論文内容の要旨

論文題目　A New Framework for Link-based
Information Retrieval from the Web
(リンクベースのWeb上情報発見手法の新しいフレームワーク)

氏名　浅野 泰仁

近年, Web上の情報発見手法として, Webページを点, ページ間のWebリンクを辺とするグラフ (Webグラフと呼ぶ) の特徴的構造を利用する手法が研究されている. 代表的な手法としては, Kleinbergによって提案されたHITSやKumarらによって提案されたTrawlingなどが挙げられる.

これらの手法は, WebページvがWebページuからリンクされているということはvがuにとって価値のある情報を含んでいると考えられるという基本的なアイディアに基づいているが, Webサイト内のリンク (ローカルリンクと呼ぶことにする) およびWebサイト間のリンク (グローバルリンク) それぞれの特徴を十分に活用しているとは言い難い.

本研究では, Web上の情報発見手法に用いるグラフとして従来のWebグラフではなく, Webサイトを点としグローバルリンクを辺とするWebサイト間グラフと, 各Webサイトについてその中にあるWebページを点としローカルリンクを辺とするWebサイト内グラフを用いるという新しい枠組みを提案し, その上で有効な情報発見手法を研究している.

Webサイト間グラフを用いることによって, これまでの情報発見手法ではWebページ間の関係しか使えなかったのに対して, Webサイト間の関係, 例えば相互リンクなどを用いることができるようになる. その理由は, 一般に相互リンクはお互いに関係の深いWebサイト間に張られているが, 多くのサイトではサイト内のリンク用のページから相手のサイトのトップページにリンクを張っていたりするため, Webサイト間に相互リンクがあっ

てもページ間に相互リンクがあるとは限らないからである. また, Webサイト内グラフは例えばLiらが提案するInformation Unitや定兼・伊川らのサイト内スコアリングなどのローカルリンクが重要な意味を持つ手法で有効に活用できると考えられる.

"Webサイト"という概念は日常において明確な定義なしに, あるひとりの個人やひとつの会社または集団などによって書かれた関連トピックを持つWebページの集合を表す語として用いられているため, Webリンクをグローバルリンクとローカルリンクに分けて解析し情報発見手法に役立てるためには, Webサイトの適切なモデルが必要となる.

既存の研究では, Webサイトのモデルとして Webサーバーが用いられてきた. この考え方はたとえば会社, 政府, 社会的組織によって制作された公的なWebサイトのようにひとつのWebサイトがひとつのWebサーバーに対応する場合は比較的うまく機能するが, レンタルWebサーバー, インターネットサービスプロバイダ, 大学などのサーバーに置かれた個人Webサイトのように複数のWebサイトがひとつのWebサーバーに対応する場合はうまく機能しない. 世間一般において比較的知られた分野に関しては公的なサイトに置かれた情報だけでも十分であることが多いが, 比較的マニアック (かつ, おそらくマイナー) な分野の情報に関しては個人サイトに置かれる情報が増えてきているため, そのような情報を得ようとするときWebサイトのモデルとしてWebサーバーを用いることは不利であると考えられる.

本研究では, 典型的個人サイトをうまく扱うために *directory-based site* というWebサイトの新しいモデルを提案する. 与えられたWebサーバー内にサーバーの管理者以外のあるアカウント X がファイル作成消去の権利を持つディレクトリが存在する場合, そのディレクトリ内にあるページの集合を X のdirectory-based siteと呼ぶ. そのようなディレクトリにないページの集合は管理者のサイトと呼ぶことにする. さらに, directory-based siteを識別する手法を提案し, *filters* と名づけた. おのおののfilterは与えられた複数のWebサーバーの一部を, ひとつしかdirectory-based siteを持たないサーバーとそうでないサーバーとかどちらかに決定し, 残りのサーバーを次のfilterに入力として引き渡す. 提案したfiltersは全部で6種類あり, URLと有名サーバーに関する知識, 簡単なヒューリスティクス, 連結成分分解, バックリンクとディレクトリの数などに基づいている. さらに後述のクリークを用いて誤り訂正をおこなうこともした. 2000年, 2002年それぞれで豊田, 喜連川によって収集されたjpドメインURLデータを用いて, この手法がほとんどのサーバーを正しく上記のどちらかに決定し, もとのサーバーの数の5倍以上のdirectory-based sitesを抽出できることを確かめた.

こうして得られたWebサイト間グラフおよび各directory-based siteにおけるWebサイト内グラフに対して, まず最初に, Webからの情報発見手法に重要な役割を果たす次数分布を調べ, 従来用いられてきたWebグラフの次数分布とは, 全Webサイト内グラフの次数分布を集めたものに非常に近く, 逆にWebサイト間グラフの次数分布とは大きく異なっているという結果を得た. これは, Webサイト間グラフとWebグラフの性質の違いを示唆していると共に, Webグラフの生長モデルは現在提案されている単一種類のWebリンクのみを考えたpower law graphモデルとは異なることも示唆している.

次に, Trawling を 2000 年および 2002 年の jp ドメイン URL データから得られた従来の Web グラフと Web サイト間グラフ両方に対して適用し, 特に *nepotistic core* の問題に関して Web サイト間グラフの方が自然にサイト間の関係を利用できることを示した. また Web サーバー間グラフと Web サイト間グラフについて得られたコアを内容によって分類することで, 情報発見のためには directory-based site の方が Web サーバーより Web site のモデルとして良いことを示した.

さらに, Web サイト間グラフを活用する新しい情報発見手法として, directory-based site とそれらの間を結ぶ相互リンクからなるグラフから, 極大クリークを列挙することによって関連した directory-based site の集合を発見する手法を提案し, 上記の URL データを用いてそのような集合を実際に発見し, 特に, Web サーバーをサイトのモデルとして用いたのでは発見できない要素である個人サイトを含む関連サイト集合が含まれていることを確かめた.

これらの手法は, Web リンクの巨大なデータと膨大な計算時間を必要し, またデータに含まれるすべてのコミュニティや関連サイト集合を発見することが目的であるがために, ユーザーが興味を持つ情報を発見する目的には向かない. そこで本研究ではこの目的に使える手法として, ユーザーが興味を持つ複数の URL(履歴やブックマークなどでもよい) を与えることによって, その近傍からなる Web サイト間グラフを実際の Web からオンデマンドで取得し, ユーザーが興味を持つであろうコミュニティを計算するシステム **Neighbor Community Finder** を開発し, 実際に有志から得たサンプルに関してその近傍のコミュニティを計算し, さらに Google の "Related Sites" サービスと比較することで我々のシステムの優位性を確かめた.

また, Web サイト内グラフの特徴を利用することで Web グラフの新しい圧縮方法が提案できることを示し, 従来の方法より圧縮率が高く, 複合にかかる時間もさほど変わらないことを計算機実験によって確かめた.

最後に, Web リンクの構造をわかりやすく描画するためにも, Web サイト間グラフおよび各 Web サイト内グラフを用いることは有効であることを示し, さらにローカルリンクとグローバルリンクを分けて描画するシステム **Web-linkage Viewer** を提案し, 既存の方法より各 Web サイト内グラフの木構造に似た構造が明確に視覚化できることを示した.