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

2009－ICS－155 (5)
2009／3／21

複雑ネットワークからの構造情報抽出

大沢 英一†, 珍田 計幸†

†公立はこだて未来大学

要 旨

近年，複雑ネットワークに関する研究が盛んである．論文の共著関係，俳優の共演関係やインターネットの端末網，線虫の神経細胞など実世界に存在する多くのネットワークがスモールワールドやスケールフリー性といった特性を持つということが知られている．WWW もその典型的な例であり，WWW のハイパーリンクのネットワークを複雑ネットワークとして捉えた研究が進んでいる．しかし，複雑ネットワークの効率性や集積性といった利点が存在することが判明してきているにも関わらず，これらの利点を実用レベルで活用するまでに至っていない．本研究では，複雑ネットワーク研究の土台とも言える社会学の知見に着目し，その知見を WWW に適用することによる情報抽出手法の提案を行なう．特に本研究においては，弱い紐帯の強みに関わる知見を利用し，WWW における弱い紐帯を手がかりにした情報抽出手法を提案する．また，WWW のハイパーリンクの構造を適度に補正することによって，構造からの抽出法に対して，精度を上げることができる．これらの抽出手法によって，従来のキーワード検索によるウェブ検索システムでは発見が困難であった有用な関連ページの抽出が可能となる．

Extraction of Structural Information from Complex Networks

Ei-Ichi Osawa†, Kazuyuki Chinda†

†Future University - Hakodate

Abstract

Recently, the research on a complex network is active. Most networks such as paper co-citation network, movie actor collaboration network, Cellular networks and the Internet are known as Complex Networks. It contains features such as Small World, Scale Free and Clustering. The WWW is being rapidly researched as Complex Networks, because the network of hyperlink at the WWW is the typical example of Complex Networks. By contrast, it's not being leveraged the features that is efficiency or clustering of Complex Networks. This research's proposal is the information extraction technique by knowledge of the social science; in addition, it applies the WWW. Especially, we focus attention for the weak ties, and we propose the information extraction technique on the WWW.

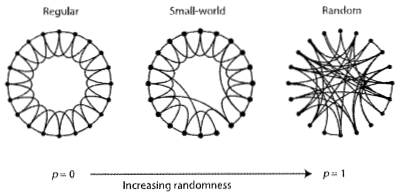

図 1: WS モデルによるスモールワールド・ネットワーク（文献 [12] Fig.1）

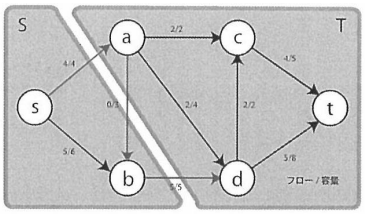

図 2: 最大フロー最小カット定理によるカット

1 まえがき

1.1 複雑ネットワーク

現在，社会的な人間関係や電力の供給網，インターネットの端末網，WWW，ニューロンのネットワークなど，実世界における実に多くのネットワークが，スモールワールドやスケールフリーといった特性を持つ「複雑ネットワーク」の構造からなるネットワークであることが知られている [1, 12].

これらのことは，1960 年代に社会学者である Milgram の実験 [8] や Granovetter の実験 [5] により人間関係におけるネットワークの特徴が明らかにされてきたことに端を発する．また，その過程において，特に「スモールワールド問題 (small world problem)」と呼ばれる話題が提起されたが，このスモールワールドに代表される複雑ネットワークが近年になって数学や物理学，生物学，計算機科学などの分野においても注目されている [13]．その契機となったのが，Watts と Strogatz により提案された WS (Watts-Strogatz) モデル [12] である．これは，図 1 に示すような 1 次元格子を用いた完全グラフを元に生成される．この完全グラフのパスを，ある (ほんの僅かな) 一定の確率により別のノードに繋ぎ変える作業を行うことで，WS モデルによるスモールワールド・ネットワークが完成する．ここで，ノードを繋ぎ換えたパスがネットワークにおけるショートカットの役割を果たすことで，このネットワークは距離のより小さな世界（スモールワールド）へと変化する．このように，規則的なネットワークとランダムネットワークとの両者の中間に位置し，ネットワークの直径が大きく，かつクラスタ係数[1]の高いネットワークのことをスモールワールド・ネットワークと呼ぶ．

[1] ネットワークにおけるノードの集積の度合いを示す．

2 ネットワーク構造からの情報抽出

2.1 フローネットワーク

最大フローアルゴリズム [2] はネットワーク計画法におけるフローネットワークの最大フロー問題を WWW に適用したものである．最大フロー問題とは，図 2 に示すような各エッジに容量 (限界流量) が与えられたネットワークに対して，入口 (source: s) から出口 (sink: t) へのフロー（流量）を最大にするような経路を計画する問題で，数理計画法の一分野である．

