社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MPS－72 (45)
2008－BIO－15 (45)
2008／12／18

リンク構造によるドキュメントの信頼度算出法

亀山周明[†] 白山晋[††]

会社組織において社内ブログ/SNSの活用が進みつつある．生産の現場においては，情報共有の目的で，そのような仕組みが利用されるようになっている．しかしながら，インターネット上のブログ/SNSと同様に情報の信頼度の問題がある．本稿では，社内ブログ/SNSの上のドキュメント群に対し，参照関係から生じるネットワークのリンク構造を用いることによって，信頼度の高い情報とそうでないものを判別する方法を提案する．

A Method for Assessing the Reliability of Documents Using a Co-citation Network

SHUMEI KAMEYAMA[†] SUSUMU SHIRAYAMA[††]

In-house blogs and/or SNS are going into use in corporate organization. Those kinds of services are being used for information sharing at production sites, and have played important roles in developing and maintaining the information system. However, there exist several issues about information handling, as well as those on blogs or SNS on the Internet. One of the most important issues is to quantify the reliability of information. In this paper, we propose a method for assessing the reliability of documents using a co-citation network occurs from any stories or reports on the in-house blogs and/or SNS.

1. はじめに

パソコンに詳しい一部の人だけが更新していたホームページであるが，近年，ブログやSNSなどの容易な情報発信手段が普及し，一般利用者がWeb上で情報を発信することが珍しくなくなっている．特にブログ利用者の伸びは目覚ましいものがあり，ブログ自体が一般化しているといえる（総務省の調査によれば2008年1月現在，インターネット上で公開されている国内のブログの総数は約1,690 万あるとされる[1]）．

それにともない，この情報発信の手段は，会社組織においても社内ブログや社内SNSといった形で利用されるようになっている．そして，社内ネットワークの上に様々な報告書も存在するようになっている．また，そのような情報には質の高いもの，有用なものが数多く存在し，社内の情報システムにとっても重要な役割を担うことが指摘されつつある[2]．

しかし一方で，粗悪なもの，信頼性の低い情報が多数含まれるのも事実である．例えば，簡易的な参照サイトの存在であり，これは一般のブログにおけるスパムサイトと部分的に似た性質をもつ．スパムサイトとは，以下のような特徴を持つものとして定義されている[3]．

(a) 機械的に作られている
(b) 情報付加価値がない（既存の情報を単純にコピーしたようなものを意味する）
(c) 収益を主とした隠れた意図がある

スパムは情報付加価値がないのでコンテンツとして無意味でユーザーに何ら便益をもたらさない．また，悪質なサイトに誘導するためのページであることも少なくない．LinらやKolariらは，スパムブログが検索の質やネットワーク資源の浪費を招いていると指摘している[3,4]．このため，産業界においては，悪質，あるいは情報の信頼度の低いスパムサイトを減少させることが喫緊の課題として取り上げられている[5,6]．社内ブログでは，悪質サイトへの誘導という点を考慮する必要性はほとんどないが，資源の浪費や検索の質を低下させるという弊害については十分に留意しなければならない．

社内ブログや社内SNSを効果的，かつ効率的に活用するためには，情報の選別，統合・融合の効率化や精度の向上，不正行為の検知などが必要になる．このためには，情報の質を確保するフィルタリングシステムが不可欠になると考えられる．スパム的な情報の考慮は極端であるが，それを含めて情報の信頼度を定量的に評価する必要がある．

後述するように，情報の信頼性は意味的内容によって判断されることがほとんどである．一方，社内ブログや社内SNSの仕組みからは情報間の参照関係が比較的容易に抽出できる．本稿では，ブログやSNSの記事とコメントを纏めて1つのドキュメントとして扱い，トラックバックのように記事同士を結びつける仕組みが参照関係を形成すると考える．この参照関係によって生じるネットワークの構造を利用した信頼度の算出法を提案する．

2. 既存研究

社内ブログや社内SNSにおいて，悪質サイトへの誘導のような性質を持つスパムサイトが存在することはほとんどないが，研究面ではそのような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．

スパムの発見には人手によるものと機械処理によるものがある．さらに，機械処理によるスパム発見は，意味内容（言語処理）を用いるもの[7,8]，ネットワーク構造を用いるもの[9,10]に大別される．

前者の主なアプローチとしては，スパム確率を与えた辞書を用いる，データセットを与えての教師付き学習，語の頻度分析，記事を要約しての比較などが挙げられる．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には2つのものがある．1つは，スパムに特徴的な2部グラフ構造などの抽出である[9]．この方法の問題には，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が発見することができないということがある．

