特　集　「Webシステムにおける情報獲得支援技術」

Webグラフの構造解析

Structural Analysis of Web Graph

廣川 佐千男　九州大学情報基盤センター
Sachio Hirokawa　Computing and Communications Center, Kyushu University.
hirokawa@cc.kyushu-u.ac.jp

池田　大輔　（同　　上）
Daisuke Ikeda　daisuke@cc.kyushu-u.ac.jp

Keywords: web graph, authority, hub, latent semantics indexing, principal component analysis, singular vector decomposition, power-law.

1. はじめに

人間の活動に関わるさまざまな文書がWeb上で入手できるようになった．Webが人間社会の縮図，電子的な現代社会そのものと呼べるようになり，Web空間からの知識発見が重要な課題となっている．Webページの検索でも，情報の量だけでなく質が問題となっている．さらに関連する複数のページ群を探すときには，それらのページ間の関連自体も重要となる．[Kleinberg 98] や [Kumar 99] では，強く関連し合うWebのページ群をコミュニティと呼び，ページ間のリンク情報を利用し，そのようなコミュニティを発見する手法を述べ，多くの例を示している．これは，Webページとその間のハイパーリンクを有向グラフとみなし，そのグラフの隣接行列に対して特異値分解を用いて主成分を求めるものである．本稿では，コミュニティ発見の可能性とコミュニティ解釈の問題点を，そのコミュニティのキーワード抽出の具体例を通して紹介する．また，大規模なWebデータに基づくWeb空間全体のサイズや構造推定の研究と，グラフ理論を用いた関連研究を紹介する．

2. 隣接行列の特異値分解によるWebグラフ解析

Webグラフとは，Webページをノード，ページ間のハイパーリンクをエッジとする有向グラフである．Webグラフの隣接行列とは，ノード v_i から v_j へリンクがあるとき，$a_{ij}=1$，そうでないとき $a_{ij}=0$ として決まる正方行列 $A=(a_{ij})$ である．例えば，図1は，1番から13番のWebページからなるWebグラフを表す．1番のノードから7番のノードへの矢印は，1から7へのハイパーリンクを表す．

Kleinberg [Kleinberg 98] は，信頼できる情報を含むオーソリティページと有用なページへのリンクを多く含むハブページという概念を導入した．この二つは，「良いオーソリティは良いハブからリンクされており，良いハブは良いオーソリティをリンクしている」という，相互に依存した関係にあり，これを利用したアルゴリズムを用いて有用なページを見つけだすことに成功した．図1の小さい世界では皆から注目されている7, 8, 10がオーソリティであり，9, 5がほかの1, 2, 3, 4, 6, 11, 12, 13のノードより多くリンクを含むという意味でハブと考えられる．各ノード v_i に対するオーソリティ度を x_i，ハブ度を y_i で表すと，各ノードのオーソリティ度とハブ度は $x_i=\sum_{(v_j,v_i)\in E} y_j$，$y_i=\sum_{(v_j,v_i)\in E} x_j$ という方程式の不動点となる．ベクトルで表現すると $x=A^t y$，$y=Ax$ となり，オーソリティ，ハブは行列 AA^t ならびに A^tA の固有ベクトルの中で最大の固有値に対応する固有ベクトルとなることを示した（A^t は A の転置行列）．[Kleinberg 98] ではさらに，主要固有ベクトル以外の固有ベクトルは，異なるコミュニティ発見のために使えることを実証した．AA^t, A^tA の固有値と固有ベクトルは A の特異値分解 $A=U\Sigma^{-1}V^t$ により求めることができる [Berry 95]．図1のグラフの隣接行列 A についての特異値分解では，Σ