最大フロー問題では，最大のフローを取るときのネットワークのカットが最大フロー最小カット定理によって与えられる．最大フロー最小カット定理によるカットの例を図 2 に示す．このカットされたエッジはネットワーク全体における均衡点であるため，WWW に適用した場合にはこのカットによって分割される 2 つのネットワークがコミュニティとして抽出可能となる．

このカットを再帰的にネットワークに対して適用することで，複数のコミュニティが同定可能である．

また，最大フローアルゴリズムによって抽出されるコミュニティは HITS algorithm に比べ，より抽象的な概念を含むようなコミュニティとなる．

2.2 GN アルゴリズム

GN (Girvan–Newman)-algorithm [10] は中心性[2]の尺度のひとつである *betweenness* を同定することでネットワークのノード，もしくはエッジを削除することが可能であるというアイディアの元に成り立っているアルゴリズムである．*betweenness* はネットワークにおいてある 2 点間のノードに対する寄与率の大きさを示したものである．

[2] 社会学の分野において Freeman が *degree*(次数中心性), *betweenness*(媒介中心性), *closeness*(距離中心性) の 3 つの中心性 (centrality) を代表的な尺度としてまとめている [3].

すなわち，*betweenness* の高いノード（エッジ）をネットワークから取り除くと，ネットワーク全体としての効率性が低下する．このように，*betweenness* の高いものから順番に取り除くことで小さなネットワークが複数出来上がることになる．

betweeneess はネットワーク全体に対して計算する必要があり，WWW のような巨大なネットワークに適用することは現実的ではなかったが，ネットワークの全体を参照せずに *betweenness* を算出する近似アルゴリズム *shortest-path betweenness* が知られている [9]．

GN-algorithm では，*shortest-path betweenness* などを適用することによってコミュニティの分割を行なう．具体的には，クラスタ分割におけるデンドログラムグラムを分割することでコミュニティが得られる．

2.3 WWW からの情報抽出

2.3.1 HITS アルゴリズム

HITS algorithm [6] では，重要なノードの重み付けに *authority* と *hubness* を導入している [7]．これらの重み値は，他のページから多くリンクされているページを *authority* が高い（権威のある）とし，また，そのページ内に記述しているリンク（内部リンク）が，重要度の高いページへのリンクであればそのページの *hubness* が高い（拠点性が高い）とする．そして，重み *authority* の値の高いページを AUTHORITY，重み *hubness* の値の高いページを HUB として定義する．つまり，優れた HUB とは，多くの AUTHORITY への内部リンクを持つページのことであり，また，優れた AUTHORITY というのは，多くの HUB からリンクを獲得しているページのことである．

HITS algorithm は，図 3 に示すような HUB, AUTHORITY からなる完全 2 部グラフが抽出可能なアルゴリズムである．しかし，HITS algorithm では，抽出可能なコミュニティは一つのみで，必ずしも適切なものが抽出されるとは限らないため，他の手法や文書解析などを併用する必要がある．

2.4 PageRank

PageRank[11] はウェブドキュメントに対して PageRank と呼ばれる重み付けを行うことにより，ウェブドキュメントのランク付けを行う．

あるドキュメント p_i の PageRank は次式によって与えられる．

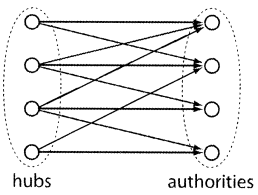

図 3: HUB, AUTHORITY からなる完全 2 部グラフ

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in \mathrm{M}(p_i)}^{n} \frac{\mathrm{PR}(p_j)}{\mathrm{L}(p_j)}$$

ここで，N を対象とする全てのウェブページ，$\mathrm{PR}(p_j)$ を各ページ p_i にリンクしているページ p_j の PageRank，$\mathrm{L}(p_j)$ をページ p_j に含まれる他のページへのリンクの総数とする．定数 $d(d \leq 1)$ は減衰因子（damping factor）であり，恣意的に PageRank を上げようとする行為に対し，より小さい値を設定することでこれに対応することができる．

PageRank は重要なページは重要なページからリンクを獲得する，という考え方から incoming link を重視する．しかし，リンク集などの多くの outgoing link を保有するページからの incoming link はさほど重要であるとはみなさない．なぜなら，確かにリンク集は有用かもしれないが，ドキュメントとしての内容が充実しているものであるとは言えないからである．

