_{\alpha\beta}=\begin{cases} l & \alpha\text{から}\beta\text{へ}l\text{本のエッジがあるとき} \\ 0 & \alpha\text{と}\beta\text{にエッジがないとき} \end{cases} \quad (2)$$

とする．Fig.2 にコミュニティグラフの一例を示す．

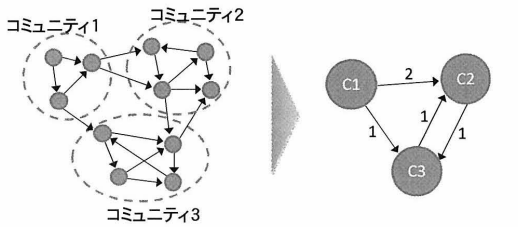


Fig. 2 コミュニティグラフの作成

2.3 コミュニティグラフに対する信頼度算出

TrustRank と BadRank を，コミュニティグラフに対して算出する．

コミュニティに対する TrustRank を$\mathbf{t_C}$とし，その成分をt^{C}_{α}とする．$\mathbf{A_C}$から遷移行列$\mathbf{T_C}$を求め，

$$\mathbf{t_C}=\omega\mathbf{T_C}\mathbf{t_C}+(1-\omega)\mathbf{d_C} \quad (3)$$

によって，TrustRank を算出する．ここで，ωは減衰ファクター，$\mathbf{d_C}$は特定ページへの加算項を表す．

また，BadRank を$\tilde{\mathbf{t}}_{\mathbf{C}}$とし次式で求める．

$$\tilde{\mathbf{t}}_{\mathbf{C}}=\omega\mathbf{U_C}\tilde{\mathbf{t}}_{\mathbf{C}}+(1-\omega)\tilde{\mathbf{d}}_{\mathbf{C}} \quad (4)$$

ここで，$\mathbf{U_C}$は逆遷移行列，$\tilde{\mathbf{d}}_{\mathbf{C}}$は特定ページへの加算項を表す．また，$\tilde{t}^{\mathrm{C}}_{\alpha}$をその成分とする．

第１に，Gyongyi らに従い，Inverse PageRank によってシードとなるコミュニティを選ぶ[5]．次章で示す実験では，10 個のコミュニティをシードとした．

第２に，そのシードを精査することによって信頼度を数値（score：スコア）として与える．具体的には，シードとしたコミュニティからノードを適当に選び，それらを精査することによってシードのコミュニティの評価とした．このとき，SMC 法を併用することによりシード精査の手間を軽減している．

第３に，コミュニティグラフに対して，式(3)と(4)によって，TrustRank の値と BadRank の値を伝播させる．

2.4 コミュニティ内への伝播

コミュニティ毎に得られた TrustRank と BadRank を，コミュニティ内に伝播させる．

元のネットワークに対する TrustRank を$\mathbf{t}$とし，その成分をt_iとする．$\mathbf{A}$から遷移行列$\mathbf{T}$を求め，

$$\mathbf{t}=\omega\mathbf{T}\mathbf{t}+(1-\omega)\mathbf{d} \quad (5)$$

によって，TrustRank を算出する．同様に BadRank を$\tilde{\mathbf{t}}$，逆遷移行列を$\mathbf{U}$とし，BadRank を算出する．

$$\tilde{\mathbf{t}}=\omega\mathbf{U}\tilde{\mathbf{t}}+(1-\omega)\tilde{\mathbf{d}} \quad (6)$$

コミュニティに対して得られた$\mathbf{t_C}$と$\tilde{\mathbf{t}}_{\mathbf{C}}$を，そのコミュニティに属すノードの初期値として，元のネットワークに式(5)と(6)を適用することによってスコアを求める．

2.5 アルゴリズムの流れ

Fig.3 はアルゴリズムの流れを図示したものである．Fig.3 の左上図の，点線で囲まれたものがコミュニティである．このネットワークに対して求めたコミュニティグラフが右上図である（図中１）．シードとして赤い部分を選択し精査する（図中２）．

左下図はコミュニティグラフに対して，信頼度を求めたものである（図中３）．右下図は、左下図で求めたスコアを初期値として（図中４），ネットワーク全体に対して，信頼度を求めたものである（図中５）．それぞれのノードに異なる信頼度が求められている．

コミュニティグラフを用いることにより，元のネットワークの特徴をある程度保ったままで，ネットワークのサイズを小さくできる．これによりシードからの距離を短縮できるので，伝播の際に発生する減衰の問題も緩和できる．また，シード選定のコストも小さくできる．これが提案手法の特徴である．

