Web Community Browser: 大規模 Web コミュニティチャートの可視化

福地　健太郎[†]　豊田正史[††]　喜連川　優[††]

Web コミュニティとは、ある共通するトピックを持った Web ページの集合である。我々はすでに国内 1400 万ページからリンク解析により 7 万個のコミュニティを抽出している。Web コミュニティチャートは、WWW 上のコミュニティ群とその関連を可視化したものである。今回我々は、こうして得られたコミュニティ群を可視化し、閲覧・探索を支援するツール「Web Community Browser」を構築した。Web Community Browser は、ユーザーが指定したコミュニティと関連の強いコミュニティを提示し、関連コミュニティ同士の関係性の発見を支援する。また、ユーザーのブラウジング履歴やブックマークを元にコミュニティリストを生成し、ユーザーの興味に合致するページ群を提供する。

論文番号 A1-4

1. 背　　景

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コミュニティを抽出する手法を研究している。4)，5) で提案した手法は、Web ページ群のリンク情報を解析するもので、現在までに国内 1700 万ページを元に、7 万個のコミュニティを発見している。

我々は今回、上記の手法で得た Web コミュニティ群を可視化し、それらを閲覧・探索する為のツール「Web Community Browser」を構築した。本ツールを使用する事で、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ができる。

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 Vierer[3] 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[2] 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。

† 東京工業大学情報理工学研究科数理・計算科学専攻
Tokyo Institute of Technology, Graduate School of Information Science and Engineering
†† 東京大学生産技術研究所
Tokyo University, Institute of Industrial Science

2. Web コミュニティチャート

我々は 4)，5) で提案する手法により、ロボットにより収集した国内 1700 万ページを基にした Web コミュニティチャートを自動的に作成している。

一つのコミュニティは、リンク解析の結果同じトピックに関心をもつ人々や団体によって作成された web ページの集合であり、二つ以上のページ (URL) から構成される。

コミュニティ間の関係は、コミュニティをまたぐリンクの数を採用している。コミュニティ A のページからコミュニティ B のページにリンクが n 本ある場合、A は B を参照するリンクを持ち、その重みは n であるとして扱う。

コミュニティと、コミュニティ間のリンクをあわせて、Web コミュニティチャートと呼ぶ。今回は 2000 年 10 月に収集した 1700 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 67949 個、コミュニティ間のリンクは 577423 本となっている。

3. Web Community Browser

Web Community Browser は、Web コミュニティチャートの部分集合を可視化し、ユーザーによる閲覧・探索を支援するツールである。図 3に画面例を示す。画面左側に Web コミュニティチャートを可視化したものが示される。画面右側に、閲覧・探索を支援するための操作パネルが置かれている。

次節でまず Web コミュニティチャートの可視化手法について説明し、その後に閲覧・探索の支援機能につい

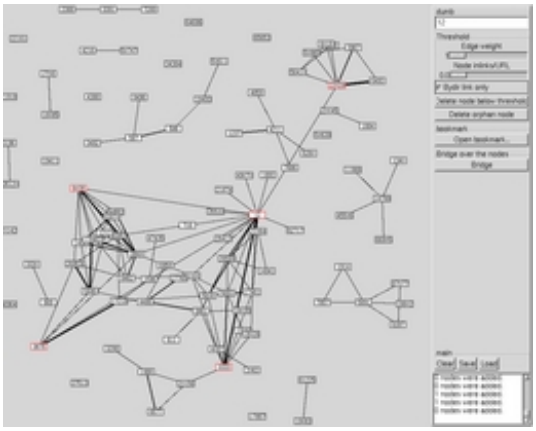

図 1　Web Community Browsr 画面例

て述べる。

3.1　Web コミュニティチャートの可視化

Web Community Browser では、各コミュニティをノード、コミュニティ間のリンクをエッジとした二次元無向グラフとして扱う。コミュニティ間のリンクは片方向だけであっても、ノード間にはエッジがあるものとする。ただし、後述する機能により、双方向にリンクがある場合にのみエッジを張るといった操作も可能である。