こういった方針によって，PageRank ではリンク集のみのページや，ある限定されたコミュニティ間のリンク結合，サイト内リンクのみからなるページなどといったものの重要度は低くなる．

3 提案

本稿では，複雑ネットワーク，特に WWW を対象とした情報抽出技術について述べる．

3.1 弱い紐帯の概念を用いた情報抽出

栗原らは，検索エンジンを用いた検索結果から，WWW における弱い紐帯 [4] にあたるページの発見手法を提案している [14]．ある検索対象に対して複数の異なる検索語句による検索を行ない，検索結果のページ集合の共通部分に位置するページを解析することで実現している．

ここで注目をするべきページは No. 3 に見られるパタンのページである．このようなページには，ど

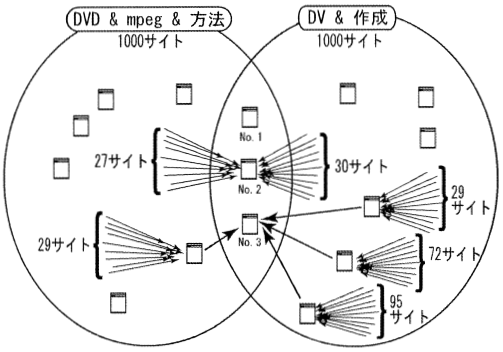

図 4: 栗原らの手法 (文献 [14] 図 5)

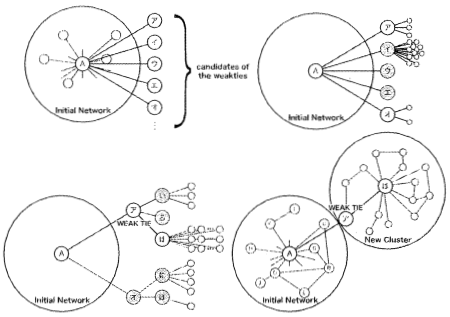

図 5: 抽出の概念図 1

ちらのクラスタからもAuthority型のページからリンクをされており，さらに検索における表示順位が低かったという特徴があった．また，このページを実際に訪問したところ，「非常に使えるWebサイトであった [16]」ことが確認されている．

このことから，WWWにおける弱い紐帯が存在することがいえる．したがって，構造情報から弱い紐帯を同定することで，情報抽出を行うことができるはずである．

ここで，問題点はネットワークにおいて弱い紐帯を同定するためには，ネットワーク構造を大局的に把握することが必要であるということである．もしくは，ネットワーク全体まで扱わずとも，対象とするノード周辺のネットワーク構造を少なくとも知る必要があるということである．具体的には，Granovetterの定義に則り，局所ブリッジ [5] を探索しその次数を求めることや，媒介中心性などの中心性の尺度 [3] をネットワークに適用することなどによって同定可能である．しかし，WWWのような巨大なネットワークに対しては計算コストやデータベースの必要性などの観点から，その導出は困難である．

3.1.1 提案

そこで，WWW上における弱い紐帯の定義を新たに与えることと，既存の検索エンジンを利用することでネットワーク全体を参照せず，また，計算コストを軽減させた，弱い紐帯の発見同定手法の提案を行なうことが可能である [15]．抽出の概念図を図 5 に示す．

表 1: 弱い紐帯の同定数

	総ページ数	Auth.	W.T.
分野 A (16 単語)	11,237	274	244
分野 B (14 単語)	11,226	281	217
分野 C (24 単語)	7,269	210	182
分野 D (20 単語)	13,603	342	327

3.1.2 実験結果

得られた弱い紐帯，Authorityのページ数を表 1 に示す．

ひとつのキーワードについて，弱い紐帯とAuthorityを同定するために必要なページ数は約580ページであった．距離2のAuthorityを探索するためには，その性質上から，探索するすべてのページを開始ページからリンクを3階層までたどることで実現可能である．

この実験から得られた弱い紐帯はひとつの単語につき約13ページである．また，得られたAuthorityについても約15ページと同等のページ数が得られた．しかし，実際に得られたAuthorityや弱い紐帯のページを訪問したところ，そのほとんどが有用であるとは言い難いページであった．したがって，この提案手法によってリンク構造のみから弱い紐帯を完全に同定することは困難であることが推察される．

3.1.3 追加実験

抽出されたAuthorityにはAuthorityにふさわしくないようなページが存在した．これは，Authorityの判別を2部グラフなどを考慮せず，被リンク数のみから行なっているためである．

しかしながら，抽出された弱い紐帯の中にはこの中には本来の弱い紐帯役を担っているページが存在するはずである．それらのページをより高い精度で

抽出するためには，見せかけのAuthorityを排除する必要がある．そこで，得られたAuthorityについて，以下の基準によるページの分類を人手により行なった．