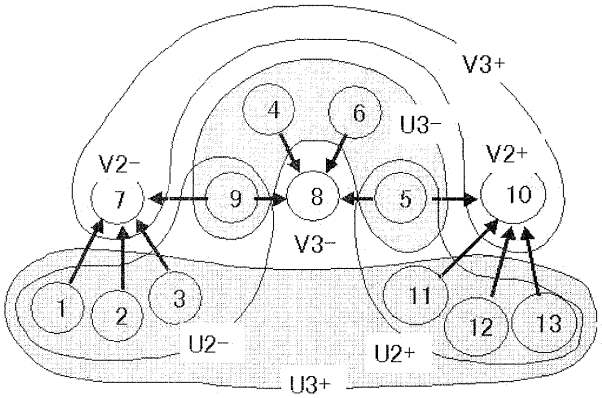

図1　Webグラフとコミュニティ

は2.327，2.000，1.608を成分として持つ対角行列となり，U，Vは次のようになる．

U_1	U_2	U_3	V_1	V_2	V_3
0.215	−0.354	0.311	0.000	0.000	0.000
0.215	−0.354	0.311	0.000	0.000	0.000
0.215	−0.354	0.311	0.000	0.000	0.000
0.304	0.000	−0.440	0.000	0.000	0.000
0.519	0.354	−0.129	0.000	0.000	0.000
0.304	0.000	−0.440	0.000	0.000	0.000
0.000	0.000	0.000	0.500	−0.707	0.500
0.000	0.000	0.000	0.707	0.000	−0.707
0.519	−0.354	−0.129	0.000	0.000	0.000
0.000	0.000	0.000	0.500	0.707	0.500
0.215	0.354	0.311	0.000	0.000	0.000
0.215	0.354	0.311	0.000	0.000	0.000
0.215	0.354	0.311	0.000	0.000	0.000

各ノードのハブ度は固有ベクトルUの1列目のU_1に，オーソリティ度は固有ベクトルVの1列目V_1に現れている．主要固有ベクトル以外のベクトルについて，成分が正で絶対値が大きいもの（正端，ポジティブエンド）U_2^+, V_3^+，ならびに成分が負で絶対値が大きいもの（負端，ネガティブエンド）U_2^-, V_3^-をグラフ中に描くと図1のようになる．各番号iについて，ハブの固有ベクトルU_iとオーソリティ固有ベクトルV_iを眺めると，同じ符号の端は，強く連結していて，符号の異なる(U_i^+, V_i^+)と(U_i^-, V_i^-)は異なるコミュニティを成している．一般的に，それぞれの非主要固有ベクトルの正端と負端は離れていることが多い．図1でも，(U_2^+, V_2^+)，および(U_2^+, V_2^+)がそれぞれ異なる一つのコミュニティとなっていることがわかる．また，(U_3^-, V_3^-)も一つのクラスタに見える．しかし，例えば(U_3^+, V_3^+)は一つの固まりには見えない．

このように，特異値分解によるコミュニティの構成は，クラスタリングとして使えるとは限らない．クラスタになっていたとしても，そのコミュニティの意味を求めるには別の手段が必要となる．そして特異値分解に計算コストがかかる．コミュニティ発見のこの手法には，この三つの問題点がある．計算のネックとなる特異値分解の高速化については，[Drineas 99, Kobayashi 00] などの研究がある．次章では，最初の2点について，具体的な実験を通して詳しく述べる．

3. 解析実験について

我々はリンク情報データベースとその可視化システムを開発しリンク情報を使ってWeb空間の解析を行っている [Ikeda 00]．本章ではコミュニティの構成とその解釈についての実験について述べる．実験の前半部分は [Kleinberg 98] と同様，次のような手順で行った．まず，100件程度のURLから成るroot setから始め，順リンク，逆リンクをそれぞれ100個以内でたどり，N = 3 000 ～ 10 000程度のURLからなるbase setを構成する．次にbase setのページで決まるWebグラフの隣接行列に対し，特異値分解を行い，固有値，固有ベクトルを求める．固有値の分布を分析し，主成分の数Kを推定する．主固有ベクトル以外の$K-1$個のそれぞれの固有ベクトルについて，成分の分布を調べ，正端コミュニティ$C_i^+ = (U_i^+, V_i^+)$と負端コミュニティ$C_i^- = (U_i^-, V_i^-)$を求める．このようにして求めたコミュニティが果たして，クラスタとなっているかどうかのテスト基準として，実験の後半部分では，有村ら [Arimura 98] のアルゴリズムを用いて，C_i^+とC_i^-を区別する特徴的キーワード抽出を行った．

