Web Community Browser: 大規模 Web グラフ探索ツールとそのインタラクション技術

Web Community Browser: A Browsing Tool of Large Web Graphs

福地 健太郎　豊田 正史　喜連川 優*

Summary. Web community is a collection of web pages created by individuals or any kind of associations that have a common interest on a specific topic. We had extracted 100 thousands communities from 40 million web pages in Japan by using link analysis technique. "Web Community Browser" is a tool to browse relationships between the communities. The browser shows communities and their relationships as an indirected graph, and its layout is optimized by physics-based graph layout algorithm.

1　はじめに

Web ページが増大する中で, それらから如何に情報を抽出するかが研究課題となっている. 我々は,Web ページ群を自動解析して, 同じトピックを共有するページ群である **Web コミュニティ**を抽出する手法を研究している. 先に提案した手法 [1] は, 与えられた Web ページ群のリンク情報を解析して, Web ページ群をコミュニティに分割し, 加えて個々のコミュニティ間の関連を重みつきの有向エッジとして出力する. 我々は同手法により,2001 年 10 月の時点で, 国内 4000 万ページを元にして 13 万個のコミュニティ・112 万本のエッジからなる巨大な有向グラフを抽出している. 我々はこのグラフを **Web コミュニティチャート**と呼ぶ.

Web Community Browser[2] は, 上記の手法で得た Web コミュニティチャートを可視化し, それらを閲覧・探索する為のツールで, 取得した Web コミュニティの中から, ユーザーが興味を持つコミュニティやその周辺との関連, グラフ構造等をインタラクティブに閲覧する事を目的としている. Web コミュニティチャートは非常に巨大なグラフであり, またその構造はハイパーリンク的で, リンク構造は複雑である. このようなグラフはその全体を適切にレイアウトするのは非常に困難である事が知られている. また,10 万個超のノードを持つグラフを通常のディスプレイ上に表示するのは, 解像度の問題があり現実的ではない. そこで,Web Community Browser では, ユーザーが全体の中から部分グラフを選択し, 動的に探索しながら表示するグラフを編集していく事で, レイアウトと表示の問題を解決している. また,Web コミュニティチャートは重みつき有向グラフの一種であるが, 一般的な重みつき有向グラフにも適用可能な可視化技法を提案する.

部分グラフの大きさは小さい程レイアウトにも表示にも有利となるが, コミュニティ構造を把握するためには, できる限り多くのコミュニティを表示する事が望ましい. しかし表示量が多くなると図形要素同士が重なりあい, 見易さを損ねる. そこで本研究では,Fisheye View[4] の概念を導入して, ユーザーが注目している部分を表示しながらグラフ全体の構造を併せて表示するための手法を提案する.

2　関連研究

Web のリンク構造をグラフを可視化する手法としては, 木構造を取り出して球体内に立体的に描画する H3 Viewer[7] があるが, コミュニティ群は一般には木構造ではない複雑なリンク構造をしており, コミュニティ群の関連を把握する目的には合致しない. WebOFDAV[8] はユーザーの訪れたページをグラフにして可視化するものである. ユーザーにとって既知のページ群の可視化には優れるが, 未発見の周辺コミュニティの探索を支援するものではない. DocSpace[9] は, 文献の関連関係を表わすグラフをバネモデルを用いてレイアウトしている.

グラフレイアウトをユーザーが動的に指示するシステムを Henry らが提案している [5]. 同研究では, ユーザーが部分グラフを選択してレイアウト手法を指示して, グラフレイアウトを編集する事で, ユーザーの関心を反映したグラフレイアウトができるとしている.

グラフレイアウトに Fisheye view を導入した例では,[6] がある. これは各ノードの位置を, 注視点からの距離に応じて変化させて注視点近辺を拡大表示するものであり, グラフの構造を利用したものではない.

