情 報 学 基 礎 61－8
自 然 言 語 処 理 142－8
（2001. 3. 5）

リンク構造の連結性に注目したコミュニティ導出に基づくWebブラウジング手法の提案

外山大介* 吉高淳夫** 平川正人**
*広島大学大学院工学研究科
**広島大学工学部第二類（電気系）
〒 739-8527 東広島市鏡山 1 丁目 4-1
E-mail: {tym, yoshi, hirakawa}@isl.hiroshima-u.ac.jp

本論文では、Web サイトのコミュニティを用いてユーザの Web ブラウジングを支援するシステムを提案する。コミュニティの導出にあたっては、ユーザがWebブラウザで閲覧しているWebサイトの周辺のリンク構造に注目する。Webサイト間のリンクによる連結の関係とそれによる関連度を重み付き偏差行列で表現し、強連結成分の抽出を行うことでコミュニティを導出する。Webサイト間の関連度に基づいてコミュニティをグラフ化し、提示することでWebブラウジング時のユーザの視界を拡張する。

Community Extraction for Web Browsing Based on Hyperlink Connectivity

Daisuke TOYAMA*, Atsuo YOSHITAKA**, Masahito HIRAKAWA**
*Graduate School of Engineering, Hiroshima University
** Faculty of Engineering, Hiroshima University
4-1, Kagamiyama 1 chome, Higashi-Hiroshima, 739-8527
E-mail: {tym, yoshi, hirakawa}@isl.hiroshima-u.ac.jp

In this paper, we propose a WWW browsing method based on communities that are sets of web sites created by the authors who have a common hobby, insistence, and interest on a specific topic. Communities are derived using hyperlink structure. A weighted deviation matrix describing the hyperlink connectivity is used to extract communities from WWW by considering strongly connected elements. The proposed system provides an additional browsing view to see not only surrounding web sites but also related communities, which assists the user in performing WWW navigation.

1. はじめに

Web ブラウジング時のユーザは、短期的な目的や趣味趣向に基づいた情報要求を持ち、リンクをたどってそれらの情報を探索する。Web ブラウジングによってユーザの要求を満たすような情報が得られるか否かは、ユーザの選択するリンクに大きく関係しており、選択したリンクによっては有効な情報源を見落とすことになる。そのため、ユーザはリンクを選択するための判断材料を必要としている。しかし、Web サイト閲覧時にユーザに与えられるリンク先に関する情報の大半は Web サイト作成者独自の主観に依存しており、リンク先についての説明が不十分な場合も少なくない。このような場合には、ユーザは機械的にリンク先を順次選択して閲覧することとなり、効率が悪い。それに対し、[1]の手法のように、リンク先の Web ページを先読みし、提示を行う支援方法がある。この手法ではリンク先の Web サイトの内容を一部確認できるが、他の Web サイトとの関係を把握できない。また、[2]のようにリンク構造をグラフ化し、可視化を手法があるが、複雑なリンク構造を提示するだけでは Web サイト同士の関係を把握するのが困難であり、有効な支援とはいえない。閲覧している Web サイトと周囲に存在する Web サイトとの関係を適切な形で提示することが必要である。

近年、様々な Web 検索の研究で、検索結果の信頼性を高めるために、リンクによる Web ページ間の繋がりが注目されている。それらの研究結果によって、リンク構造を利用することで検索結果の適合率、再現率が共に増加することが示された[3][4]。これは、Web 空間中のリンク構造は Web ページの内容と密接な関係があり、関連のある Web サイト群はその特徴に基づいた集合を自ずと形成していることを表している。まず、本研究では同一 Web サイト内の Web ページを参照しているリンクを内部リンク、他の Web サイトの Web ページを参照しているリンクを外部リンクと定義する。一般的に、外部リンクは Web サイト作成者が他の Web サイトとの間に何らかの関連を認めたときに作成される。また、外部リンクは作成者の意図によって選別されているため、多くの Web サイトにリンクされた Web サイトはそれだけの支持を得ているといえる。また、パスの距離をパスに含まれる外部リンクの数とし、Web サイト間の距離を、Web サイト間の最短パスの距離と定義すると、Web サイトの距離と関連には相関があり、互いの距離が一定値以下である Web サイト群は何らかの意味的まとまりを持っていると考えられる。