*† 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

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 空間での信頼度の算出そのものに利用されている．代表的なものが Gyongyi らによって提案された TrustRank である[10]．TrustRank では，「良いページは良いページへリンクしている可能性が高い」という考え方にもとづき，信頼度を算出する．はじめに，いくつかのページ（シード：seed と呼ぶ）に対して有識者らによって数値として信頼度が与えられる（スコアと呼ぶ）．それを伝播させることで，Web 空間上で信頼度を得るというものである．一方，信頼度の低いものを伝播させるものがある．BadRank と呼ばれている[16]．これらの方法には，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．

本研究では，TrustRank と BadRank を改良し，併用することによってドキュメントの信頼度の算出を試みる．

3. 提案手法

提案手法は，TrustRank [10]と BadRank[16]の考えにもとづき，参照関係のネットワークに含まれる特徴的な構造のパターンを利用し，より正確かつ効率的に信頼度を算出するものである．ドキュメントの存在する場所をページと呼ぶことにする．

3.1 TrustRank と BadRank の併用

TrustRank は「良いページは良いページへリンクしている可能性が高い」という原理にもとづく．Fig.1 に TrustRank の概念図を示す．ノードが Web ページ，矢印がリンクを表している．白丸ノードが良いページ，黒丸ノードが悪いページである．

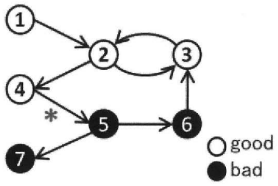

Fig. 1 TrustRank 概念図 [10]

TrustRank のアルゴリズムは以下である．

1. 有識者などが Web ページの一部を精査し，信頼度を数値として与える．
2. 次式を繰り返し，信頼度を Web 全体に伝播させる．

$$\boldsymbol{t} = \alpha \cdot \mathbf{T} \cdot \boldsymbol{t} + (1-\alpha) \cdot \boldsymbol{d} \qquad (1)$$

ここで，$\boldsymbol{t}$ は TrustRank，$\mathbf{T}$ は遷移行列，α は減衰ファクター，$\boldsymbol{d}$ は特定ページへの加算項を表す．上式は Biased PageRank として知られた式である．TrustRank は自分を指すページの信頼度のスコアと，ある特定のページとの和で表現される[10]．Fig.2 にネットワーク上にスコアを重ねた例を示す．図中で 1 を与えられたノードがシードを表す．BadRank は TrustRank とは逆に信頼度の低いもののスコアを伝播させるものである．方法は，TrustRank と同様である．

TrustRank, BadRank ともに，シードから遠いページの精度が低下するという問題がある．TrustRank では，アルゴリズム的にスコアの伝播が起こりやすい，出次数が多いページの選択が推奨される．したがって，そのようなページから中味の精査を行い，シードを選択するのが効果的とされている[10]．一方，ブログの場合，出次数の多いページはスパム群である可能性が高い．よって，せっかく精査してもその結果が無駄になる場合も多いと考えられる．そこで，TrustRank と逆の概念である BadRank を併用し，精査の結果を無駄なく活かすようにする．

また，TrustRank では，信頼度のスコアを伝播させていく際に伝播経路内にスパムサイトが含まれても，通常のサイトとして加味されてしまうが，BadRank の併用によってそのようなものを避けることができる．BadRank においても同様のケースがある．

ただし，TrustRank でも BadRank でも，特別に扱われる特徴的なネットワーク構造は少ないことに留意すべきである[a]．また，ネットワークが大規模になると，計算時間と記憶容量の問題が生じる．

本稿では，これらの問題点を解決するために，TrustRank と BadRank の拡張を行う．具体的な手法を次節から示す．

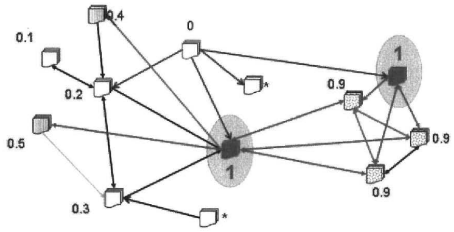

Fig. 2 TrustRank の例

3.2 コミュニティ構造を利用した拡張

Fig.3 に示すようにコミュニティ構造を考慮し，重み付けなどを行った上で信頼度のスコアを算出する．

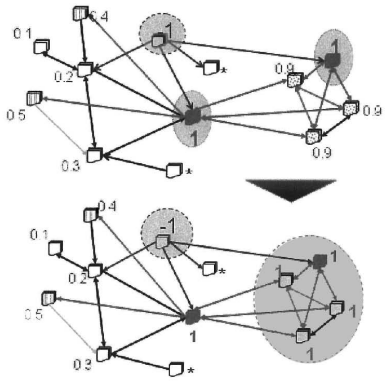

Fig. 3 コミュニティ構造の考慮

本稿では，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．前者にも後者にも，有向グラフに対するものが提案されているが，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．

a ネットワーク構造として遷移行列に反映される部分はある．

前者として，Clauset ら[14]と Blondel ら[15]の方法を利用する．得ら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，著者らの SMC 法[12]を用いる．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．