* Kentaro Fukuchi, 東京工業大学 情報理工学研究科 数理・計算科学専攻, Masashi Toyoda and Masaru Kitsuregawa, 東京大学生産技術研究所

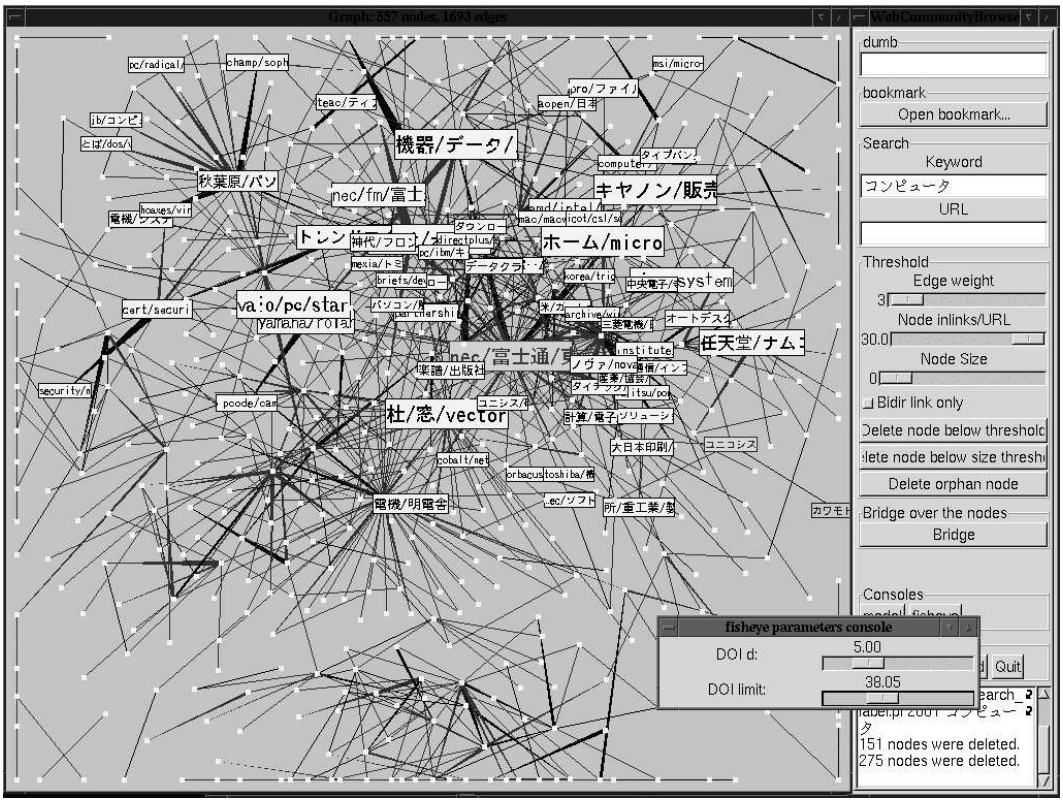

図 1. Web Community Browser 画面例

3 Web コミュニティチャート

我々は [1] で提案する手法により, ロボットで収集した国内の Web ページ 4000 万ページを基にして, 日本の Web ページの分類を行っている. この手法は, リンク集のようなページの中で列記されるページ同士には関連があるものとし, 強い関連をお互いに持つページ同士は一つの **Web コミュニティ**としてまとめるものである. 詳しいアルゴリズムについては文献 [1] を参照されたい.

抽出された Web コミュニティは一般に, 同じトピックを共有するページの集まりである事が確認されている. また, 例えばコンピューター企業の集合とユーザーコミュニティといったように, その属性が異なる場合は異なるコミュニティとして分割される傾向にある事が確認されている.

同手法では,Web コミュニティ間の関連を, 重みつきの有向エッジとして表わす. 有向エッジになる理由は, 例えば二つのコミュニティ A と B があり, A に含まれるページを指すリンク集では B に含まれるページも一緒に扱われる事が多いが, B に含まれるページを指すリンク集には A のページはあまり扱われないといった場合には, A から B への関連リンクがあるものとし, 逆方向のリンクは与えないためである.

