データベースシステム 128−63
（２００２．７．１９）

Web Community Browser における探索機構の実装と評価

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 4000 万ページからリンク解析により 13 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。可視化には力学シミュレーションを用いた、ばねモデルを応用した。

An interactino technique of Web Community Browser and its evaluation

Kentarou Fukuchi,[†] Masashi Toyoda[††]
and Masaru Kitsuregawa[††]

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**コミュニティ**を抽出する手法を研究している。4) 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事ができる。

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事ができる。

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 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[1] 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。

3. Web コミュニティチャートの抽出手法

我々は文献 4) で提案する手法により、ロボットにより収集した国内 4000 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。以下にチャートの抽出手法の手順を述べる。

まず、国内の Web サイトから Web ページをロボットにより収集し、蓄積する。ここから、各ページの URL と、そのページに含まれるアンカーのリンク先 URL のデータを得る。

次に、各ページを seed ページとして、seed ページと他のページとの関連を調べる。まず seed ページを指しているページの集合を A とする。次に、A のページが指しているページの集合を取得し、これを B する。これらの和集合 $A \cup B$ に seed ページを加えた集合と、集合内のページ間のリンクを併せたグラフの事を、本論文

† 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

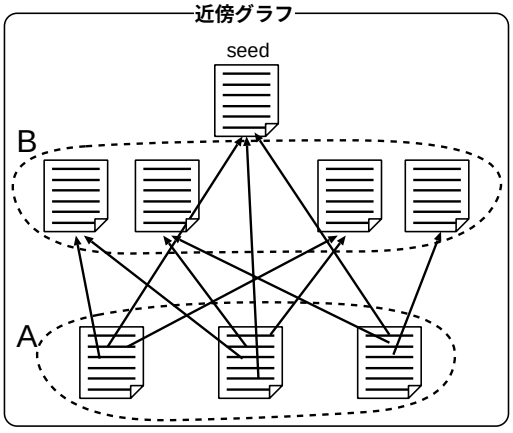

図 1 近傍グラフの定義

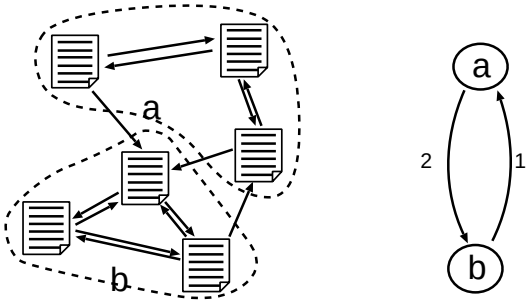

図 2 コミュニティの抽出過程: ADG から、双方向エッジで接続するページ同士を抽出し、コミュニティとする (左図)。二つのコミュニティに跨がる片方向エッジは、コミュニティ間の関連リンクとして扱う。

では seed ページの**近傍グラフ**と呼ぶ (**図 1**)。

次に、近傍グラフ内で各ページの Hub&Authority スコアを計算し、 Authority スコアの高いページは、seed ページに関連があるものとして、 seed ページから Authority ノードへ有向エッジを張る。この操作を全てのページに行って得たグラフを、 **Authority Derivation Graph**(ADG) と呼ぶ。

最後に、 ADG のうち双方向にエッジを持つページ同士をグルーピングし、これを **Web コミュニティとする** (**図 2** 左)。また、別々の Web コミュニティに含まれるページ間のエッジは、その Web コミュニティ間に関連があるものとし、**関連リンク**とする。エッジが複数ある場合は、その本数を関連リンクの重みとして扱う (図 2 右)。結果として、 Web コミュニティとその関連リンクからなる有向グラフを得る。これを **Web コミュニティチャート**と呼ぶ。

上記手法によって得られた 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た。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た。

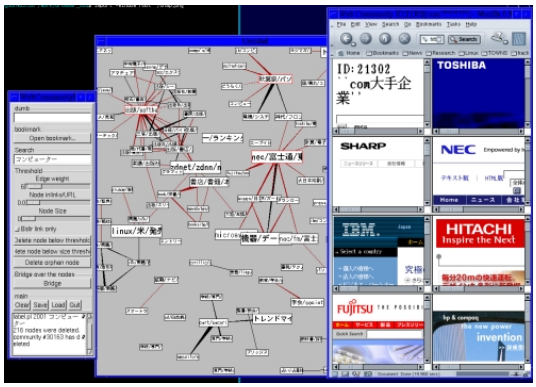

図 3 Web Community Browsr 画面例

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。

4. 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事ができる。

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て述べる。

4.1 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。

4.1.1 グラフレイアウト

グラフはバネモデル[2] 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、斥力が働く。

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、inlink 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い、その過程は動的に提示する。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。