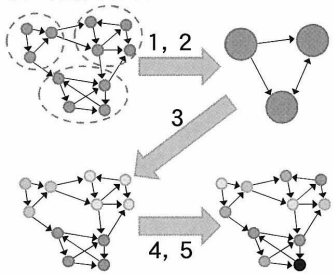


Fig. 3 アルゴリズムの流れ

3. 実験と結果

3.1 概要

本研究では，Web 上のブログに対してハイパーリンクを用いたネットワークを作り，そのネットワークに対して提案手法を適用し，信頼度のスコアを求める．いくつか無作為に抽出したページ群を調べ，スコアの妥当性について検討し，提案手法の有効性を示す．

用いたデータは，ソネットエンタテインメント（株）から提供されたブログデータ[b]である．ブログの記事をノード，記事中のリンクをエッジとして作成したネットワーク（ノード数: 141,356，リンク: 196,122）に対して提案手法を適用する．

3.2 コミュニティグラフに対する信頼度算出

はじめに，Clauset ら[8]の方法を用いてコミュニティグラフを作成した．ノードは 219 個に縮約された．これらのコミュニティを見るとブログ URL や内容が似通っており，コミュニティ縮約の過程において同質なものをまとめていることがわかった．

次に，この 219 個のコミュニティグラフに対して信頼度を求めた．10 個のシード選択し，精査した結果，4 個を信頼できるもの，6 個を信頼できないものと判断した．

このとき，コミュニティ内ノードを SMC 法によってクラスタリングし精査の手間を削減している（Fig.4）．精査した 10 個のシードに基づいて，コミュニティグラフの信頼度を求めた．Table.1 に信頼度の上位と下位の２つの値

b リコメンデーションコンテスト: http://www.so-net.ne.jp/web2/compe2008/contest.html

を示す．ここで，com#はコミュニティのインデックス，TrustRank，BadRank はそれぞれのスコアである．

表中の Total は信頼度であり，信頼度$\mathbf{t}^*$とし，

$$\mathbf{t}^* = \mathbf{t} - \tilde{\mathbf{t}} \qquad (7)$$

で算出している．この$\mathbf{t}^*$の分布を信頼度マップと呼ぶことにする．また，値の差をわかりやすく表示させるため，次式のように対数をとった値にしている．

$$\mathbf{t}_{\ln} = \log(1000 \cdot \mathbf{t} + \mathbf{1}) \qquad (8)$$

各々のコミュニティ内を精査すると信頼度の高いものにはほとんどスパムが含まれておらず，低いものにはスパムが数多く含まれていることがわかった．

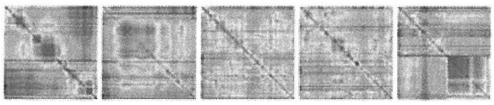


Fig. 4 コミュニティグラフの一部シードを可視化したもの

Table. 1 信頼度の例

com#	TrustRank	BadRank	Total
207	4.641	1.581	4.603
212	4.589	0.888	4.575
・・・	・・・	・・・	・・・
176	0.863	3.938	-3.911
184	0.322	4.049	-4.042

Fig.5 に，コミュニティグラフに対する$\mathbf{t}$，$\tilde{\mathbf{t}}$，$\mathbf{t}^*$の分布を示す．Fig.5 下は信頼度の分布図である．信頼度の値を高い順に並べたもので，縦軸は信頼度の値，横軸はコミュニティの信頼度の順位を表している．Fig.5 上は TrustRank，BadRank の分布図である．信頼度の高い順に並べている．なお，BadRank は縦軸負方向に示している．Fig.5 から TrustRank と BadRank の双方の値が大きいものが存在することがわかる．これらは注意すべきコミュニティであり，双方の値を評価する$\mathbf{t}^*$の意味が重要になる．ここで，数は少ないが，TrustRank と BadRank ともに 0 になるコミュニティがある．これらは値が伝播しなかったコミュニティである．

3.3 コミュニティ内の信頼度算出

前節で求めたコミュニティに対するスコアを初期値として，各コミュニティ内のノードに与え，ネットワーク全体のスコアを求める．Fig.6 に，典型的な 2 つのコミュニティ内の TrustRank，BadRank の分布を示す．

Fig.6 左は com#が 146 のコミュニティで，499 個のノードからなっている．Fig.5 と同じく，信頼度の高い順にノードを並べている．縦軸は TrustRank，BadRank の値で，横軸はノードの信頼度の順位を表している．このコミュニティの TrustRank は 4.016，BadRank は 2.186 である．全体で 3 番目に信頼度の値が高い．