多くのページは，掲載数の少ないリンク集，アクセス解析やランキングシステムなどの管理システムへのログインページ (その他) であった．

また，それら以外のページについてもAuthorityであるのにも関わらず，ほとんどのページが有用であるとは言えない内容のページであった．

さらに，Authorityとなっている要因が同一ドメインやサブドメイン内の相互リンクからの内部リンク（有用ではないリンク集がAuthorityと見なされていたのもこの理由による）が多いことによるものであるためであり，一般的に有用なコンテンツを含むページであるとは言えないものであった．

そこで，弱い紐帯と得られたAuthorityが同一ドメインによるもの，Authorityがパスワード入力を求める管理画面ならびにリンク集のページをそれぞれノイズとして除去した．

この処理によって，それぞれ約85%の「見かけ上の」Authorityが除去された．これにより，当初の実験よりも少数のページが弱い紐帯と同定され，また弱い紐帯が同定されなくなった語句も存在する．当然，弱い紐帯が存在しないケースもあるはずであり，この点は問題ではないと考えられる．

3.1.4 評価

実際に実験で得られた弱い紐帯の妥当性を，得られたAuthorityが有用なものであるかを考慮し，検証を行なう．ただし，ここでは具体的な事例を挙げるに留める．

- 事例1

Authorityとして国内自動車メーカAのトップページを起点とした場合，得られたAuthorityは2つのサイトが存在し，どちらも個人が運営する日記形式のページ (Weblogではない) であった．これらのページは科学技術一般について広く取り上げたサイトであり，起点としたメーカについても多くの話題を取り上げていた．また，この例における弱い紐帯はこの自動車メーカの新製品を取り上げたニュース記事であった．

この例で得られた弱い紐帯を経由したAuthorityについて，リンクで接続したページ群をクラスタとみなすと，このクラスタにはプログラミングや通信機器，PCなどの新製品の話題などを扱った日記やWeblog, Wikiなどのページ群が多く所属していることが分かった．

- 事例2

また，例1と同様に，国内自動車メーカBのトップページを起点として得られた場合，次のようなAuthorityが得られた

得られたAuthorityのページは切手の収集サイトで，特に昆虫の切手を扱ったページである．このページ自体は実際に訪問してみても，情報が整理され情報量も多い有用なページであった．このページから得られるクラスタは，切手を扱ったページか昆虫を扱ったページがほとんどであり，自動車に関するページは存在しなかった．

このようなページが抽出された要因のひとつとして，弱い紐帯と同定されたページが実は弱い紐帯の役割を担っていないページであったということが考えられる．この弱い紐帯のページは複数の話題をページ内で扱っており，偶然どちらのAuthorityからもリンクされていたページであったため抽出された例である．この例から，リンク解析のみで情報抽出を行う場合，たとえ有用であってもユーザが望む内容ではない情報を含む可能性がある．

3.2 統計情報に基づく確率モデルの導入

WWWからの情報抽出において，リンク構造の観点からの情報抽出について述べた．この手法による情報抽出では有用な情報抽出が可能であることを示す一方で，関連性の薄い情報が抽出されてしまうという問題点が存在した．そこで，この点に対する改善として，ネットワーク構造におけるエッジの存在確率を用いたネットワークの修正モデルを提案する．

3.3 統計情報に基づく確率モデルの導入

図6のように3ページ間のリンク関係を調査することにより，ネットワークにおけるリンクの存在確率が算出される．この情報に基づき，ネットワークにおいて不足しているエッジや，冗長なエッジを追加，削除を行なうことで，より適切なネットワーク構造を推定することが可能である．

WWWにおいては，本来付加されていないページ間のリンクを自動付加することで，より適切なネットワーク構造を推定でき，さらに情報抽出に役立てることが可能となる．

3.3.1 実験

WWWにおける確率モデルの導入に先立ち，予備実験を行なった．これは，WWW全域に対する

表 2: 3 ページ間のリンク割合

1	2	3	4	5	6	7
14.1	9.8	16.5	18.3	9.5	25.0	6.8

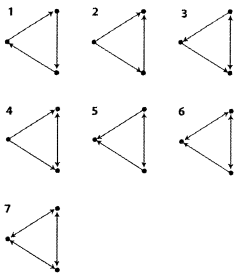

図 6: リンク構造

構造情報の抽出が困難であることから，ある特定の限定されたドメインにおける構造情報から本提案手法の妥当性を検証するためのものである．

3.3.2 予備実験

`http://www2.fun.ac.jp/` における調査結果を表 2, 図 6 に示す．

3.3.3 結果

以上の結果から模擬的に構成したネットワークに対して，自動リンク付加実験を行った．図 7 の上部は，ネットワークに対して，コミュニティ抽出法を適用した例である．

実験のため，このネットワークのエッジをある確率で除去したエッジの欠損したネットワークを生成した．破線で示したエッジはリンクが欠損しているエッジである．図 7 の左下の図は欠損したネットワークに対して，コミュニティ抽出法を適用した例である．望ましい抽出例に対して，コミュニティの範囲が小さくなっていることが分かる．