本稿では，(1)“＋java＋introduction”に対するキーワード検索結果（Alta-Vista），(2)“WBT”に対するキーワード検索結果（google日本語），(3) ヒューマンインタフェースに関するリンク集，(4) インターネットと知識推論に関するリンク集，の四つのroot setについて実験結果を紹介する．(3)，(4) は人工知能学会の「私のブックマーク」*1にあるリンク集である．表1にroot set，base setの要素数を表した．

表1 対象URLの個数

	root set	base set	リンク数
(1)	190	9169	9464
(2)	170	1212	1478
(3)	87	9169	11410
(4)	25	3338	3525

(1) のJava入門について，固有値を大きい順にプロットしたグラフは図2のようになる．最大固有値205.49の次は，150番目あたりまでは固有値が50でほぼ同じになっている．これは因子分析として考えると，主成分の個数が多い成熟した空間といえる．

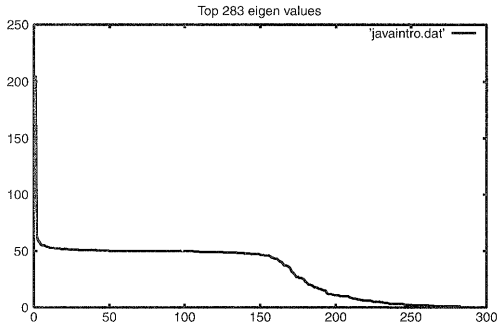

図2 Java入門の空間の固有値

(1) の上位10個の固有値に対する固有ベクトルについて，オーソリティ，ハブの各固有の正端，負端の個数を求めたものが表2である．オーソリティ（0番の正端）

*1 wwwsoc.nacsis.ac.jp/jsai/journal/mybookmark/

表1 対象URLの個数

	オーソリティ		ハブ		固有値
i	+	−	+	−	
0	52	0	6	0	205.497275
1	41	0	0	148	62.912842
2	0	2	0	91	59.000000
3	9	30	16	149	58.063427
4	32	9	96	0	55.726074
5	52	18	18	164	55.364432
6	17	29	98	2	55.022854
7	26	49	151	198	54.450700
8	35	38	13	245	53.984467
9	31	45	202	106	53.447097
10	34	41	114	67	53.066337

としてはinternet.com, javascript.com, webdeveloper.comなど29のドットコムのサイトの52のURLが得られた．特に52の内25件についてはinternet.comのページであった．ハブは，wdvl.comが2件，stars.comが2件，ほかにjavaboutique. internet.com, webdevelopersjournal.comがそれぞれ1件，合計6件であった．非主要固有ベクトルの1番目の正端には，sun, netscape, ibm, javacats, javaworld, microsoft, zdnet, stars, wdvl, webreviewなどのドットコムが24件，clarku, cs. princeton, ecst.csuchico, ei.cs.vt, genomve-www.stanfordなどの大学の個人のページが10件，その他7の合計41件であった．2番目の負端のオーソリティはeastjava.comのサイトのページが二つのURLだけ，抽出できたキーワードは，indonesia, llpadding, border, ffffff, about, ation, ction, ional, tions, from, ight, ment, ness, page, site, this, tingだった．特徴的なキーワードはindonesiaしかない．つまり，2番目の負端はインドネシアのジャワ島についての観光案内といえる．キーワード抽出について他に，11番目，12番の負端がほとんどフランス語であることが特徴的だった．16番の負端には，hokkaido, nikkei, asahi, keio, kobeなどの日本語が見受けられた．

キーワードWBTについてgoogleで日本語での検索結果107件をroot setとした場合の固有値のグラフが図3である．最大固有値に対する固有ベクトルの正端は，想定していたWeb Based Trainingについて，高度映像処理センター*2のサイトに2件と日本パーソナルコンピュータソフトウェア協会*3にあった1件*4であった．しかし，他にウィンドウズベースターミナルについての個人のページ*5も含まれていた．1番目のオーソリティの正端は，日本パーソナルコンピュータソフトウェア協会の先ほどのページと，残りの2件は気象業務支援センターの気象新聞Weather Business Today *6に関連するものだった．これは，1番目の正端の141件のハブの中に，一つのページ中に気象とパソコンの両方へのリンクを含む資格についてのページが多数あり，その結果2種類のWBTが同じコミュニティと認識されたと考えられる．このように，正端と負端が必ずしも内容的に意味のあるクラスタと解釈できるとは限らない．リンク情報を使ったコミュニティ発見では，リンクをたどる際のこのようなテーマのズレをいかに防ぐかが重要な課題といえる[村田 00, Murata 01]．