Web コミュニティとコミュニティ間の関連をあわせた有向グラフを, 我々は **Web コミュニティチャート**と呼ぶ. 今回は 2001 年 10 月に収集した 4000 万ページを元に抽出したデータを使ってコミュニティチャートを生成した. その内訳は, コミュニティが 129537 個, コミュニティ間のリンクは 1120227 本となっている. なお,4000 万ページのうち, どこからもリンクされていないページやどのコミュニティにも属さなかったページは省かれている. コミュニティに属するページの総数は 64 万ページとなっている.

3.1 用語

各コミュニティが持つ関連リンクのうち, 他のコミュニティからそのコミュニティへと張られたリンクを, 元のコミュニティの **inlink**, その逆のリンクを **outlink** と呼ぶ. 二つのコミュニティ間に inlink か outlink またはその両方がある場合, その二つのコミュニティ間のリンクをまとめて**エッジ**と呼ぶ.

各コミュニティの重要度を表わす数値を**スコア**と呼ぶ. 本研究の実装では, コミュニティの inlink の数を, コミュニティに含まれるページの数で割って正規化した値を用いた. スコアの算出方法は議論の余地があろう.

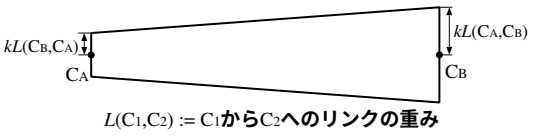

図 2. コミュニティとエッジの可視化

4 Web Community Browser

我々はWebコミュニティチャートを可視化し, 閲覧･探索するためのツール **Web Community Browser** を構築した. 図 1 にその画面例を示す. 画面左側はチャートを可視化したものが示されており, ユーザは表示されているグラフをマウスで直接操作・編集する事ができる. 右側にはグラフのレイアウトに関する各種パラメータを操作したり, 検索・編集をするためのインターフェースが置かれている. 個々のコミュニティに含まれている Web ページは, 別画面にある Web ブラウザによって確認する事ができる.

4.1 可視化技法

Web Community Browser では, 各コミュニティをラベル付けされた矩形で, エッジを等脚台形で表わす.

コミュニティの表示

各コミュニティは, ラベル付けされた矩形で提示される. 矩形の大きさは, コミュニティのスコアに比例させて大きくする事で, 有力なコミュニティを目立たせている. ラベルは, コミュニティに含まれるページへのリンクに付されたテキスト (アンカーテキスト) から自動生成している.

コミュニティが混み入っている箇所では, ラベルがエッジを隠してしまうためコミュニティ同士の関係構造が見えにくくなる. そこで, ユーザーの指示に応じてラベルの描画を省略し,5 ピクセル四方の矩形で置き換えて表示する. 詳しくは 5 章で述べる.

エッジの表示

一般に有向グラフを可視化する際には, エッジは矢印で表わされる. しかし,Web コミュニティチャートを操作する過程では inlink を多く持つコミュニティが表われる事が多く, 矢印を描画した場合はそうしたコミュニティ近辺の描画が混み入ったものになり, 線分の向きがユーザーには把握し難くなりやすい.

また, コミュニティA とコミュニティB を結ぶエッジは,A から B への関連リンクの重みと B から A への関連リンクの重みの, 二つの値を持つ. Web コミュニティチャートでは多くの場合, 二つの値は不均衡である事がわかっている. こうした非対称な関係を可視化するために,Web Community Browser ではエッジを等脚台形で表わす事により, エッジの向きとその重みとを同時に表わしている (図 2).

台形の一方の底の中心はコミュニティA の中心に位置し, その底の長さは B から A への関連リンクの重みに比例する. 同様に, 台形のもう一方の底の中心はコミュニティB の中心に位置し, その底の長さは A から B への関連リンクの重みに比例する.

なお, 双方向のリンクを含むエッジと片方向のリンクのみをエッジとで, エッジの色を変えて識別できるようにしている. こうしたエッジは全エッジの半数以上を占める事が多く, 有益な情報となる.

エッジの長さは, 両端のコミュニティのスコアの最小値に適当な係数を掛けた値を採用した. スコアの高いコミュニティ同士を離して配置する事で, グラフの局所的な密度を低下させている.