Web Community Browser では、画面をブロック分割して、ノード間の斥力計算の負荷を軽減させている。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☆。また、XGA サイズの画面を 3×2 に配列した、3072×1536 ピクセルの画面も充分な速度で動作している☆☆。

4.1.2 コミュニティの可視化

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。まずコミュニティに含まれるページのアンカーテキスト全てを対象に、日本語形態素解析ツール JUMAN で形態素単位に分割する。それらを出現頻度の高い順に 10 個取り出して、間に '/'(スラッシュ) を挟んでつなげたものをラベルとして使用した。この際、「ホームページ」「会社」のような、多くのコミュニティに現われ、コミュニティの特性を表しにくい語は排除している。

コミュニティを表わす矩形の大きさは、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する事で、有力なコミュニティを目立たせている。

4.1.3 エッジの可視化

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。

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、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。

4.2 閲覧・探索支援機能

4.2.1 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。

これらの機能により初期グラフが提示される。

4.2.2 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、Mozilla 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、その中から 7 つのページを分割表示したウィンドウが表示される (図 3)。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。

4.2.3 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ノードが

☆ PentiumIII500MHz 機を使用

☆☆ Xeon1.7GHzx2 機を使用

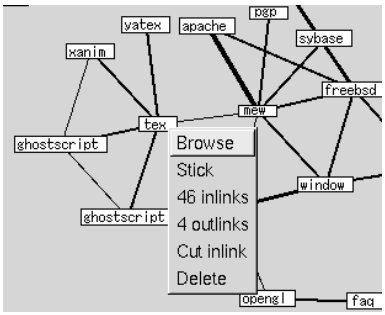

図 4 inlink/outlink 展開の選択。新たに追加されるノードの数が示される。

現われる事が多い。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 inlink 展開と outlink 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。

4.2.4 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。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 inlink 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。また、コミュニティに含まれるページの数にも閾値を設け、小さなコミュニティを削除する機能を持つ。

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。また、閾値操作は可逆であり、閾値を下げると、除去されたエッジは元に戻る。

4.2.5 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 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。

5. 使 用 事 例

5.1 コンピューター関連のコミュニティ

図 5 は、コンピューターに関係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、手動で適宜レイアウトを整えた。