Fig.6 右は com#が 110 のコミュニティで，291 個のノードからなっている．コミュニティの TrustRank は 0.104，BadRank は 3.216 であり，全体の中でも 7 番目に信頼度の値が低いコミュニティである．

3.4 信頼度マップの表示

ここまで成分として調べてきた信頼度マップをネットワーク可視化によって表現する．はじめにコミュニティグラフの信頼度マップを示す（Fig.7，Fig.8）．描画には Pajek[c]を用いている．

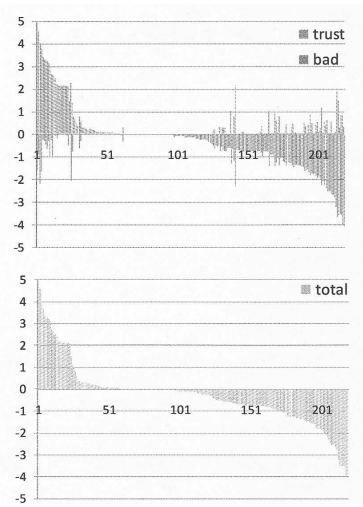


Fig. 5 コミュニティグラフに対する分布図

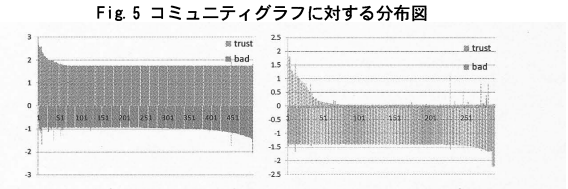


Fig. 6 各コミュニティ内の TrustRank, BadRank の分布
左：com#146，右：com#110

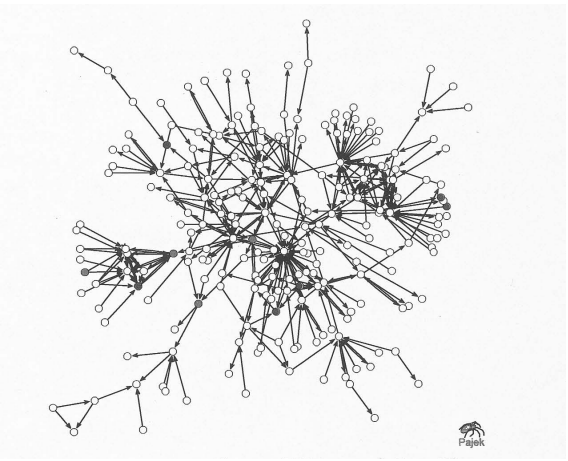


Fig. 7 コミュニティグラフの信頼度マップ（シード）

Fig.7 はコミュニティグラフに与えたシードで，赤色のノードが良いシード，青色のノードが悪いシードである．これらのシードに対して TrustRank，BadRank，信頼度の各指標を求めることによって，作成された信頼度マップが Fig.8 左下，Fig.8 右下，Fig.8 上である．それぞれ値の高いノードを大きく，低いノードを小さく表示している．

次に，各コミュニティ対して信頼度マップを示す．信

c Pajek: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

頼度マップは 3.3 節で分布を示した典型的なコミュニティを含む 2 例であり，それぞれ 2 つのコミュニティを示している．Fig.9 上段は高い信頼度のコミュニティのマップ，下段は低い信頼度のコミュニティのマップである．

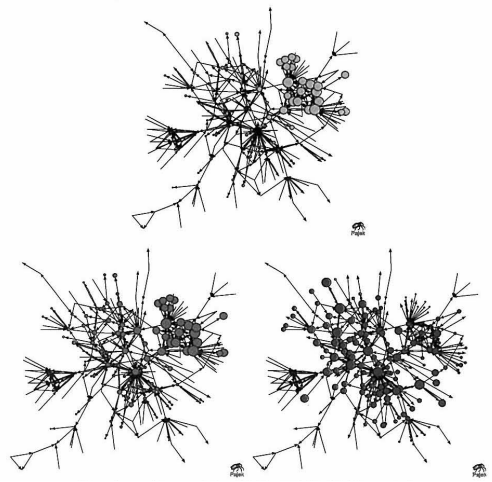
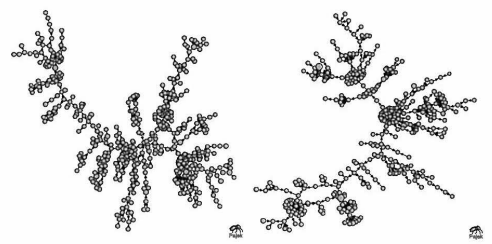
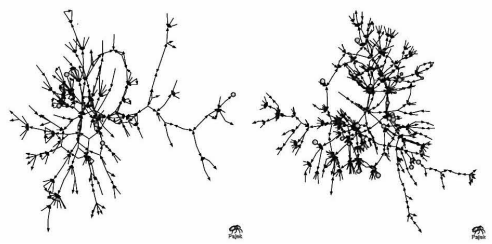