3.3 コミュニティグラフの利用

コミュニティをノードとし，コミュニティ間の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[13]を作成する（Fig.4）．そして，階層的な TrustRank と BadRank の算出を行う．

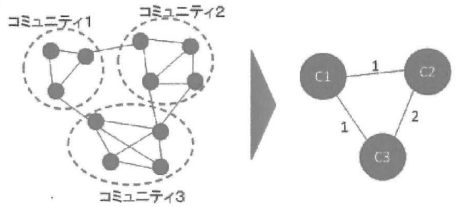

Fig. 4 コミュニティグラフの作成

はじめに，コミュニティ内のスコアの代表値を求める．単純平均による方法，コミュニティ内での伝播を利用する方法，コミュニティ内の特徴的な構造を利用する方法によって代表値を算出する．

コミュニティ内での伝播を利用する方法は，各々のコミュニティに対して式(1)によってスコアを求める方法である．コミュニティ内の特徴的な構造を利用する方法は，Fig.5 のように特徴的な構造を利用してスコアを求める方法である．

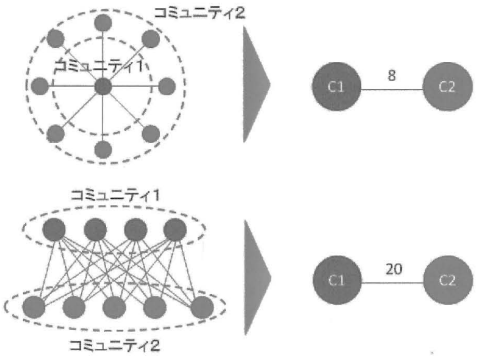

Fig. 5 特徴的な構造を縮約したコミュニティグラフ

次に，コミュニティグラフに対して，TrustRank の場合は，式(1)によってノードのスコアを伝播させる．

最後に，伝播後の結果をコミュニティ内にフィードバックさせる．この際にも，コミュニティ構造を利用する．

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．これにより計算時間，記憶容量の問題が緩和できる．また，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ものと考えられる．

3.4 モチーフの利用

モチーフとはネットワークに頻繁に現れる特徴的な構成要素のことである[11]．モチーフの構造はネットワークの持つ機能，ネットワーク上で起こる現象と関係していると考えられている．

著者らは大規模ネットワークからの効率的なモチーフの抽出法を構築している[12]．この抽出法では，Fig.6 下図に示すようにモチーフ内における類似した構造のグルーピングもできる．この方法は，Fig.5 に示したネットワークの特徴的な構造の抽出に利用できる．一方，本節では，特徴的な構造の直接的な利用方法について説明する．

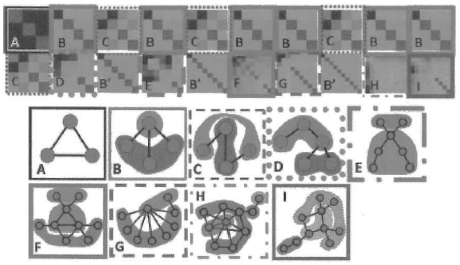

Fig. 6 モチーフの抽出 [11]

3.4.1 不要ページの除外

複数のページを指すページは関連するドキュメントの在処をまとめたポータル的なページである可能性が高い．しかしながら，Fig.7 のように，そのページがどのページからも指されていない場合は役に立つページとはいえない．TrustRank の場合，このようなページを不要ページとして除外する．

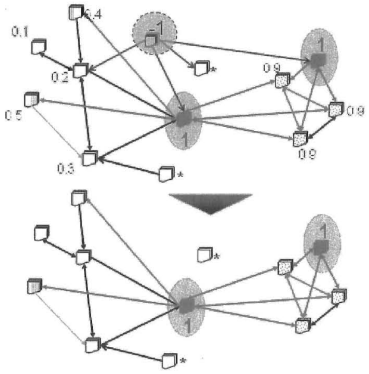

Fig. 7 不要ページの除外

3.4.2 モチーフによるスコアの補完

シードから遠いページの信頼度の予測精度が低下するという問題をモチーフによって解決する．

シードから離れてスコアが減衰したページがモチーフに含まれる場合は，同様のモチーフでスコアの精度が保たれていると思われるものを探し，その値から減衰したスコアを補正する．例えば，Fig.8 の右のモチーフ上では，信頼度が曖昧であるとする．一方，左で示すモチーフでは高い精度で信頼度が求められているとする．このとき，