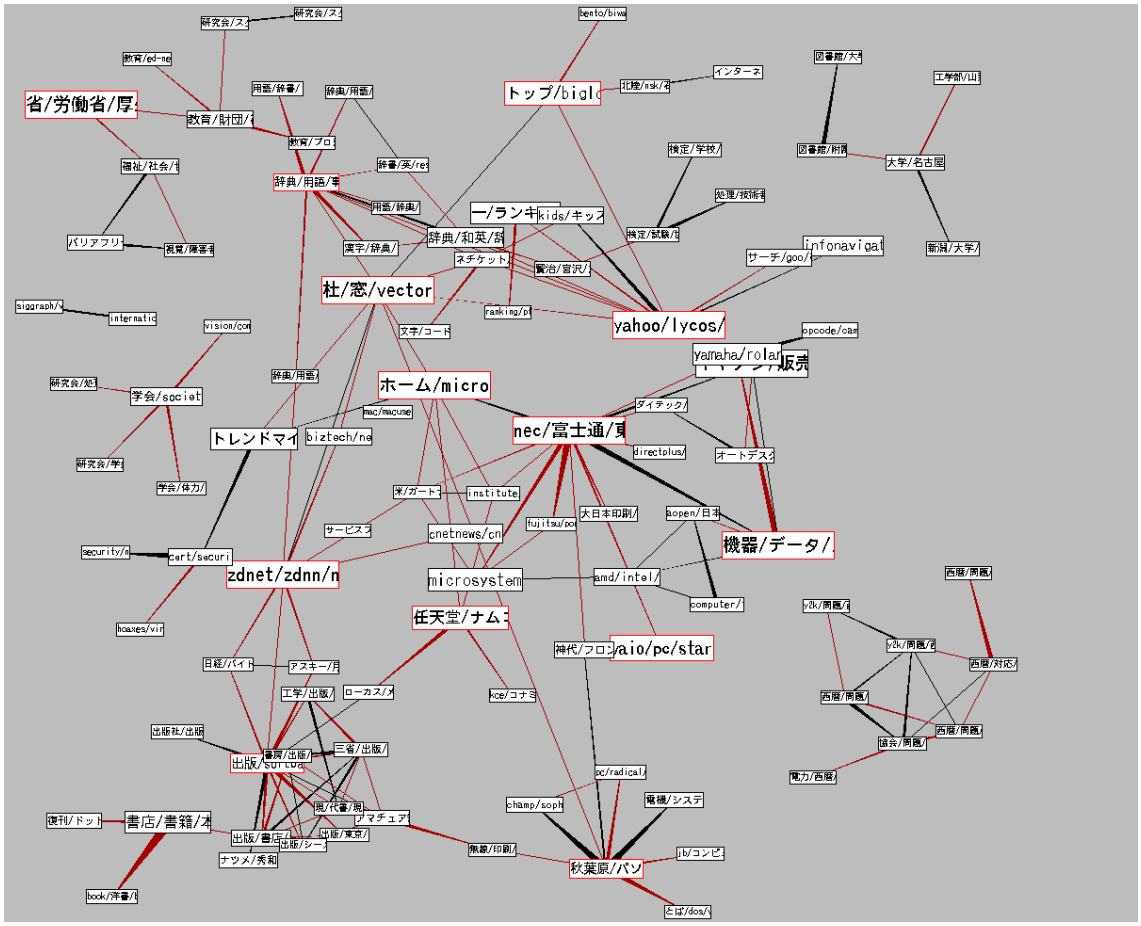

図 5　コンピューター関連のコミュニティの構造図

5.2　証券業関連のコミュニティ

図 6 は、キーワード「証券」で検索した後、エッジの重みの閾値を 9 に上げ、孤立コミュニティを除去してからレイアウトを整えた状態の図である。大きく分けて三つのスター型の構造が見えるが、左上から右下に向って、順に証券会社のコミュニティ、証券業協会等関連団体のコミュニティ、証券取引所のコミュニティとなっている。それぞれの中心的なコミュニティがあって、その周囲に規模のやや小さい、関係する団体が接続しているという構造が見てとれる。例えば証券会社のコミュニティは、中心に大手企業のコミュニティがあり、周囲は中小の証券会社のコミュニティが現われている。

このように、同じ証券というキーワードを共有するコミュニティでも、その業種や形態の違いによって、違うコミュニティに分類する事ができており、またその関連構造も抽出できている事がわかる。

5.3　プロ野球関連のコミュニティ

図 7 は、プロ野球団 12 チームの公式ページからなるコミュニティを inlink 展開した後に、双方向リンクのエッジのみを取り出し、重み閾値を 4 にして孤立コミュニティを除去した後に手動でレイアウトを整えた状態の図である。双方向リンクに限定すると、公式ページやマスコミ関連のコミュニティはほぼ孤立し、除去される。これは、こうしたコミュニティは inlink は数多く持つものの、outlink をほとんど持たない為で、一般に Authority としての度合が高いコミュニティはこうした傾向を持つ。

逆に、球団ファンのコミュニティは相互に関係しあう事が多く、双方向の関連性を持つ事が多い。その為、プロ野球関連のコミュニティで片方向リンク除去の操作を行うと、ファンコミュニティ間の関係が見えてくる。図からは、阪神・中日・広島・ダイエーのファンコミュニティはそれぞれで強固なコミュニティ間の結合を持っている事がわかる。巨人ファンコミュニティはヤクルト・横浜ファンコミュニティとは結合があるが、阪神ファンコミュニティとは繋っていない。