WBTについての検索結果の空間については，固有値のグラフ（図3）から，上位5個の固有値の累積寄与率が80％となり，この空間の次元が5と推定できる．一方，3番目，6番目のオーソリティに含まれていたボトルメール（What's BoTtle mail）とWorld Beach Trip（WebCamera関連のページ）まで入れると，WBTについて合計5種類の特徴的キーワードが得られる．また，Java入門についての固有値のグラフ（図2）と比較すれば，前者が後者より格段に広い世界であることがわかる．

ヒューマンコンピュータインタフェースに関するリンク集についての固有値のグラフ（図4）は，主要主固有値と非主要固有値の差が大きくJava入門の固有値グラフと類似している．一方，インターネットと知識推論についてのリンク集の固有値グラフ（図5）は，WBTの

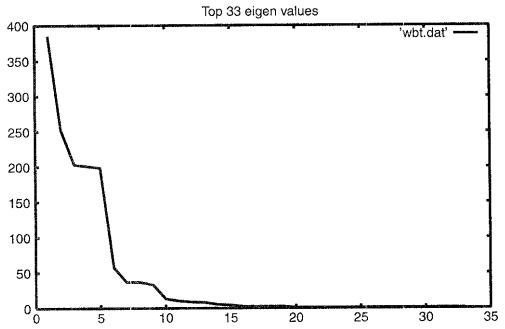

図3 WBTで決まる空間の固有値

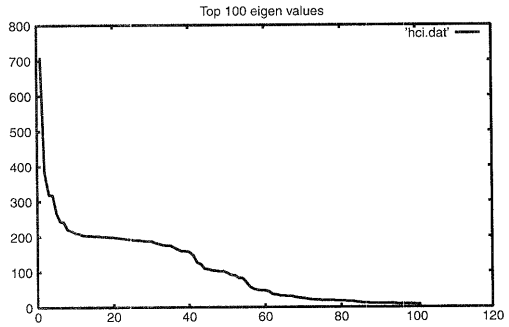

図4 ヒューマンコンピュータインタフェースの空間の固有値

*2 www.avcc.or.jp
*3 www.psa.or.jp
*4 www.jpsa.or. gl 1102.htm
*5 www.awave.or.jp/home/kenj1756/pc/pc4.html
*6 www.jmbsc.or.jp/wbt/new.htm

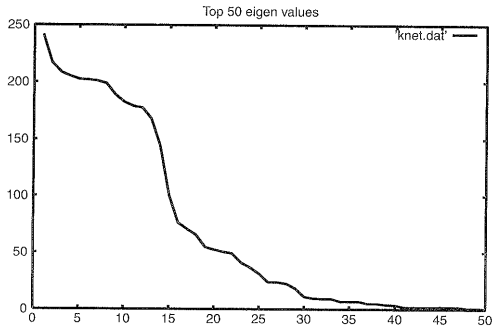

図5 インターネットと知識推論の空間の固有値

それに似ている．このように，空間の広がりの解析については，隣接行列の特異値分解による解析が有効であることがわかる．しかし，同じ正端に異なる内容のページが含まれることもあり，一般的にコミュニティを意味的なクラスタとみなせるとは限らない．また，常に正端と負端の両方があるわけではないので，コミュニティの特徴的キーワードの抽出には，正端と負端の比較だけでなく，異なる固有値についてのコミュニティ間の比較など必要で，コミュニティの解釈はまだ未解決といえる．