グラフはバネモデル[1)] を用いて配置を最適化する。バネモデルではノードを質点、エッジある長さを持ったバネとして扱い、力学モデルに従って反復計算する事で適当なグラフ配置を求める。今回は各ノードは全て同じ性質を持ったものとして扱う。エッジの長さは、リンクの重みから決定する。コミュニティ g に含まれる URL の数を $W(g)$、コミュニティ p からコミュニティ q へのリンクの重みを $L(p,q)$ で表すと、コミュニティ A,B 間のエッジの長さ $E(A,B)$ は次式で決定する。

$$E(A,B) = k \min(\frac{L(p,q)}{W(q)}, \frac{L(q,p)}{W(p)})$$

kは適当な係数(単位ピクセル)

こうする事で、in-link の多いコミュニティ同士は離れて配置され、グラフの局所的な密度が低下される。なお、エッジの長さには下限 (30 ピクセル) を設けている。

こうして決定されたグラフを、バネモデルを用いて配置する。反復計算はプログラムの実行中常に行われ、その過程は動的に提示される。 Web Community Browser ではブロック分割を施して計算負荷を軽減させており、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できる☆。

各ノードは、ラベル付けされた矩形で提示される。ラベルは、コミュニティ内のページを指すリンクのアンカーテキストを基に自動生成したものを採用したが、ユーザーがラベルの内容を書き換える事もできる。各エッジは線分で示される。線分の太さは、リンクの重みに応じて太く表示する。エッジの向きは表示には反映していない。エッジを矢印で提示する事も試みたが、エッジの本数が増えるに従って見易さを損うと感じたためである。

3.2　閲覧・探索支援機能

3.2.1　グラフの生成

Web Community Browser では Web コミュニティチャートの一部を表示・閲覧できる。ユーザーはまず最初に表示させるグラフを指示する必要がある。 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　各コミュニティは、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 Netscape Navigator により生成されたもののみサポートしている。

ブラウジング履歴　ユーザーが普段使用している Web ブラウザのブラウジング履歴に基き、最近閲覧した URL を含むコミュニティを表示する。現在は Netscape Navigator のみサポートしている。

これらの機能により初期グラフが提示される。それぞれのコミュニティは複数の Web ページを含んでいるので、ユーザーの興味に合致し、なおかつユーザーにとって未知であるページを提供できる。

3.2.2　グラフ操作

ユーザーは表示されているグラフのノード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。

ユーザーは任意のノードを固定する事ができる。ノードやエッジが多くてグラフが混みあっている場合、いくつかのノードを固定してから拡げてやる事で、グラフの密度が下がり、見易くなる。

ユーザーはコミュニティに含まれるページを、Netscape Navigtor を通じて閲覧できる。ノードを右クリックして “Browse” を選択すると、コミュニティ内のページへのリンクを並べたウィンドウと、その中の上から 4 つのページを分割表示したウィンドウが表示される。また、このページから、コミュニティにつけられるラベルを編集する事ができる。

3.2.3　グラフの展開

Web コミュニティチャートを眺めていると、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気になる。前述の機能だけでは全ての周辺コミュニティ

☆ PentiumIII500MHz 機を使用

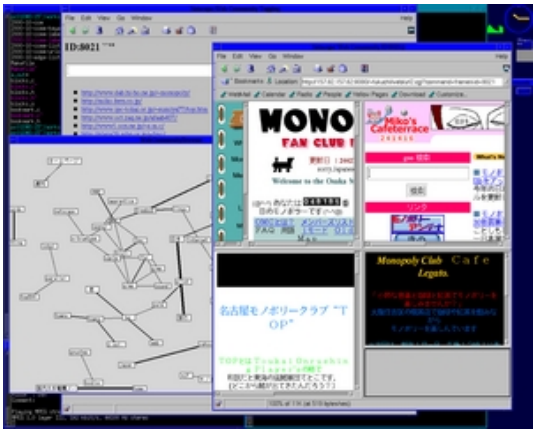

図 2　Web ブラウザでコミュニティを表示

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、in-link/out-link 展開を実装した。 in-link 展開は、指定したコミュニティへのリンクを持つコミュニティを、out-link 展開は、指定したコミュニティが参照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ノードが表われる事が多い。こうした孤立ノードとその他のノードとの関連を調べる為に、ブリッジ展開機能を提供する。ある二つのノードは直接にはリンクを持たないが、別のノードを介して関連するような場合、そのノードをブリッジノードと呼ぶ。ブリッジ展開機能はこうしたブリッジノードを探して追加表示する。