Fig. 8 コミュニティグラフの信頼度マップ
上:信頼度，左下:TrustRank，右下:BadRank



高い信頼度のコミュニティ（左:com#146，右:com#121）



低い信頼度のコミュニティ（左:com#110，右:com#177）

Fig. 9 コミュニティ内の信頼度マップ

3.5 比較実験

従来手法（コミュニティグラフ，コミュニティ構造を用いないもの）と提案手法を比較する．

従来手法において，シード数を提案手法と同じにするために，提案手法の中で精査したコミュニティに含まれるノードをシードとする．評価は，以下のカバー率によって行う．

$$\text{カバー率} = \frac{t_i\text{の値，}\tilde{t}_i\text{の値のいずれかが 0 でないノードの個数}}{\text{ネットワーク全体のノードの個数}}. \quad (9)$$

Table.2 で示すように，既存手法でのカバー率は 2.32%に対し，提案手法では 98.5%である．

従来手法では，シードの個数が 3,282 個に対し，スコアが伝播したものが 3,286 個である．これは，ブログネットワークでは明確なコミュニティ構造が存在しており，コミュニティ間の伝播が促進されないことを意味している．比較実験からは，このようなネットワークに対する信頼度の算出には，十分にコミュニティ構造を考慮する必要があることがわかる．

Table. 2 カバー率の比較

従来手法	提案手法
2.32%(=3,286/141,356)	98.5%(=139,299/141356)

4. 結論

Web 上の情報の信頼度算出において代表的な手法である TrustRank と BadRank の課題（シードからの距離が遠いページの精度が低下する，シードを精査する必要がある）を解決し，情報の信頼度を少ない人的コストで定量的に評価する方法を提案した．

提案手法は，TrustRank と BadRank の算出に，コミュニティグラフを利用し，多段階で信頼度算出を行うものである．またシード選択の際には SMC 法を併用し，シード選択の手間を削減している．

提案手法を実際のブログデータに適用した結果，少ないシードであっても，ネットワーク中のノードへの信頼度の伝播効率を示すカバー率が高くなることを示した．

また，提案手法によって作成した信頼度マップによって，ページの信頼度の分布が直感的に俯瞰できることを示せた．

参考文献

[1] 土屋大洋，浜屋敏，吉田倫子: ブログ・SNS の創発的特性と組織へのインパクト，富士通総研研究レポート No.269 (2006)

[2] 中村健二，田中成典，古田均，北野光一，寺口敏生: カテゴリ分類と時系列情報に基づくブログスパム判定手法の提案，情報処理学会論文誌，情報処理学会，**49.3**, pp.1119-1130 (2008)

[3] Takeda, T. and Takasu, A.: UpdateNews: a news clustering and summarization system using efficient text processing, Proc. of the 2007 Conf. on Digital libraries, pp. 438-439 (2007)

[4] 石田和成: スパムブログの定量的調査と分離の試み，データベースと Web 情報システムに関するシンポジウム論文集，DBWeb2007,5B (2007)

[5] Gyongyi, Z., Garcia-Molina, H. and Pedersen, J.: Combating web spam with TrustRank, Proc. of the 30th Intl. Conf. on Very Large Databases (VLDB), pp. 576-587 (2004)

[6] Wu, B. and Goel, V. and Davison, B.D.: Propagating trust and distrust to demote web spam, Models of Trust for the Web (MTW) (2006)

[7] 今藤紀子，喜連川優: Max-Flow コミュニティグラフとその特徴分析，日本データベース学会 Letters, **3.1**, pp.69-72 (2004)

[8] Clauset, A., Newman, M.E.J. and Moore, C.: Finding community structure in very large networks, Physical Review E, **70.6**, 66111 (2004)

[9] Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Mech, E.L.J.S.: Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, Vol. 2008, P10008 (2008)

[10] Kameyama, S., Uchida, M. and Shirayama, S.: A New Method for Identifying Detected Communities Based on Graph Substructure, Proc. of the 2007 IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology, pp. 263-267 (2007)