6.　議　　論

Web Community Browser を使ったグラフ編集過程

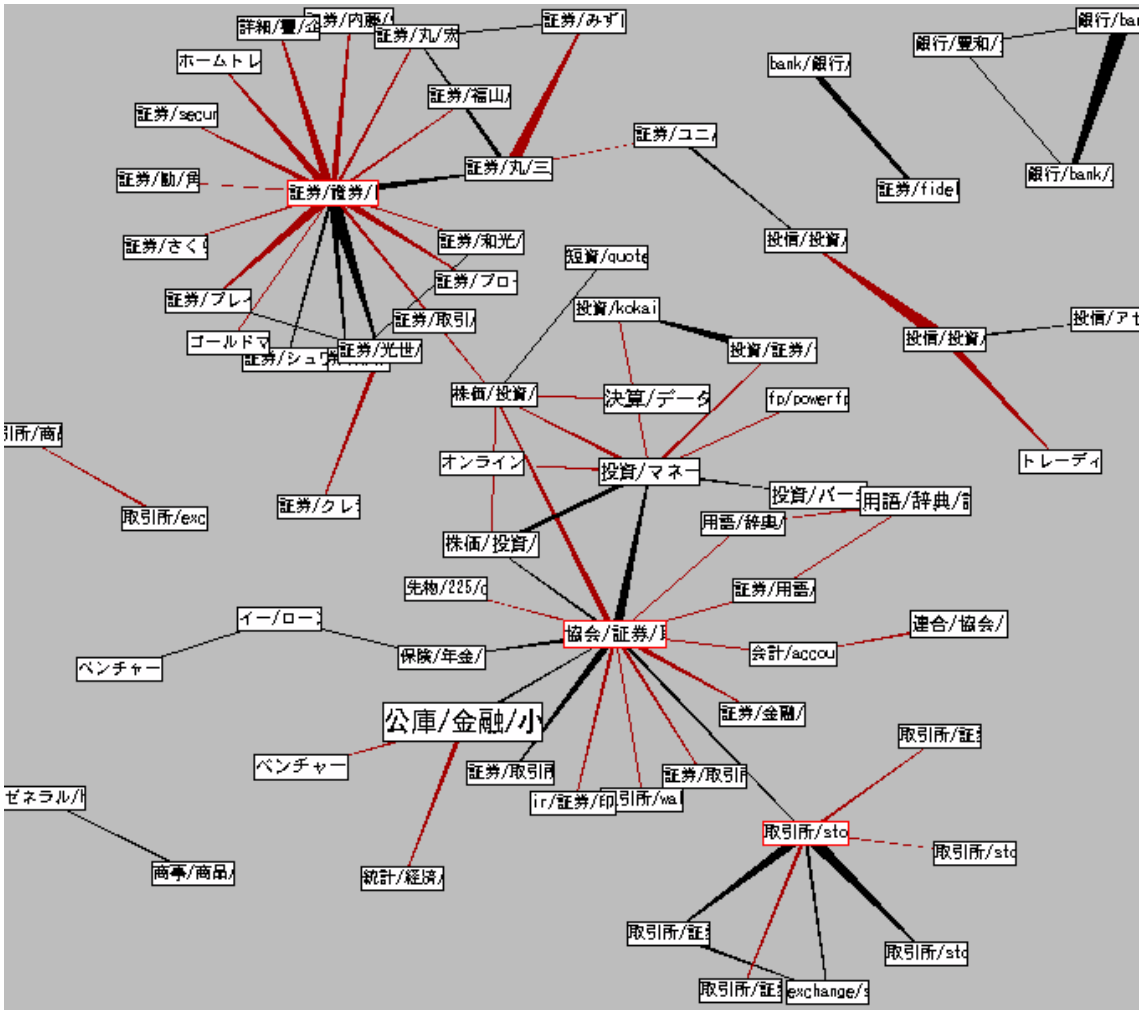

図 6　証券業関連のコミュニティの構造図

で、初期配置でノード数が多かった場合は、不要ノード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、孤立ノードを除去する作業が主で、コミュニティのスコアの閾値調整による除去はあまり使わない。これは、採用したコミュニティスコアの算出方法の妥当性が検証できておらず、重要なコミュニティが除去されてしまう可能性がある為である。一方、一般に他のコミュニティとの関連の薄いコミュニティはそもそもコミュニティ間の関連構造を見る際には関係が薄い事が多く、除去しても影響が小さい。コミュニティスコアの算出方法の改善は今後の課題である。