ブリッジ展開のアルゴリズムを述べる。まず表示されているノード全てに対して in-link 展開と out-link 展開をする。次に、新たに追加されたノードのうち、二つ以上の既存ノードへのリンクを持つノードを残し、その他の追加ノードを除去する。また、孤立ノードへのリンクを持たないノードも除去する。こうする事で、ブリッジノードを得る事ができる。なお、二つの孤立ノードに対し、ブリッジノードが複数ある場合に、数が多過ぎて問題となる事がある。これらは何らかの基準で選別除去する必要があり、今後の課題である。

3.2.4　ノード・エッジの除去

Web Community Browser では基本的に、表示されているノード間に存在するエッジは全て表示する。ノード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、除去する機能を持つ。

また、全てのエッジの中から、双方向リンクのもののみ残し、片方向リンク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

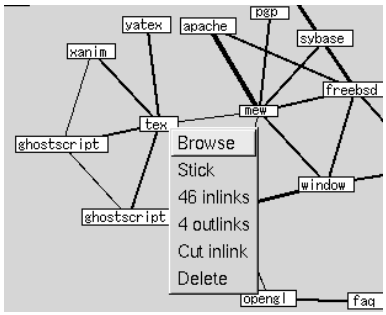

図 3　in-link/out-link 展開の選択。新たに追加されるノードの数が示される。

のあるコミュニティである場合が多い。また、検索エンジンのような有名サイトへのリンクはユーザーにとってあまり意味のある情報ではなく、これを除去する意味は大きい。

ユーザーは、表示するノードを抑制する事ができる。各コミュニティは in-link の本数を、含まれるページの数で割ったものをスコアとして持つ。スコアに閾値を設け、閾値より低いスコアのコミュニティを除去する事ができる。

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 in-link 展開をすると、閾値以下の重みの in-link を持ったコミュニティは展開されない。また、閾値操作は可逆であり、閾値を下げると、除去されたエッジやノードは元に戻る。

3.2.5　特定コミュニティへの操作

検索エンジンからなるコミュニティのように、有名サイトを多く含んだコミュニティは非常に多くの in-link を持つ。しかしこれらの in-link 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、in-link の表示を抑制させる事ができる。

4.　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 in-link/out-link 展開や bridge 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 コミュニティは 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。

現在の実装では、全てのノードはその内容に関わらず同等に扱っている。例えば含まれるページ数であるとか、スコアを表示やグラフレイアウトに反映させる事が考えられる。また、コミュニティは一つのノードとして扱っているが、コミュニティ内部にはどのようなページがあり、どのページにリンクが張られているかが知りたいという要求があり、コミュニティ内部の詳細を表示させる機能が求められている。

グラフ密度が濃くなるとノード同士が接近し、また、エッジが交差しあって見易さを損う。現在の実装では 1280x1024 ピクセルの画面一杯にウィンドウを拡げた状態で、500 ノード以上表示させると見易さを著しく損う。Web コミュニティチャートを生成する段階で、クラスタリングの階層化を施したり、可視化の段階で Focus+Context 技術を導入する等の処置が必要であろう。

参 考 文 献

1) Eades, P.: A heuristic for graph drawing, *Congressus Numerantium*, Vol. 42, pp. 149–160 (1984).
2) Huang, M. L. and Eades, P.: WebOFDAV - Navigating and Vsualizing the Web On-line with Animated Context Swapping, *Proceedings of the 7th World Wide Web Conference*, pp. 636–638 (1998).
3) Munzner, T.: Drawing large graphs with h3viewer and site manager, *Proceedings of the 6th Graph Drawing*, pp. 384–393 (1999).
4) Toyoda, M. and Kitsuregawa, M.: Finding related communities in the Web, *Poster Proceeding of 9th International WWW Conference*, pp. 70–71 (1999).
5) Toyoda, M. and Kitsuregawa, M.: A Web community chart for navigating related communities, *Poster Proceeding of 10th International WWW Conference*, pp. 62–63 (2000).