Web リンクの可視化によるグラフ構造の発見

浅野 泰仁*, 今井 浩†, 豊田 正史‡, 喜連川 優‡

{asano, imai}@is.s.u-tokyo.ac.jp, {toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

*東京大学大学院理学系研究科 †東京大学大学院情報理工学系研究科 ‡東京大学生産技術研究所

概要

近年, Web 上の情報発見手法として, Web のリンク構造を用いた手法が研究されてきている. 代表的なものとして HITS, Trawling などがあげられるが, これらの手法で用いているグラフ構造は, グラフ理論的見地に立てば, 非常に小さいグラフの性質しか用いていないため, さらに有用なグラフ構造を見つけることが重要だと考えられる. このようなグラフ構造を発見するためには, Web リンクのグラフ構造をわかりやすく可視化する手法が有効であると考えられるが, 既存の手法は基本的にブラウジング支援が目的であり, 全ての Web ページを対等に扱っているなどの理由により, グラフ構造の判別が困難である. 本論文では, Web のリンクをローカルリンクとグローバルリンクに分け, 両者を明確に区別しながら同じ空間上に配置できるように 3 次元球面上にグローバルリンクを, その外側に広がる円錐内にローカルリンクを表示するシステム **Web-Linkage Viewer** を作成した. このシステムを用いて, jp ドメインの URL データから構築された Web 部分グラフを描画することで Web 上の情報発見手法に有用なグラフ構造を発見する方法を提案する.

1 はじめに

現在, Web 上での情報検索は多くの人々の生活の一部となっている. 代表的な検索手法としては yahoo などのディレクトリ型検索, google, goo, infoseek などのロボット型検索エンジンなどが挙げられる. これらは基本的にテキスト検索を用いている.

これに対して, 言語の曖昧さを持たない検索手法として, Web ページのリンク情報を用いた検索手法が近年研究されてきている. これはテキスト検索に取って代わる手法というよりは, テキスト検索と互いに弱点を補完し合う手法であるといえる. Web のリンク情報のみを用いた検索手法の代表的なものとしては, HITS, trawling などが挙げられる.

Kleinberg の HITS([5]) は, あるトピックに関する authority(権威のある) ページと, hub(よいリンク集) ページを見つけることができる. authority は多くの hub からリンクされているページ, hub は多くの authority にリンクしているページとして定義される.

Kumar らの trawling([6]) は, Web 上のコミュニティをセンターページ集合とファンページ集合から成る完全 2 部グラフとして特徴付け, Web グラフ全体からコアと呼ばれる小さなサイズの完全 2 部グラフを見つけることによって, コミュニティを抽出している. この手法では Web グラフ全体が必要であるため, 記憶量と計算時間が膨大になってしまうという欠点がある. 村田 [8] は検索エンジンのバックリンク検索を利用して, Web グラフ全体を使うことなく trawling と似た手法で必要なコミュニティを抽出している.

一方, これらの手法で用いているグラフ構造は, いずれも距離 1 程度の関係しか用いておらず, グラフ理論的見地から見れば, まだ Web リンクの持つグラフ構造を十分に活用しているとは言い難い. 逆に, 新しい有用なグラフ構造を見つければ, 新しい検索手法が構築できる可能性は高い.

グラフ構造を見つけるのに有効な手段のひとつと考えられるのが, Web リンクの可視化である. Web リンクを, そのグラフ構造が人間の目で見てわかりやすく表示されるように描画することで, 意味的に関連の強いページ集合がもつリンクのグラフ構造の特徴をとらえやすくなるからである. 既存の代表的な Web リンク可視化手法としては, 球体内に木構造を描画する H3 Viewer([7]), 木構造を平面に描画する Astra Site Manager([1]), WebOFDAV ([3]), 3D 空間の底平面にリンクを描画し, さらに注目する点とその近傍を上方向に引き上げて目立たせる描画の納豆ビュー ([11]) などがある. 既存の多くの可視化手法は基本的にはユーザーのブラウジング支援を目的として開発されたものであり, ユーザーの訪れたページを順にたどって構築される木構造を描画するのに向いている. しかしながら, そういった手法は Web グラフの構造を見るという目的のためには使いにくい. 主な理由は: (1) 木構造を

なさないリンクを描画しようとすると極端に見にくくなる, (2) 全てのページを対等に扱っているため, サイト間のリンク (グローバルリンクと呼ぶ) とサイト内のリンク (ローカルリンクと呼ぶ) が判別しづらい, という 2 点である.