本研究では、このような Web サイト群をコミュニティとして導出し、提示することで、ユーザの Web ブラウジング時の視界を拡張する。コミュニティの導出にあたっては、ユーザの閲覧している Web サイトの周辺のリンク構造に注目し、Web サイト間のリンクによる連結性とそれによる関連度を重み付き偏差行列で表現することで、強い関連を持った連結成分の抽出を行う。

2. 連結性

本研究では、Web サイト間の繋がりの強さを把握するため、Web サイトを頂点、Web サイトを結ぶ外部リンクを枝とした有向グラフで Web サイトの接続関係を表現する。有向グラフには連結性の型が存在し、連結性の強い順に、完全連結、強連結、半強連結、準強連結、弱連結となっている[5]。これらの連結性の型は頂点間のパスの有無によって決定され、パスの距離や本数に依存しない。そこで、そのようなパスの距離や本数を考慮することで、同じ連結性の型を持つグラフに対しても、より詳細な頂点間の連結の度合いを判断する。なお本研究では、強連結を形成している連結成分を強い繋がりを持った成分として注目する。強連結とはリンクによる双方向のパスが存在する関係である。双方向のパスが存在している頂点同士は互いに到達可能であり、その距離が短く、本数が多ければより深い関連を持っている。そのような高い連結性を持った頂点群をコミュニティとして導出する。

3. 関連度

同じ強連結の型を持つ頂点間でも連結の強さを比較するために、パスの経路による関連度を考慮する。関連度は、強連結のグラフを半強連結のグラフに変えるために取り除かなければならない枝の数で表す。これは、頂点間に存在する基本パスの本数に比例し、リ

ンクをランダムにたどるユーザが、ある頂点から別の頂点に辿り付く到達確率によって表現する。全ての頂点に対しての相対的な到達確率の関係を考慮することで、あるWebサイトが他のWebサイトに対してどれだけの関連度があるかを把握する。本研究では、この関連度を要素の値とする重み付き偏差行列を算出し、そこから強連結成分の抽出を行う。このようにリンク構造を利用することで、Webサイト作成者の他のWebサイトに対する関心の度合いを考慮でき、Webサイト中に含まれるキーワードからだけでは抽出し難い特徴も取り扱うことができる。

4. システム構成

提案する手法の妥当性を確認するためにプロトタイプシステムの構築を行う。図1にシステムの概要図を示し、以下にそのシステム構成について説明する。

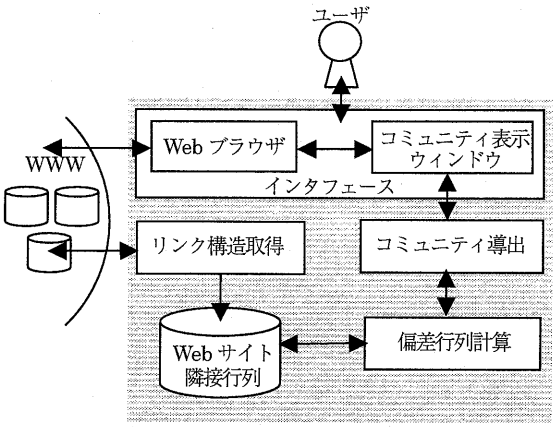

図1 システム概要

4.1 Webサイトのデータセット

あらかじめ、Web空間中のWebサイト間のリンク構造を調べ、その接続関係を表現した隣接行列を蓄えておく。なお、本研究におけるWebサイトとはある個人、組織がWWW上で提供しているWebページの集合であり、一般的に各個人、組織によって提供されていると認識される単位を指す。例えば、同一ドメイン名のURLに存在するWebページであっても、アカウントの異なる作成者によって作られたWebページ群はそれぞれが別のWebサイトであり、そのルートディレクトリのWebページ群（ドメイン名所有者作成）はそれらを含むことなく、別のWebサイトとする。