グラフレイアウト

表示されているグラフはバネモデルにより配置される. すなわち, コミュニティ間には斥力が働き, エッジはバネとして扱い, 設定されたエッジの長さに近付けるための力が働く. バネモデルによるレイアウト処理は,Web Community Browser の実行中常に行い, その過程は動的に提示する. また, コミュニティの移動速度には上限を設けた. これは, これはユーザーによるグラフの編集があった際に, 急激なグラフの形状の変化によりユーザーが注目している構造を見失うのを防ぐ為である.

Web Community Browser では表示画面をブロック分割して管理している. こうする事で, コミュニティ間の斥力計算の負荷を大幅に軽減させられる. 現在の実装では 1000 コミュニティ程度のチャートであればストレスをあまり感じさせる事なく計算・描画する事ができる[1]. また,XGA サイズの画面を 3×2 に配列した,3072×1536 ピクセルの画面でも充分な速度で動作している[2].

4.2 閲覧・探索支援機能

グラフの生成

Web Community Browser は, 以下の情報に基いた部分グラフ生成機能を提供する.

キーワード検索 各コミュニティについて, アンカーテキストに基いたキーワードが抽出されている. ユーザーがキーワードを入力すると, そのキーワードを含んだコミュニティが表示される.

ブックマーク ユーザーが普段管理している Web ブラウザのブックマークを読み込み, ブックマークに登録されている URL を含んでいるコミュニティを表示する. 現在は Mozilla により生成されたもののみサポートしている.

[1] PentiumIII500MHz 機を使用
[2] Xeon1.7GHz2CPU 構成機を使用

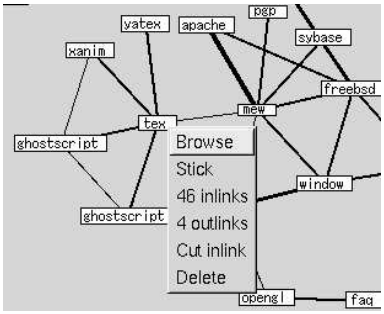

図 3. inlink/outlink 展開の選択. メニューには新たに追加されるコミュニティの数が示される.

グラフ操作

ユーザーは表示されているコミュニティを自由に動かす事ができる. バネモデルによるレイアウトだけでは最適な配置を得るのは難しく, ユーザーの判断でレイアウトに手を加えられる機能が必要である.

ユーザーは任意のコミュニティを固定する事ができる. コミュニティやエッジが多くてグラフが混みあっている場合, いくつかのコミュニティを固定してから拡げる事で, グラフの密度が下がり, 見易くなる.

グラフの展開

Web コミュニティチャートの解析では, 興味のあるコミュニティの周辺にどんなコミュニティがあるかが関心の対象となる. 前述の機能だけでは全ての周辺コミュニティが表示されているとは限らない. そのため, 周辺コミュニティを追加表示する機能として,inlink/outlink 展開を実装した. inlink 展開は, 指定したコミュニティへの関連リンクを持つコミュニティを, outlink 展開は, 指定したコミュニティがリンクしているコミュニティを追加表示する.

また, 初期グラフではエッジを持たない孤立コミュニティが現われる事が多い. こうした孤立コミュニティとその他のコミュニティとの関連を調べる為に, ブリッジ展開機能を実装した. ある二つのコミュニティにおいて, 直接にはリンクを持たないが別のコミュニティを介して関連するような場合, そのコミュニティを追加表示する.

エッジの除去

Web Community Browser では, 表示されているコミュニティ間に存在するエッジは全て表示する. コミュニティ数に対してエッジ数が過度に多い場合, バネモデルの特性によりグラフ全体が小さく縮まる傾向がある. 見易さの面からも, エッジを適当に除去する必要がある. Web Community Browser では, エッジの重みに閾値を設け, 閾値以下の重みのエッジを除去する機能を持つ. この操作は可逆であり, 閾値を下げると, 除去されたエッジは元に戻る.