本論文では, Web リンクをローカルリンクとグローバルリンクに分け, 両者を明確に区別しつつ同じ空間上に配置できるよう 3 次元球面上にグローバルリンクを, その外側に広がる円錐内にローカルリンクを表示するシステム **Web-Linkage Viewer** を作成した.

このシステムを用いて, [9], [10] と同様にして作成された jp ドメインの約 2300 万 URL からなるデータベースから構築された多くの Web 部分グラフを描画する実験をおこなうことで, Web 上の情報発見手法に有用なグラフ構造のデータベースが蓄積でき, その解析につながっていくと考えられる.

2 ローカル・グローバルリンク

本論文では, Web リンクをサイト内部のローカルリンクとサイト間のグローバルリンクに分けて扱う. これはローカルリンクは疑似木構造を持つが, グローバルリンクはより一般的なグラフ構造を持つということが予想され, それぞれの構造を分けて解析した方がよいと考えられるためである. たとえば, [2] は Web グラフ構造解析実験によって Web グラフ全体が蝶ネクタイ構造を持つことを示したが, このような実験もローカルリンクとグローバルリンクそれぞれに適用することによって, また違った構造が見えてくると考えられる.

サイトとは, Web における会社や個人のページ集合を表す概念であるが, その定義ははっきりしていない. 本論文では, サイトの定義を以下のように提案する. ページ v の制作者と呼べる人格 (個人, 法人, 共同製作者からなるグループなど) が与えられているとする.

定義 1 ページ v の属するサイトとは, *Web* リンクによって構成される u から v(または v から u) への, そのパス上の全てのページの制作者が v の制作者に等しいようなパスが存在するページ u の集合のうち極大なページ集合である.

ここで, ページのあるサーバーについては問題としていない. サーバーによるサイト分けは, 上の定義の近似にはなるが, たとえば個人サイトの場合, 掲示板などの CGI は別のサーバーに置いてあることも多く, サーバーでサイトを分けると, 意味的には同一人のページ集合なのに別のサイト扱いになってしまうし, geocities などホームページスペースを提供するサーバー内には, geocities そのものによって作成されたページと, 多くの個人ユーザーが作成したページとがあるが, これらをまとめて単一のサイトとするより, geocities が作成したサイトと, 多くの個人ユーザーのサイトとに分ける方が自然なことと考えられる.

サイトを判別する手法ができれば, 以下の定義に従って Web グラフをローカルリンクとグローバルリンクとに分けて解析することが可能になる.

定義 2 (1) サイト A に属するページ u からサイト $B \neq A$ に属するページ v へのリンクがあるとき, サイト A からサイト B への**グローバルリンク**がある, という. (2) サイト A に属する 2 つのページ u, v 間のリンクを, サイト A 内部の**ローカルリンク**という.

だが, サイトのトップページや掲示板などが持つリンクの性質からある程度の判別は可能としても, Web サイトの構成は多種多様であるため, リンク情報だけを用いて統一的に全てのサイトを判別する方法を構築するのは難しい. 今回作成したシステムでは, 原始的にサーバーだけによるサイトの近似的な区分をおこなっている. リンク, テキストなど様々な情報を用いたサイト判別手法の構築は, 今後の研究課題である.

3 Web-Linkage Viewer

3.1 基本的な考え方

本論文では, ユーザーのブラウジング支援にも使える上に, Web リンクのグラフ構造が人間の目で見てわかりやすく表示されるように描画するために, (1) ローカルリンクとグローバルリンクを分けて同一空間上に表示する, (2) あるサイトのローカルリンクを消したり付け加えたりしてもグローバルリンクの描画が乱れない, (3) 各サイトを平等に扱うために特殊な意味を持つ配置場所 (中心や端) がないようにする, という 3 つの要求を満たすような描画を考える. 平面上に描画することを考えると, (1), (2) を満たすためには平面上にひとつの境界をつくり, 境界上にグローバルリンクとローカルリンクの共通点 (サイトの代表点) を配置し, 生じる 2 つの領域の片方にグローバルリンク, もう一方にローカルリンクを描画するのが自然であり, さらに (3) を満たすためには境界は円であるのが良いので, 平面上に円をひとつ配置し, 円周上にサイトの代表点を配置し, 円内にグローバルリンクを描き, サイトごとに円周から外に広がる木としてそのサイト内の

ローカルリンクを配置する描画が考えられる. 木はたとえば円に接する三角形の内部に描く. しかしこの描画は, 円周上に点を配置する際の自由度が低くグローバルリンクが多いとすぐ見にくくなる. 実際, 交差を許さない直線だけを用いた場合, K_4 も描けない.