まず、第一にURL中のドメイン名が異なるWebページは別のWebサイトのものであるとする。例外として、ミラーサイトの存在が挙げられるが、これは2つのページがリンクをどの程度共有しているかで判断する。また、Webサイトの移動元ページや、CGIの使用制限やディスク容量等の理由で別のドメイン名のURLに同一作成者によるWebサイトの一部が存在する場合には、いずれかの外部リンクが少なく、互いの関連度が極めて高い場合には、同一のWebサイトと見なす。関連度が低く、リンクの連結にも相違が見られない場合には同一作成者によるWebサイトであっても主題の異なるWebサイトであると判断し、別のWebサイトとする。

第二に、リンクによって直接参照されているWebページからリンクをたどることによって閲覧可能なWebページで、参照されているWebページ以下のディレクトリ階層に存在するWebページは同一Webサイト内のものとする。

第三に、URL中に"~"（チルダ）を含んでいる場合には、チルダを含んだディレクトリ以下に存在するWebサイト群を1つのWebサイトとして抽出する。これはWWWサーバとして多く用いられているUNIX系のOSにおいてチルダが各ユーザのホームディレクトリを示すという特徴を利用したものである。また、URL中にチルダを含まない場合でも、一般トップレベル・ドメイン名がcom、netの場合や、国別セカンドレベル・ドメイン名がac、co、ne、orの場合には、下層のディレクトリに複数のユーザのWebサイトが存在している可能性がある。このようなドメイン名のURLに対しては、ルートディレクトリから内部リンクの本数よりも外部リンクの本数が多くならない深さのディレクトリまでをWebサイト単位とし、それ以下は各ユーザのWebサイトであると見なす。これは企業、団体によるWebサイトは内部リンクに比べ、外部リンクの数が非常に少ないという特徴を利用したものである[6]。以上のようにWebサイトを抽出し、外部リンクを選別した後、Webサイト間の接続の関係を表した隣接行列を作成し、そこから

Web サイト間の関連度を要素の値とする重み付き偏差行列を算出する。

4.2 コミュニティ導出

ユーザが閲覧している Web サイトの URL データを受け取り、あらかじめ準備された重み付き偏差行列から、そのWebサイトへのパスが存在しているWebサイト群とその Web サイトからのパスが存在しているWebサイト群、つまりユーザの閲覧しているWebサイトに対する半強連結成分を取得する。そこから閾値以上の関連度を持つ Web サイト群を抽出し、その部分行列に対してコミュニティ導出を行う。本研究では、任意の Web サイトの対に対して、双方へのパスが共に存在し、そのパスの本数と距離による関連度が閾値を越えるような Web サイトの連結成分をコミュニティとする。N 個の Web サイトを含む Web 空間グラフ G において、Web サイト V_i からの外部リンク数を n_i、Web サイト V_i から Web サイト V_j への基本パス（同じ Web サイトを二度通過しないパス）の数を $path(V_i, V_j)$ としたとき、V_i が距離 r の基本パスによって V_j に与える関連度を値とした重み付き隣接行列 $A^{(r)}(G)=a_{ij}^{(r)}(1\leq i\leq N, 1\leq j\leq N)$ を作成する。

$$a_{ij}^{(r)}=\begin{cases}1/n_i & : r=1,\ path(V_i, V_j)>0\\ \sum_{k=1}^{N} a_{ik}^{(r-1)}a_{kj}^{(1)} & : r>1, i\neq j, \sum_{k=1}^{N} a_{ik}^{(1)}a_{kj}^{(r-1)}\neq 0\\ 0 & :otherwise\end{cases}$$

