論文審査の結果の要旨

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と，URLとWebリンクに関する知識を用いてこれを識別する手法を提案している．また，2000年7月から8月にかけて豊田，喜連川によって収集された jp ドメインの2300万以上のURLおよび1億以上のWebリンクからなるデータを用いた計算機実験によって11万以上のサーバーのうちおよそ 66% のサーバーが複数の directory-based site を含むかどうか識別して 50 万以上のdirectory-based siteと約400万のグローバルリンクを実際に抽出している．さらに，2002年2月のjp ドメインのURLデータに対しても同様にdirectory-based siteの抽出をおこなっている．

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ている．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，計算機実験により比較をおこない，Webサイト間グラフを用いた方がより多くのコアを見つけることができ，コアのサイズも小さくなることを確かめている．さらに2000年と2002年のデータに対する結果を比較することで，その差がどのように変化するかも確かめている．

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，極大クリークやそれに似た密な部分グラフ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，jpドメインの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ている．

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抽出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関

連サイト集合を計算するシステムを開発している．また，このWebサイト間グラフおよび各Webサイト内グラフを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 によって，ユーザーにこのグラフ構造と関連サイト集合の情報をよりわかりやすい形で提供している．最後に，Webサイト間グラフとWebサイト内グラフそれぞれ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，従来の方法と計算機実験による比較をしている．

以上のように本論文では，情報科学的に緻密な考察を経て，Web グラフの情報整理のため directory-based site という新しい概念を定式化することに成功している．しかも大量の現実のデータに対して，本手法の有効性を検証しており，説得力の高い内容になっている．

なお本論文の内容は，今井浩・豊田正史・喜連川優を共著者として既に外部のいくつかの学会において公開されているが，論文提出者の寄与が十分であると判断する．

従って，博士(理学)の学位を授与できるものと認める．