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。

現在は、コミュニティ間の構造のみが可視化されているが、コミュニティ間構造のさらなる解析のためには、コミュニティ内部の ADG の構造を調べる必要がある。コミュニティ間構造と ADG の両方を同時に閲覧できるような可視化が求められる。

7.　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

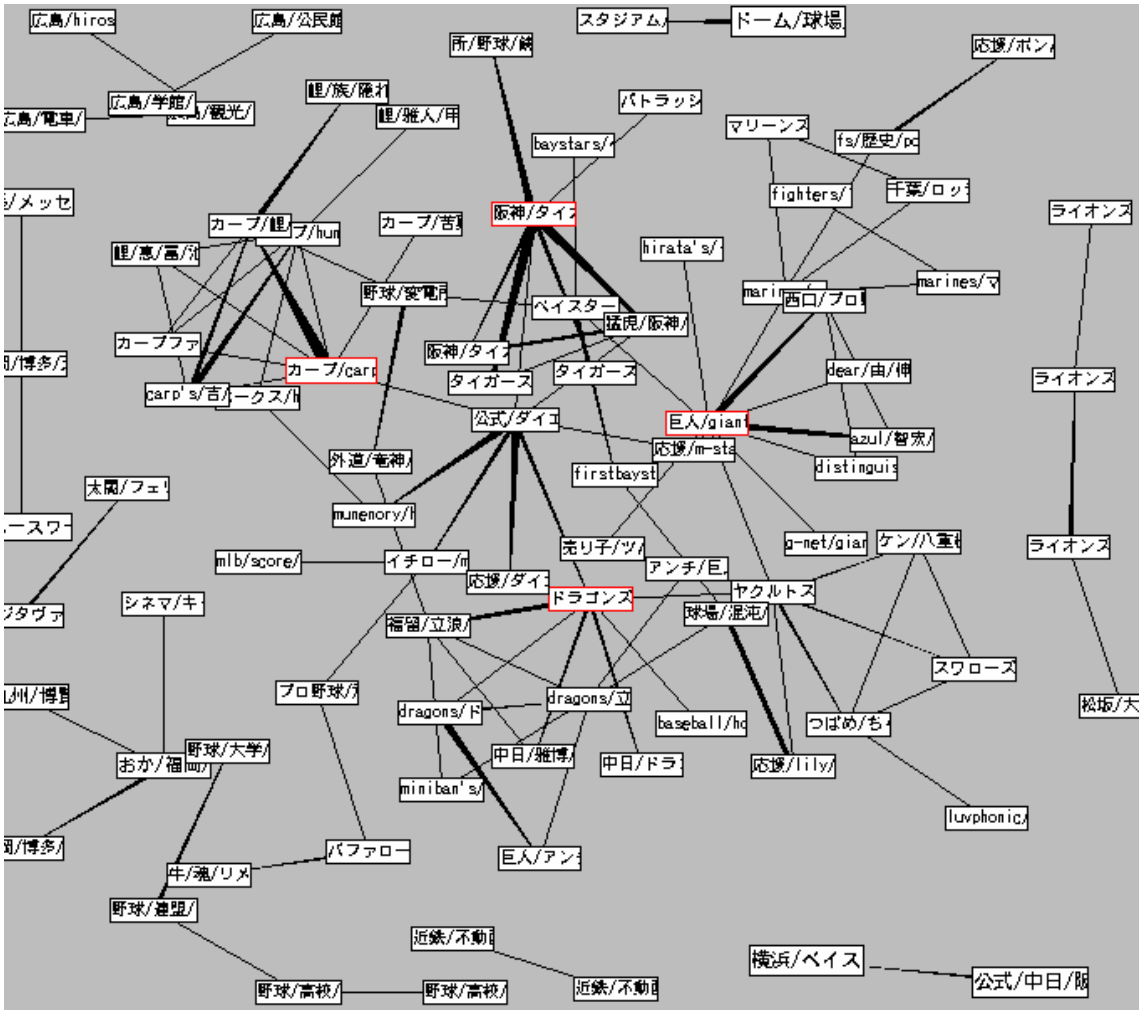

図 7　プロ野球関連のコミュニティの構造図

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。

今後は、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課題とする。また、コミュニティ構造の経年変化を捉えるための機構が必要である。

参 考 文 献

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