$A^{(r)}(G)$ の計算は、まず $A^{(1)}(G)$ を求めた後、$A^{(1)}(G)$ を r 回だけべき乗することによって算出する。ただし、$i=j$ となる項の値 $a_{ij}^{(r)}$ は、その Web サイトからその Web サイト自身に向けての自己推薦としての関連度を表しているため考慮しない。また、閉路を含むパスを全て考慮しないようにする。これらによって、基本パスのみによる関連度を考慮することができる。

N 個の Web サイトを含む Web 空間グラフ G における最長の基本パスの距離を L としたとき、$A^{(L)}(G)$ まで計算することで重み付き偏差行列 $D(G)=d_{ij}(1\leq i\leq N, 1\leq j\leq N)$ を作成する。$D(G)$ は全ての Web サイト同士が与え合う関連度の総和を表す行列であり、それぞれの値は重み付き隣接行列の要素の和を求めることで算出する。

$$d_{ij}=\sum_{k=1}^{L} a_{ij}^{(k)}$$

この $D(G)$ の要素 d_{ij} から連結性の型を判定する。$D(G)$ において、任意のサイト群が $i\neq j$ を満たす全ての要素に対して d_{ij} かつ対称要素 d_{ji} の両方に要素を持つとき強連結である。この値が閾値以上であるとき、コミュニティとする。

上述した手法によりコミュニティを導出した後、コミュニティの評価を行う。コミュニティ C_p の中の Web サイト数を N_p とするとき、C_p の評価値 $EvC(C_p)$ を C_p 内の Web サイト対の関連度の平均値として以下のように定義する。

$$EvC(C_p)=\frac{\sum_{(V_i,V_j)\in C_p} d_{ij}}{N_p^2-N_p}$$

以上のようにして導出したコミュニティの評価値と Web サイトの関連度を、コミュニティとあわせて視覚化し、ユーザに提示する。

4.3 インタフェース

導出されたコミュニティと Web サイトの関連を視覚化し、ユーザに提示する。本システムのインタフェースは Web ブラウザとコミュニティ表示ウィンドウからなる。図2にコミュニティ表示ウィンドウのインタフェースを示す。Web ブラウザでは Web ページの内容を表示し、コミュニティ表示ウィンドウ上では Web サイトを頂点、Web サイト間のリンクを枝としたグラフを表示する。両者間で URL データのやり取りを行うことで、ユーザが Web ブラウザで閲覧しているWebサイトから半強連結のWebサイト群をコミュニティ表示ウィンドウ上に表示する。頂点に各Webサイトのタイトルを表示し、コミュニティ毎に異なる配色を行うことで、Webサイトの属するコミュニティを直感的に把握できるようにし、ユーザの Web ブラウジングを支援する。また、ユーザが Web サイトを選択することで、そこからのリンクが強調表示され、同じコミュニティ内の Web サイトのタイトルがタイトルリストに関連度順に表示される。コミュニティ表示ウィンドウ上の Web サイトはユーザが指定するこ