また, 全てのエッジの中から, 双方向にリンクを持つもののみ残し, 片方向リンクのエッジを除去する事ができる. 一般にコミュニティ間のリンクが双方向リンクであれば, それらのコミュニティは同じトピックを共有する, 強い関連性のあるコミュニティである場合が多い.

検索エンジンからなるコミュニティのように, 普遍的な有名サイトを多く含んだコミュニティは非常に多くの inlink を持つ. しかしこれらの inlink はユーザーにとって意味のある情報ではない場合が多く, 一般には表示させる意味がない. そこで, ユーザーはそうしたコミュニティに対し,inlink の表示を抑制させる事ができる.

これらの除去機能は, 前節で説明した周辺コミュニティの展開機能に対しても働いており, 例えばエッジの重みに閾値を設けた状態で inlink 展開をすると, 閾値以下の重みの inlink を持ったコミュニティは展開されない.

5 Fisheye view の応用

コミュニティ数及びエッジ数が多いグラフでは, ラベルがエッジを覆い隠してしまい, コミュニティ同士の関係構造が見えにくくなる事が多い. しかし, ラベルの描画は省略して, ユーザーがマウスで指したコミュニティのみラベルを描画するという手法では, ユーザー注目するコミュニティの周辺のコミュニティとの関連を読み解くのは困難である.

そこで本研究では, ユーザーが注目するコミュニティと関連の強いコミュニティのみラベルを表示し, それ以外のコミュニティのラベルは省略表示するという手法を導入して, この問題の解決を試みた. 以下の手法で対象グラフ内のコミュニティの **DOI** (Degree Of Interest: 注目度) を求め, その値がユーザーが設定した閾値を越えたコミュニティを, 注目されているコミュニティと関連の強いコミュニティとしてラベルを表示する. なお, ここでいうユーザーが注目しているコミュニティとは, カーソルを当てたコミュニティを指す.

まず, 注目されているコミュニティ v の DOI を適宜定める (ここでは 100 とする). v との間にエッジを持つコミュニティ u は以下の式で計算する.

$$DOI_u = DOI_v \frac{w(u,v) + w(v,u)}{d + w(u,v) + w(v,u)} \qquad (1)$$

$w(u,v)$ は u から v へのリンクの重み
d は減衰率

u の周辺コミュニティについても同様に DOI を計算する. DOI が閾値未満になったらそれ以上の探索は打ち切る.

また, 注目コミュニティ周辺の描画密度を低下させるために, 注目コミュニティは斥力の係数を高く

設定する. こうする事で, 注目コミュニティから画面上で近距離にあるコミュニティは注目コミュニティから離れ, 結果として描画密度が低下する.

図 4 は,「野球」というキーワードを含むコミュニティを対象に, Fisheye view の効果を示したものである. 画面 (a) は, ラベルを全て表示させたもの, 画面 (b) は全てのラベルを省略表示したものである. 画面 (c)(d) はともに, プロ野球団公式ページからなるコミュニティ(画面右上) にフォーカスをあて, (c) では DOI の閾値を 35 に,(d) では 25 にそれぞれ設定して表示したものである.

ラベルが全て表示されている状態ではグラフ構造の一部が隠れている. DOI を有効にした図では, フォーカスされたコミュニティの周辺はラベルが表示されているが, 離れたコミュニティでは, ハブになっているコミュニティとその周辺の一部のコミュニティのラベルが表示されている. 特に, 画面左下には少年野球チームのコミュニティ群があり, (c) ではそのハブとなっているコミュニティのみが表示されているのに対し, 閾値を下げた (d) では周辺コミュニティの一部が加わっているのがわかる.

マルチフォーカスへの拡張

前節では, 注目しているコミュニティを一つとし, 式 (1) により DOI を求めた. これをマルチフォーカスへ拡張する手法について述べる.