上の 3 つの要求を満たすという性質はそのままに, グローバルリンクの描画の自由度を上げるために, これを 3 次元空間へ拡張し, 3 次元球面上にサイトの代表点を配置してグローバルリンクを描き, サイトごとにその代表点を頂点として球の外側に広が円錐内にローカルリンクを描くことにした. 球面上のグラフ描画は, 過去に研究されてきた平面上のグラフ描画の手法を拡張して用いることができるという点でも扱いやすい. また, 球面上にはある 2 点 a, b からそれぞれ与えられた距離 r_1, r_2 にある点が複数存在する場合があることに基づいた平面にはない描画に関する性質もあり, 今後活用できると考えられる. 欠点は, 球面の表と裏が同時に見られないことと, 3D プログラミングが複雑になるという点である. また現在, サイト内部のページから他のサイトのページへのリンクは全てグローバルリンクで表しているが, サイト内部のどのページからリンクが出ているのかがわかった方がよい場合もある. この場合にはサイト内部のページを示す点から他のサイトのページを示す点への線分を描画すればよいし, この描画を付け加えてもグローバルリンクの描画領域を乱すことはない.

我々は, この描画を実際に Windows 上で実装し, **Web-Linkage Viewer** と名付けた. 実装には Borland C++ Builder 5.0 と Microsoft DirectX 8.0 を使用している.

3.2 ローカルリンクの描画

ひとつのサイト内のローカルリンクは, 基本的に木構造をなしていることが多いため, 幅優先探索によって木構造を見つけ, 3 次元円錐の内部に広がる木として配置したあと, 木構造をなさないリンクも付け加えている. ページに対応する各点は立方体で, 各リンクは 2 点間を結ぶ線分で描画する.

3.3 グローバルリンクの描画

現在, **Web-Linkage Viewer** は [4] のスプリングモデルを平面から球面に拡張した描画を用いている.

平面上のスプリングモデルは, 任意の 2 頂点間にバネが存在すると想定し, バネの伸び縮みによって生じるエネルギーがもっとも減少する方向に点を少しだけ動かし, またエネルギーを計算して動かすことを繰り返して安定な状態になったらそれを最終的な配置とする.

球面においては, 距離の定義が 2 通り考えられる. ひとつは 3 次元空間上での (球体内を通る)2 点間の距離をそのまま用いる定義であり, もうひとつは球面上での最短距離を用いる定義である. 今回は, 前者を採用して, 極座標系を用いた 2 次元の回転角度の成分のエネルギー差分を計算している. 後者の方がより自然に思えるが, 前者が満たすような距離の法則 (ピタゴラスの定理など) が満たされない上に極座標を用いたエネルギーの計算が非常に複雑になるためである. こうして得られた配置に従って各点を立方体で, 各リンクは球面上での最短距離を実現する曲線で表示する.

この手法は, 3.5 節で見るようにある程度グラフ構造を見やすく表示することができる. また, 本来調べたい関連度の強いページ集合は密なリンク構造をしていると考えられるため, 密なリンクを持つ頂点集合を近くに表示する工夫をした描画について研究中である.

3.4 その他の仕様

Web-Linkage Viewer は描画する Web グラフとして, URL データファイルを読み込むほかにも, ブラウザの履歴と, その中のページにリンクしているかリンクされているページ集合からなるグラフを作ることができる. これは, 履歴はユーザーが興味を持つページを表し, リンクをたどることでユーザーが次に見るべきページに到達することができると考えられるからである. また, あるグラフを描画している状態から, さらにファイルを読み込んでグラフを大きくしたり, 現在のグラフ構造をファイルに保存したりもできる.

その他, マウスによる拡大縮小・回転, クリックによる URL の表示・ブラウザでその URL を開くなどの操作が可能である. また, 球体を半透明にすることもできる.