とによってWebブラウザ上で直接閲覧できる。また、今までに閲覧したサイトは履歴リストに表示される。

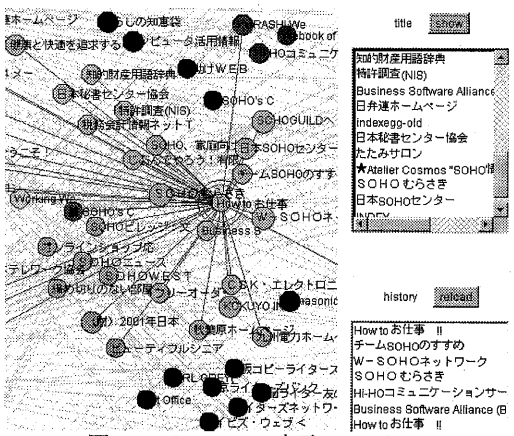

図2　コミュニティ表示ウィンドウ

4.4　スプリングモデル

コミュニティ表示ウィンドウの大きさとWebサイト同士の関連度に従ってウィンドウ上に表示する2頂点間の距離を算出し、頂点を配置することにより、Webサイトの関連を視覚的に表す。関連度によって与えられる頂点間の表示距離を最適距離と定義すると、N個のサイトとその間の関連度による最適距離が与えられているとき、これを完全に表現しようとすると、最低でも$(N-1)$次元の空間が必要になる。そのため、頂点が4つ以上の場合、2次元平面で頂点とその関連度を完全に表現しようとすると誤差が生じる。本研究では同じコミュニティ内のサイト間の表示距離と関連度の誤差を軽減するため、スプリングモデルを用いて配置を決定する。スプリングモデルとは、2頂点間で局所的に決まる最適距離をばねの自然長と見なし、各頂点を質点に見立て、ばねの力学系が安定するように質点を動かすことで全体のばねの力が均衡するように質点の位置を決定する手法である。このような力指向アルゴリズムを用いることで、各頂点間の関連度を距離に反映させることができる。また、頂点同士が重ならないように斥力を与えることで、適切な視覚化が可能になる。頂点間に働く引力f_s、斥力f_rは、disを頂点間のコミュニティ表示ウィンドウ上における表示距離、dis_0を頂点間の最適距離（ばねの自然長）、ばねの強さを決める係数をk_s、k_rと表すとき、次のように定義する。プロトタイプではk_s、k_rの値にサイト対の関連度の和を用いた。

$$f_s(dis) = k_s \log \frac{dis}{dis_0}$$

$$f_r(dis) = k_r \frac{1}{(dis)^2}$$

引力はdisの値とdis_0の値が離れているときには大きく働き、逆に近いときにはほとんど働かない。また、斥力は頂点間の距離の2乗に反比例する。このような力を各頂点に対して算出して全ての合力を計算し、各頂点を移動させつつ再計算を繰り返すことで、全体の均衡がとれるように各頂点を移動させ、それぞれの位置を決定する。頂点の初期配置はコミュニティ間の関連度に基づいて行う。コミュニティC_p、C_qの中のサイト数をそれぞれN_p、N_qとしたとき、C_pからC_qへの関連度$EvD(C_p, C_q)$は、C_p内のサイトV_iとC_q内のサイトV_jの関連度の平均値として以下のように定義する。

$$EvD(C_p, C_q) = \frac{\sum_{(V_i)\in C_p, (V_j)\in C_q} d_{ij}}{N_p N_q}$$

ユーザが最初に閲覧していたWebサイトの頂点を中心に、そのWebサイトを含むコミュニティからの関連度によって他のコミュニティに含まれるWebサイトの初期配置を決定する。

5.　考察

プロトタイプのデータセットとして、Webロボットを使って収集したWebサイトを用いた。収集の基点となるURLを12個決定し、それぞれ幅優先探索によって、約2000サイト、約50000のWebページ（HTML文書のみ）を収集した。収集した全てのWebサイトからリンクを抽出し、重み付き偏差行列を作成した後にコミュニティ導出を行った。

本手法の有効性を評価するために、プロトタイプシステムに対してユーザテストを行った。実際にプロトタイプシステムを用いてそれぞれのコミュニティに対して10サイトのブラウジングを行い、それらの

Webサイトが初期のWebサイトにどの程度関連しているかを判断する主観評価を依頼した。評価基準は「主題が同じ」「関連するが主題は異なる」「関連しない」の3つとなっている。表1に主観評価をまとめたリストを示す。○は「主題が同じ」を選んだ数、△は「関連するが主題は異なる」を選んだ数、×は「関連しない」を選んだ数、－はページがサーバ上に存在しなかった数を表す。精度は「主題が同じ」を2点、「関連するが主題は異なる」を1点としたときの満点に対する合計点の割合で表している。図3にはこの精度のヒストグラムを示した。