モチーフに着目すると，紫で囲まれた部分は同じ構造をしている．同じ構造をしているということは，周辺のノードの役割も似通っていると考えられる．精度が高いモチーフから低いもののスコアを補完する．

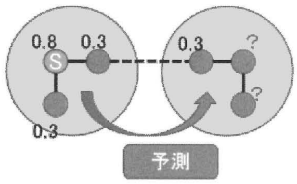

Fig. 8 モチーフによるスコアの補完

4. 実験と結果

社内ブログ/SNS のデータを入手することは難しい．そこで，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調べる．

本研究では，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の最大連結成分群（ノード数: 141,356，リンク: 196,122）を用いる．Fig.9 は，Clauset ら[14]の方法を用いて作成したコミュニティグラフである．ノードは 219 個に縮約されている．

これに対して信頼度を求める．Table.1 に，信頼度の上位 5 位，下位 5 位の値を示す．ここで，Total は TrustRank-BadRank として求めている．

それぞれ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．

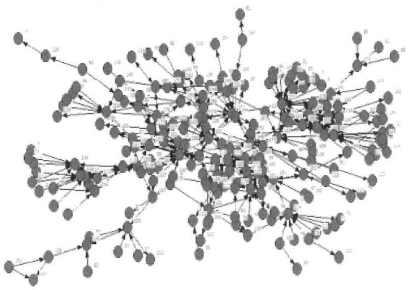

Fig. 9 コミュニティグラフ

5. 結論と今後

社内ブログ/SNS のドキュメント間の参照関係から生じるネットワークの構造を利用した TrustRank と BadRank にもとづく信頼度の算出法を提案した．

TrustRank と BadRank の併用によって，シードの効率的な選択が可能になり，信頼度スコア伝播時にそれぞれを補うことが示唆できた．また，２つの課題（シードから遠いページの精度の向上，ネットワークの大規模化にともなう計算時間と記憶容量の増大の軽減）の解決策として，コミュニティグラフとネットワークモチーフを用いた方法を提案した．

Table. 1 信頼度の例

com#	TrustRank	BadRank	Total
211	0.10848	0.006747	0.101733
212	0.09846	0.005406	0.093054
188	0.01239	0.000774	0.011616
86	0.00895	0	0.00895
144	0.01154	0.002799	0.008741
・・・	・・・	・・・	・・・
207	0.01048	0.038885	-0.02841
176	0.00081	0.03753	-0.03672
146	0.00146	0.038811	-0.03735
121	0.00243	0.041585	-0.03916
184	0.00023	0.042039	-0.04181

参考文献

[1] ブログの実態に関する調査研究，総務省 (2008)
[2] 社内ブログ活用研究会：http://shanaiblog.com/
[3] Lin, Y.R., Sundaram, H., Chi, Y., Tatemura, J. and Tseng, B.L.: Splog detection using self-similarity analysis on blog temporal dynamics, Proc. of the 3rd Intl. workshop on Adversa- rial information retrieval on the web, pp. 1-8 (2007)
[4] Kolari, P., Java, A. and Finin, T.: Characterizing the Splogosphere, Proc. of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wide Web Conference (2006)
[5] ニフティ，スパムブログのフィルタリング技術を開発：http://www.nifty.co.jp/cs/07shimo/detail/080326003337/1.htm
[6] サイトのリンク構造から有害コンテンツを判定，東大とトレンドマイクロ：http://internet.watch.impress.co.jp/cda/news/2008/01/28/18251.html
[7] 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)
[8] Takeda, T. and Takasu, A.: UpdateNews: a news clustering and summarization system using efficient text processing, Proc. of the 2007 conference on Digital libraries, pp. 438-439 (2007)
[9] 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)
[10] 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)
[11] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U.: Network Motifs: Simple Building Blocks of Complex Networks, Science, Vol. 298, no. 5594, pp. 824-827 (2002)
[12] Kameyama, S. and 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)
[13] 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，第 15 回データ工学ワークショップ(DEWS2004)，電子情報通信学会 (2004)
[14] Clauset, A., Newman, M.E.J. and Moore, C.: Finding community structure in very large networks, Physical Review E, Vol. 70, no. 6, 66111 (2004)
[15] 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)
[16] 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)