3.5 Web-Linkage Viewer を用いたグラフ構造の発見

我々は [9], [10] と同様の手法で 2000 年 7-8 月に収集された jp ドメインページからのリンクデータから作成された約 2300 万 URL, 1 億リンクからなる Web グラフから部分グラフを作成し, 関連するサイトの集合を含む小さなグラフを描画して, それらがどのようなグラフ構造を持っているのか研究中である. 図 1 は

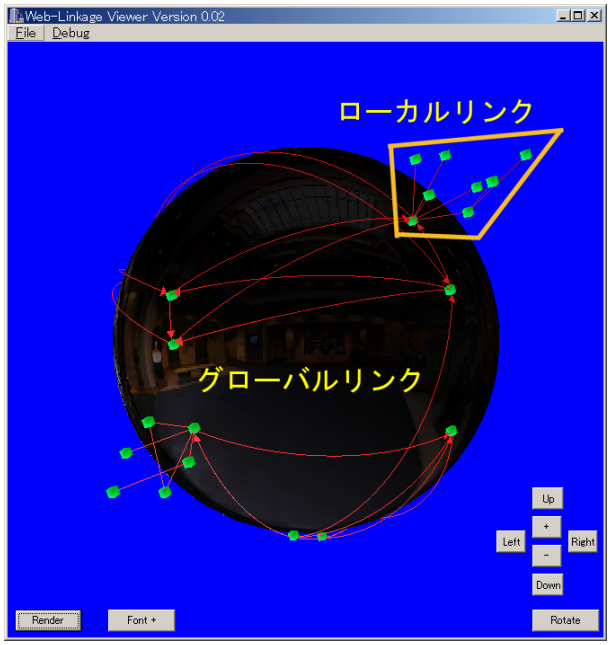

図 1: **Web-Linkage Viewer**

描画例のひとつである. 紙面上では縮小されてわかりにくい面もあるので, 画像を加工してローカルリンクを線で囲い, またローカル・グローバルリンクの描画領域を文字で示してある. この例では, グローバルリンクに K_4 があるのがよくわかる. ローカルリンクは数が多いので, これをグローバルリンクと同一の面に描けばこのようにわかりやすく表示することはできないと考えられる. またいくつかの頂点は裏側に描画されていて見えないが, 回転することで見ることができる. また球体を半透明にすることで, リンクや頂点は重なって見やすさは損なわれてしまうが, 裏側の様子をある程度知ることはできる.

このように, このシステムを用いて, 多くの Web 部分グラフを描画する実験を通して, 関連するページ集合が持つグラフ構造のデータベースを構築しそれを解析するという手法で, 情報検索に有用なグラフ構造が発見できるものと考える.

4 まとめと今後の予定

本論文では, **Web-Linkage Viewer** というグローバルリンクとローカルリンクを 3 次元球面上とその外側に広がる円錐内に表示するシステムを作成し, Web リンクのグラフ構造を見やすく表示する描画を通して情報検索に役立つグラフ構造を見つける手法として提案した. 今後は, この描画手法をさらに発展させてその有効性を検証するとともに, 自動生成されたコミュニティデータベースなどから Web 部分グラフを自動生成し, それを描画する実験を通して, 情報検索に有用な Web グラフの構造の解析をおこなうことを考えている.

参考文献

[1] Astra site manager. `http://www.mercury.com/`

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. In *Proceedings of the 9th International WWW Conference*, 2000.

[3] M. L. Huang, P. Eades, and R. F. Cohen. WebOFDAV - navigating and visualizing the Web on-line with animated context swapping. In *Proceedings of the 7th International WWW Conference*, 1998.

[4] T. Kamada. *Visualizing Abstrat Objects and Relations - A Constraint-Based Approach*. World Scientific, 1989.

[5] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th SODA*, pp. 668–677, 1998.

[6] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. In *Proceedings of the 8th International WWW Conference*, 1999.

[7] T. Munzner. *Interactive Visualization of Large Graphs and Networks*. PhD thesis, Stanford University, June 2000.

[8] T. Murata. Discovery of Web communities based on the co-occurrence of references. In *Proc. of the 3rd International Conf. on Discovery Science, LNCS1967*, pp. 65–75, 2000.

[9] M. Toyoda and M. Kitsuregawa. Finding related communities in the Web. In *Poster Proceedings of 9th International WWW Conference*, pp. 70–71, 1999.

[10] M. Toyoda and M. Kitsuregawa. A Web community chart for navigating related communities. In *Poster Proceedings of 10th International WWW Conference*, pp. 62–63, 2000.

[11] 塩澤秀和, 西山晴彦, 松下温. 「納豆ビュー」の対話的な情報視覚化における位置づけ. 情報処理学会論文誌, Vol. 38, No. 11, pp. 2331–2342, 1997.