表1　コミュニティの評価リスト

URL	○	△	×	－	精度
www.iodata.co.jp/sohot/enjoy/how_to/	4	3	1	2	0.69
kbic.ardour.co.jp/~kaijoken/	6	3	1	0	0.75
www2c.biglobe.ne.jp/~takesako/	3	7	0	0	0.65
www.campus.ne.jp/~rung/yorikiri/	6	1	3	0	0.65
www.biwa.ne.jp/~minoura/	6	2	1	1	0.77
www2u.biglobe.ne.jp/~Ryosuke/	5	5	0	0	0.75
www.seaple.icc.ne.jp/~k-ichino/amuro/	5	4	0	1	0.78
www04.u-page.so-net.ne.jp/fg7/ko-sei/	2	4	4	0	0.67
www.bee-project.com/fmiti/	1	3	6	0	0.25
www10.cds.ne.jp/~niko/	7	3	0	0	0.85
www1.odn.ne.jp/~cbu94260/	9	0	0	1	1.00
home.catv.ne.jp/dd/meidai/	8	0	2	0	0.80
平均	5.2	2.9	1.5	0.3	0.72

○：主題が同じURLの数、△：関連するが主題は異なるURLの数、
－：ページがサーバ上に存在しないURLの数

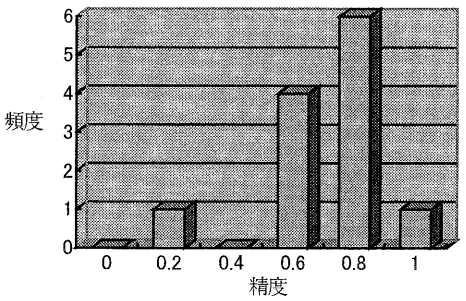

図3　精度のヒストグラム

精度の低い部分は、サイト作成者がYahoo!やテレビ局のページといった著名なサイトをリンクした場合に、そのWebサイトの特徴を特に示さないような連結性が生じているためである。このような現象を防ぐために、多くのデータを収集し、上記のようなWebサイトを様々なWebサイトからリンクされているという特徴から取り除いておくことが重要である。

情報探索という観点でブラウジングを考えた場合、ユーザの要求を満たすことのできる情報をどれだけ高速に得られるかということが重要になる。本システムは、ユーザが閲覧しているWebサイトから直接リンクされていない周囲のWebサイトであっても他のWebサイトからどれだけリンクされているかを把握できる。また、そのWebサイトを直接閲覧できる機能を提供しているため、ユーザがWebサイトを移動する間に無駄なWebページサイトを閲覧する時間や、マウスクリックの回数を軽減することができ、Webブラウジング効率を向上させることができた。

6. まとめ

本研究では、リンク構造の連結性に注目したコミュニティ導出に基づくWebブラウジング支援の手法を提案した。導出したそれぞれのコミュニティとサイトの関連を提示することによって、個々のサイトの関連を把握することが可能になった。

参考文献

[1] T. Joachims, D. Freitag, and T. Mitchell, "Webwatcher: A tour guide for the World Wide Web", In Proceedings of the 15th International Joint Conference on Artificial Intelligence, 1997.

[2] A. Wood, R. Beale, N. Drew, and R. J. Hendley, "Hyperspace: a Worldwide Web Visualiser and its implications for Cooperative Browsing and Agents", In Proceedings of the Human-Computer Interaction, 1995.

[3] L. Page, "PageRank: Bringing order to the Web", Stanford Digital Libraries working paper, 1997.

[4] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[5] C. Flament, "Application of Graph Theory to Group Structure", Prentice-Hall, 1963. (山本國夫訳, "グラフ理論と社会構造", 紀伊国屋, 1974.)

[6] 中川 格, 石塚 英弘, 山本 毅雄, "日本の World Wide Web 情報空間", 図書館情報大学第9回ディジタル図書館ワークショップ, 1997.