4. Web空間の大局的解析

Webページはつくった人の考えでそれらの間にリンクが張られるので，ランダムなグラフになるように思われる．しかし，Webグラフはランダムなグラフとは異なる性質をもつことが知られている．Barabasiら[Albert 99]は，NotreDame大学の全ページ（約33万ページ，146万リンク）からなるWeb空間を分析し，リンクがベキ分布に従うことを示した．すなわち，出次数（outdegree）がxであるページの個数yは，ある定数a, bについて，$\log(y) = a - b * \log(x)$となる．$x, y$の代わりにlogをとった値$x, y$で考えると，$Y = A - B * X$となる．$(X, Y)$をプロットすると，$x$切片が$A$，傾きが$-B$の直線となる．具体的に係数$B$の値は$-2.45$である．これは，あるページから$k$個のリンクが出ている確率が$k^{-2.45}$であることを意味する．入次数（indegree）についてBは-2.1となる．このことから，Web空間は，どのページもたかだか19回のリンクで互いにつながっている“small world”であることを示した．リンクがベキ分布に従うことは，[Kleinberg 99, Kumar 99]も独立に発見している．最近，Broderら[Broder 00]は，より大規模なWebデータ（2億ページ，15億のリンク）についてWeb空間の構造を解析し，蝶ネクタイ理論（Bow-Tie Theory）を発表している．それによれば，非常に強力に連結する「結び目の部分」，結び目にたどり着けるが逆はできない「始まりのページ」，結び目からたどれるが逆はできない「終りのページ」からなるほぼ同規模の三つの部分が全体の90％を占め，そして，結び目とつながっていない「つる」の部分10％を占める．LawrenceとGiles [Lawrence 99]は，複数のサーチエンジンの検索結果の重複を計算することにより，世界のWebページの総数を8億ページと推定している．日本国内を対象とする同様の実験は，来住ら[来住 99]によっても行われている．Henzingerら[Henzinger 99]は同様な手法を，サーチエンジンの質の評価に応用した．

一般的に，隣接行列の固有値をグラフの分割や連結成分の規模評価などに用いるスペクトルグラフ理論が知られている[Chun 94, Pothen 90]．これをWebグラフの解析に適用する研究動向がある．[Ng 01]では，[Kleinberg 98]によるページのオーソリティ度の安定性が対象とするWebグラフの隣接行列の最大固有値と2番目の最大固有値の差（eigen gap）により評価できることを示している．[Aiello 00]ではベキ分布に従うランダムグラフについて，この条件だけから，最大度数，ノードの総数，エッジの総数，さらに最大連結成分などの評価ができることを示している．大規模な実データが入手できるので，因子分析やグラフ理論の分野としても新しい展開が期待される．これらは理論的な研究としてだけでなく，例えば豊田，菊池[豊田 01]のような特定の領域に対する高品質なポータルサイトの効率的構築にも応用できる．

5. おわりに

Webページ間の隣接関係を表す行列の特異値分解と主成分分析により，対象とするWeb空間の構造を解明するKleinbergらの手法についての分析事例を紹介した．固有値の寄与率を評価することにより，空間の複雑さである次元が把握できることを具体例で示した．また，固有ベクトルの正端と負端を類別するキーワードの抽出を行い，特異値分解により得られるコミュニティ解釈の可能性と問題点を述べた．そして，グラフ理論を用いた新しい研究動向を紹介した．

情報社会そのものであるWeb空間の解析と，Web空間からの知識発見は，今後ますます重要な課題となる．Web空間の解析の研究成果がサイエンスやネーチャーに掲載されているように，Webは情報科学だけの研究対象ではなくなったといえるのではないだろうか．

◇ 参考文献 ◇

[Aiello 00] Aiello, W., Chung, F., and Lu, L.: A Random Graph Model for Massive Graphs, in Proc.of the 32nd ACM Symp.on Theory of Computing, pp. 171-180 (2000)

[Albert 99] Albert, R., Jeong, H., and Barabasi, A.: Diameter of the World Wide Web, Nature, Vol. 401, pp. 130-131 (1999)

[Arimura 98] Arimura, H., Wataki, A., Fujino, R., and Arikawa, S.: A Fast Algorithm for Discovering Optimal String Patterns in Large Text Databases, in Proc. of the 9th Int. Workshop on Algorithmic Learning Theory, LNAI 1501, pp. 247-261 (1998)