$V = \{v_1, v_2, ..., v_n\}$ をフォーカスされているコミュニティの集合とする. フォーカスされているコミュニティ全てに同じ DOI を与え, それらのコミュニティから出発して, エッジを辿って見付かるコミュニティの DOI を式 (1) により求める. コミュニティ u の, コミュニティ v_i を元にして計算した DOI を, $D(u, v_i)$ と置く. コミュニティ u の最終的な DOI を,

$$D(u) = \max_{i=1}^{n} D(u, v_i)$$

として計算する.

6 まとめと今後の課題

Web コミュニティチャートを可視化し, 閲覧・探索を支援するツール,Web Community Browser を構築した. まずユーザーのブックマークやキーワード検索機能によりユーザーが興味関心を持つコミュニティ群を提示する. 一般的な検索エンジンは, キーワードにマッチしたページを単発的に提示するが, Web コミュニティチャートを基にした本ツールでは, コミュニティという形で結果を提示するので, 周辺情報の探索を容易にしている. また, 主要サイトとそのファンサイトというような関連が図示されているので, ユーザーは, その web ページが提供するであろう情報の質を把握しながらブラウジングできる.

ユーザーは inlink/outlink 展開やブリッジ展開機能により, 表示されているコミュニティの周辺のコミュニティを追加表示させていく事ができ, authority コミュニティや hub コミュニティの発見に役立つ. また, 各種閾値を調節する事で, 重要なグラフ構造を浮き立たせる事ができる. さらに,Fisheye view を応用したラベル表示の抑制により, これまでよりも, 関心のあるコミュニティの周辺コミュニティに特化した表示を得る事ができるようになった. DOI の伝播は, 現在の実装ではエッジの向きを無視しているが, inlink または outlink のみを辿らせた場合には異なる結果を得る事が予想される. こうした制約の指定や,DOI の閾値の指定も同時にできるようなインターフェースを加える事を予定している.

Web Community Browser を使ったグラフ編集過程で, 初期配置でコミュニティ数が多かった場合は, 不要コミュニティの除去が作業の中心となる. 現状では, エッジ重みの閾値を調整して, 孤立したコミュニティを除去する作業が主であり, 望みの部分グラフを得るのは困難である. DOI を用いて注目する部分グラフの抽出ができているので, これを部分グラフの編集に用いる事が考えられる. 例えば, DOI の閾値を調整して, 適当な大きさの部分グラフになったら他のコミュニティを除去するといった操作が挙げられる.

参考文献

[1] Masashi Toyoda and Masaru Kitsuregawa. Creating a Web Community Chart for Navigating Related Communities. *In proceeding of Hypertext 2001*, pp. 103-112, 2001.

[2] 福地 健太郎, 豊田 正史, 喜連川 優. Web Community Browser: 大規模 Web コミュニティチャートの可視化. 第 13 回データ工学ワークショップ (DEWS2002) 論文集, 2002.

[3] Kamada, T. and Kawai, S. An algorithm for drawing general indirect graphs. *Information Processing Letters*, No. 31, pp. 7-15, 1989.

[4] George Furnas. Generalized Fisheye Views. *In Proceedings of CHI'86*, pp. 16-23, 1986.

[5] Tyson Henry and Scott Hudson. Interactive Graph Layout. *In Proceedings of UIST'91*, pp. 55-64, 1991.

[6] Manojit Sarkar and Marc H. Brown. Graphical fisheye views of graphs. Technical Report 84, Systems Research Center, 1992.

[7] Tamara Munzner. Drawing large graphs with h3viewer and site manager. *In Proceedings of the 6th Graph Drawing*, pp. 384-393, 1999.

[8] Mao Lin Huang and Peter Eades. WebOFDAV - Navigating and Vsualizing the Web On-line with Animated Context Swapping. *In Proceedings of the 7th World Wide Web Conference*, pp. 636-638, 1998.

