

Web Community Browser: Web コミュニティ構造の可視化と探索機構の実現

Web Community Browser: a browsing tool of a large Web community chart

福地 健太郎[†] 豊田 正史^{††} 喜連川 優^{††}

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コミュニティを抽出する手法を研究している。2) 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、2001 年 10 月の時点で、国内 4000 万ページを元に、13 万個のコミュニティを発見している。また、個々のコミュニティ間の関連も取得することができる。

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ことができる。

2. Web コミュニティチャート

我々は文献 2) で提案する、Web ページ間のリンク解析を基にした手法により、ロボットにより収集した国内の Web ページを基にした Web コミュニティ抽出を行っている。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 2) 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、左記手法では、Web コミュニティ間の非対称な関連を得ている。非対称な関連とは、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。こうしたコミュニティ間の関連を含めたものを、我々は Web コミュニティチャートと呼ぶ。

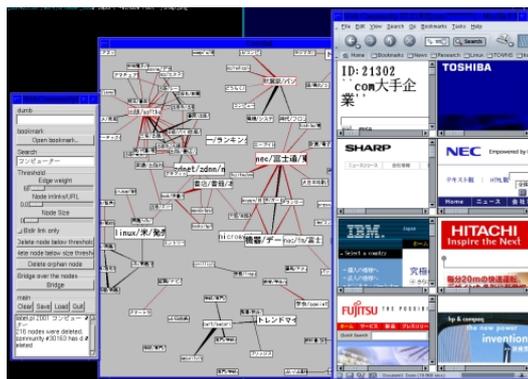


図 1 Web Community Browser 画面例

2001 年 10 月には、国内 4000 万ページを基に Web コミュニティチャートを作成した。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。

3. Web Community Browser

Web Community Browser(WCB) 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 1 に画面例を示す。画面中央に Web コミュニティチャートを可視化したものが示される。画面左側に、閲覧・探索を支援するための操作パネルが置かれている。ユーザーはいくつかのパラメータを操作したり、表示するコミュニティを追加・削除してグラフを適宜編集して、関心のあるコミュニティの周辺構造を調べる事ができる。コミュニティに含まれるページは、画面右側にある Web ブラウザで確認することができる。

3.1 Web コミュニティチャートの可視化

WCB では、各コミュニティをノード、コミュニティ間の関連リンクをエッジとした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であっても、ノード間にはエッジがあるものとする。

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。エッジで結ばれたノード同士は各ノード間は接近し過ぎないように、斥力が働く。

反復計算はプログラムの実行中常に行い、その過程は

[†] 東京工業大学情報理工学研究所 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

動的に提示する。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。

各コミュニティは、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。矩形の大きさは、コミュニティへの 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に比例させて大きくしている。

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。多くの場合、二つの値は不均衡である事がわかっている。非対称な関係を可視化するために、各エッジは線分ではなく、等脚台形で示す。

なお、A から B への関連リンクはあるが B から A への関連リンクはない場合には、エッジの色を変えて識別できるようにした。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。

3.2 閲覧・探索支援機能

WCB 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、キーワード検索により指示する。キーワード検索は、各コミュニティのアンカーテキストの集合を対象とする。これらの機能により初期グラフが提示される。表示されたコミュニティ間に存在する関連リンクは全て表示する。

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し、配置を修正する事ができる。各コミュニティに含まれるページは、図 1 に示すように、Web ブラウザに表示される。

WCB は、ユーザーが関心を持つコミュニティの周辺コミュニティを探索するための機能として、inlink/outlink 展開を提供する。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。

ノード数に対してエッジ数が過度に多い場合、パ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。WCB では、エッジの重みに閾値を設け、除去する機能を持つ。

4. 使用事例

図 2 は、コンピューターに関連したコミュニティの構造を発見した例である。中央付近の「nec/富士通」とラベル付けされたコミュニティは、NEC・富士通・東芝・SHARP 等、国内のコンピューター系大手企業からなる。周囲には、「機器/データ」のラベルで表わされている、コンピューター関連の周辺機器を扱うメーカーのコミュニティや、「ホーム/micro」とラベル付けされた、大手ソフトウェア企業 (Microsoft・Adobe・ORACLE 等) からなるコミュニティが周囲にある。

画面下部には「秋葉原/パソ」というコミュニティがあるが、これは主に秋葉原に店舗を持つパソコン系

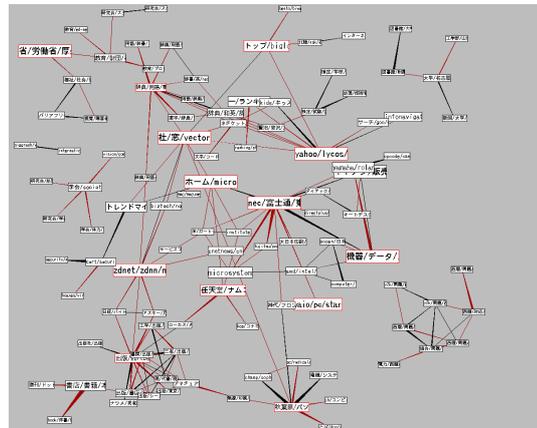


図 2 コンピューター関連のコミュニティの構造図

ショップからなる。その周囲には、やや規模の小さいショップのコミュニティが現われている。

画面左下には、大手出版社のコミュニティや、コンピューター系雑誌、書店等のコミュニティが固まって現われているが、それらはコンピューター系企業のコミュニティ群へのエッジはそれ程強くなく、重み閾値が 4 の状態では消えている。しかし、その上部にある「zdnet/zdnn」とラベル付けされた、情報系ニュースサイト (ZDNet・CNET・PCWatch 等) のコミュニティを経由して接続している事がわかる。

主要コミュニティ群とはこの状態では切断されているが、右下には西暦 2000 年問題に関するコミュニティ群が現われており、2001 年 10 月の段階でもまだコミュニティを為している。画面左側には情報処理学会を始めとする学術関係のコミュニティを中心に、研究会等のコミュニティが現われている。

このグラフを得るまでの手順を述べる。まず、「コンピューター」「コンピュータ」の二つのキーワードで OR 検索をする。この時点で 986 個のコミュニティが表示される。次に、エッジの閾値を 14 まで上げて、孤立ノードを全て除去した後に、再び閾値を 4 まで下げ、手で適宜レイアウトを整えた。

5. 最後に

今後の課題として、特徴のあるコミュニティ構造を自動で発見するような機能を加える事を検討している。また、コミュニティ内部のページ間の関係構造の可視化・閲覧機能を実装する事を課題とする。

参考文献

- 1) Kamada, T. and Kawai, S.: An algorithm for drawing general indirect graphs, *Information Processing Letters*, No. 31, pp. 7-15 (1989).
- 2) Toyoda, M. and Kitsuregawa, M.: Creating a Web Community Chart for Navigating Related Communities., *Conference Proceedings of Hypertext 2001*, pp. 103-112 (2001).