一方，確率情報を用いてエッジを自動付加した例を図 7 の右下に示す．左下に比較してより，望ましいコミュニティに近いことが分かる．

4 まとめ

本稿では，複雑ネットワークから構造情報を利用した情報抽出法について述べた．特に WWW における抽出法では，原理的に構造情報を用いている

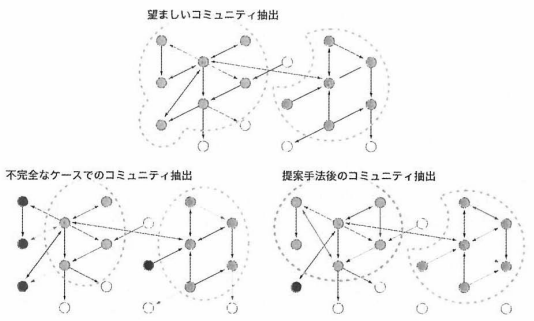

図 7: 実験結果

アルゴリズムが多いが，その検索精度は自然言語処理に頼っているのが現状である．本稿は，大規模なデータベースが必要となる自然言語処理を行うことなく，構造情報から関連情報の抽出が可能であることを示した．

また，キーワード一致による抽出法とは異なり，得られる情報の質が異なることも構造情報からの抽出方位における特徴であり，情報推薦としての活用も考えられる．

参考文献

[1] Barabási, A. and Albert, R.: Emergence of Scaling in Random Networks, *Science*, Vol. 286, No. 5439, p. 509 (1999).

[2] Flake, G., Lawrence, S., Giles, C. and Coetzee, F.: Sell-organization and identification of Web communities, *Computer(Long Beach, CA)*, Vol. 35, No. 3, pp. 66–71 (2002).

[3] Freeman, L.: Centrality in social networks: Conceptual clarification, *Social Networks*, Vol. 1, No. 3, pp. 215–239 (1979).

[4] Granovetter, M.: The Strength of Weak Ties, *The American Journal of Sociology*, Vol. 78, pp. 1360–1380 (1973).

[5] Granovetter, M.: Economic Action and Social Structure: The Problem of Embeddedness, *The American Journal of Sociology*, Vol. 91, No. 3, pp. 481–510 (1985).

[6] Kleinberg, J.: Authoritative sources in a hyperlinked environment, *Journal of the ACM (JACM)*, Vol. 46, No. 5, pp. 604–632 (1999).

[7] Kleinberg, J.: Hubs, authorities, and communities, *ACM Comput. Surv*, Vol. 31, No. 5 (1999).

[8] Milgram, S.: The small world problem, *Psychology Today*, Vol. 2, No. 1, pp. 60–67 (1967).

[9] Newman, M.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, *Physical Review E*, Vol. 64, No. 1, p. 16132 (2001).

[10] Newman, M. and Girvan, M.: Finding and evaluating community structure in networks, *Physical Review E*, Vol. 69, No. 2, p. 26113 (2004).

[11] Page, L., Brin, S., Motwani, R. and Winograd, T.: The pagerank citation ranking: Bringing order to the web (1998).

[12] Watts, D. and Strogatz, S.: Collective dynamics of 'small-world' networks., *Nature*, Vol. 393, No. 6684, pp. 409–10 (1998).

[13] マーク・ブキャナン：複雑な世界，単純な法則 – ネットワーク科学の最前線, 草思社 (2005).

[14] 栗原聡：スモールワールド性を利用した WWW からの情報発見構想, 日本ソフトウェア科学会インターネットテクノロジワークショップ 2003, 10 (2003).

[15] 珍田計幸，大沢英一：弱い紐帯の概念を利用した WWW 上からの情報抽出手法の提案, 信学技報, 第 108 巻 of *AI2008-81*, pp. 99–104 (2009).

[16] 福田健介, 栗原聡："ネットワーク" の科学（<特集> 複雑系と集合知），人工知能学会誌, Vol. 18, No. 6, pp. 716–722 (2003).