[9] 舘村 純一. DocSpace: 文献空間のインタラクティブ視覚化. インタラクティブシステムとソフトウェア IV. pp. 11-20, 近代科学社, 1996

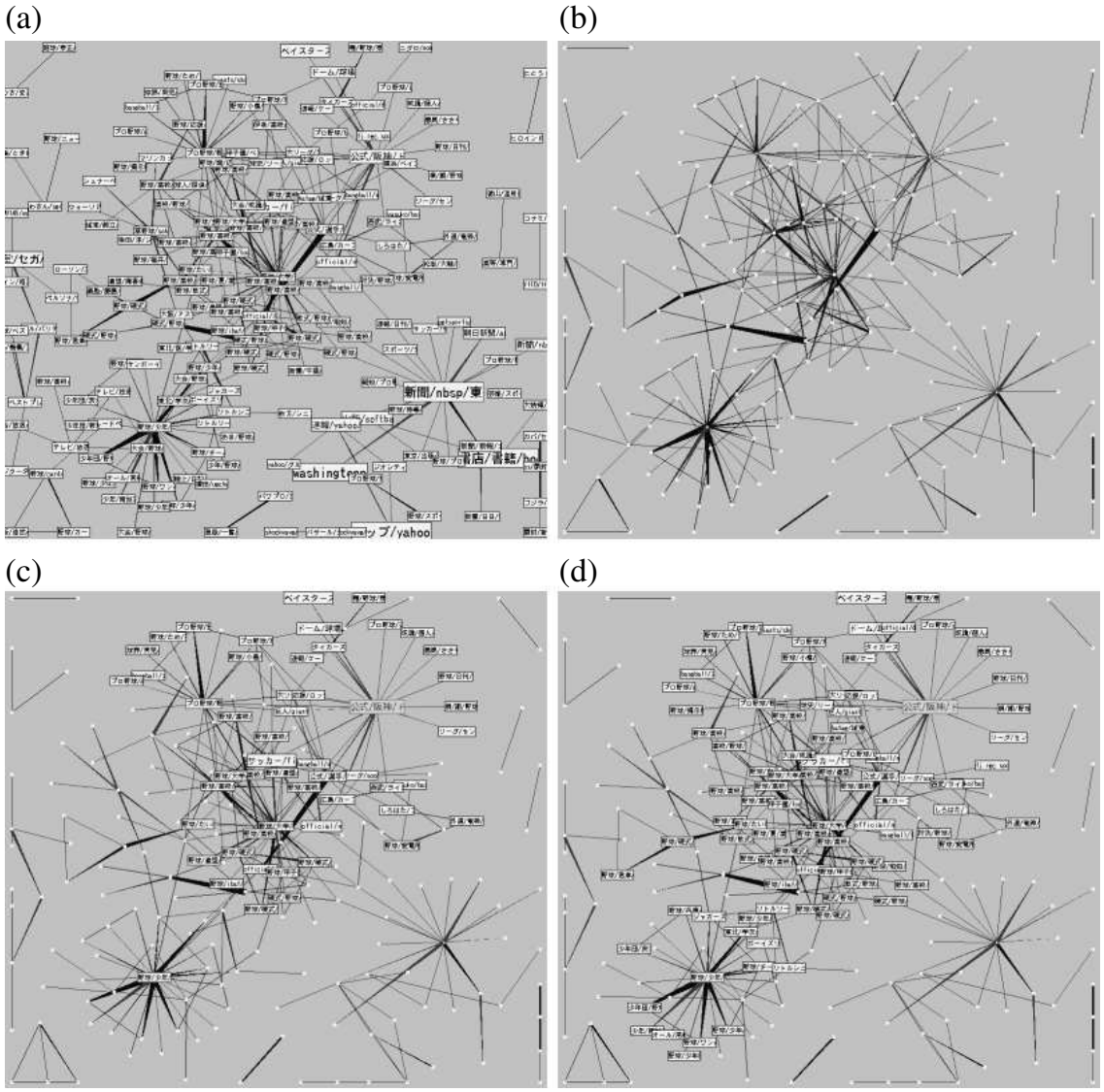

図 4. ラベル表示量の調整例: (a) 全ラベルを表示,(b) 全ラベルを省略表示, (c)(d) はプロ野球団公式ページのコミュニティにフォーカスしたもので, (c) は DOI の閾値を 35 に,(d) は 25 に設定した.