[Berry 95] Berry, M., Dumains, S. T., and O'brien, G. W.: Using

Linear Algebra for Intelligent Information Retrieval, SIAM Review, Vol. 37, pp. 573-595 (1995)
[Broder 00] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J.: Graph Structure in the Web, in Proc. of the 9th World Wide Web Conf. (2000)
[Chun 94] Chung, F. R. K.: Spectral Graph Theory, American Mathematical Society (1994)
[Drineas 99] Drineas, P., Frieze, A., Kannan, R., Vempala, S., and Vinay V.: Clustering in Large Graphs and Matrices, in Proc. ACM-SIAM Symp. on Discrete Algorithms, pp. 291-299 (1999)
[Henzinger 99] Henzinger, M. R., Heydon, A., Mitzenmacher, M., and Najork, M.: Measuring Index Quality Using Random Walks on the Web, in Proc.of the 8th International World Wide Web Conf., pp. 213-225 (1999)
[Ikeda 00] Ikeda, D. and Hirokawa, S.: Extracting Positive and Negative Keywords for Web Communities, in Proc. of the 2nd International Conf. on Discovery Science, Springer LNAI, Vol. 1967, pp. 299-303 (2000)
[来住 99] 来住伸子, 大森貴博, 笹塚清二, 近藤晶子, 水谷正大, 小川貴英: 統計的推定による日本語Webの調査, インターネットコンファレンス99論文集, pp. 21-28 (1999)
[Kleinberg 98] Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment, Proc. ACM-SIAM Symp. on Discrete Algorithms, pp. 668-677 (1998)
[Kleinberg 99] Kleinberg, J., Kumar, S. R., Raphavan, P., Rajagopalan, S., and Tomkins, A.: The web as a graph: Measurements, models and methods, Proc. of the International Conference on Combinatorics and Computing (1999)
[Kobayashi 00] Kobayashi, M., Dupret, G., King, O., and Samukawa, H.: Efficient Estimation of Singular Values for Searching Very Large and Dynamic Web Databases, IBM RT-0347 (2000)
[Kumar 99] Kumar, S. R., Raghavan, P., Rajagopala, S., and Tomkins, A.: Extracting Large Scale Knowledge Bases from the Web, in Proc. IEEE International Conference on VLDB, pp. 639-650 (1999)
[Lawrence 99] Lawrence, S., and Giles, C. L.: Accessibility of information on the Web, Nature, Vol. 400, pp. 107-109 (1999)
[Murata 00] Murata, T.: Discovery of Web Communities Based on the Co-occurence of References, in Proc. of the 3rd International Conf. on Discovery Science, Springer LNAI, Vol. 1967, pp. 65-75 (2000)
[村田 01] 村田剛志: 参照の共起性に基づくWebコミュニティの発見, 人工知能学会論文誌, Vol. 16, No. 3, pp. 316-323 (2001)
[Ng 01] Ng, A. Y., Zheng, A. X., and Jordan, M. I.: Link Analysis, Eigenvectors and Stability, in Proc. of IJCAI 2001 (2001)
[Pothen 90] Pothen, A., Simon, H. D., and Liou, K. -P.: Partitioning Sparse Matrices with Eigenvectors of Graphs, SIAM J. Matrix Anal. Appl., Vol. 11, No. 3, pp. 430-452 (1990)
[豊田 01] 豊田正史, 菊地時夫: リンク解析を用いた地球環境ポータルサイト構築の試み, データ工学ワークショップDEWS2001 (2001)

2001年5月7日 受理

著者紹介

廣川 佐千男 (正会員)

1977年九州大学理学部数学科卒業. 1979年同大学院修士課程修了. 1979年静岡大学工学部情報工学科助手, 1988年九州大学教養部情報科学教室助教授, 1994年同大学理学部物理学科助教授, 1996年同大学院システム情報科学研究科教授, 1997年同大学大型計算機センター教授, 2000年同大学情報基盤センター教授, 現在に至る. 博士 (理学). Web空間の構造解析, 半構造化データの分析と統合, 型理論と証明論の研究に従事. 情報処理学会, 電子情報通信学会, 日本ソフトウェア科学会, 数学会, ASL各会員.

池田 大輔

1994年九州大学理学部物理学科卒業. 1996年同大学院総合理工学研究科情報システム学専攻修士課程修了. 同年同大学院システム情報科学研究科情報理学専攻博士課程退学. 同年同大学大型計算機センター助手. 2000年同大学情報基盤センター講師. 現在に至る. テキストマイニング, Webマイニング, 形式言語理論の研究に従事.