社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DBS－139 (11)
2006－ＦＩ－ 83 (11)
2006／5／31

Web構造マイニングとWebコミュニティ発見

村田剛志
東京工業大学 大学院情報理工学研究科 計算工学専攻
〒152-8552 東京都目黒区大岡山 2-12-1 W8-59
murata@cs.titech.ac.jp

あらまし：Webは膨大であり、2005年1月現在、インデックス可能なWebページは115億ページ以上と推定されている。多くのWebページにおいては、関連性のある他のページにハイパーリンクを張ることで参照を行っている。個々のWebページを頂点、ハイパーリンクを辺とみなすと、Webページ集合は全体としてグラフ構造とみなすことができる。このようなハイパーリンクの構造は、時としてコンテンツが表現する以上の情報をもたらすことが少なくない。ハイパーリンクのグラフ構造に注目してページ間の関連性を見出したり、重要ページのランキングを行ったりするアプローチはWeb構造マイニングと呼ばれ、近年盛んに研究されてきている。本稿ではWeb構造マイニングにおけるアプローチを紹介するとともに、二部グラフに基づいた筆者のWebコミュニティ発見の手法について述べる。

キーワード：Web構造マイニング、ハイパーリンク、グラフ構造、Webコミュニティ

Web Structure Mining and Discovery of Web Communities

Tsuyoshi Murata
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
W8-59 2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552, Japan
murata@cs.titech.ac.jp

Abstract: The Web is huge; the estimated size of indexable Web is more than 11.5 billion pages as of January 2005. Most of the Web pages refer related other pages with hyperlinks. If we regard each Web page as a node and each hyperlink as an edge, a collection of Web pages can be regarded as a graph. Such hyperlink structures often bring more information rather than information from the contents of Web pages. Approaches focusing on graph structures of hyperlinks, such as finding relations among pages and ranking important pages, are called Web structure mining. This paper explains approaches of Web structure mining and our method for finding Web communities based on bipartite graphs.

Keywords: Web structure mining, hyperlink, graph structure, Web community

1. はじめに

Web マイニングは、自然言語処理や機械学習、データマイニングなどの人工知能の分野にとどまらず、情報検索やデータベースなど幅広い分野と関連する複合的な研究分野である。注目する Web データの種類によって、Web マイニングは以下の 3 つに分類される。

1) Web ページのコンテンツに注目し、自然言語処理やデータベースのアプローチを用いて、テキストマイニングによる情報抽出や半構造データにおける検索のモデル化などを目指す Web 内容マイニング

2) Web ページ間を結ぶハイパーリンクによって構成されるグラフ構造に注目し、関連ページの発見や重要ページのランキング、グラフ構造のモデル化などを目指す Web 構造マイニング

3) Web ページの閲覧によって生じる(サーバー側やクライアント側の)ログデータに注目し、ユーザプロファイルの学習やサイトにおける閲覧パターンの学習などを目指す Web 利用マイニング

本稿では、ハイパーリンクのグラフ構造に注目する Web 構造マイニングについて述べるとともに、筆者の Web コミュニティ発見の試みについて紹介する。

2. Web 構造マイニング

Web ページの多くはハイパーリンクによって他のページと結合している。Web ページ集合は、個々の Web ページを頂点、ページ間のハイパーリンクを辺としたグラフ構造とみなすことができる。このような構造に注目した Web 構造マイニングにおける目標としては、以下のものがあげられる。それぞれについて以下で説明する。

1. 重要ページのランキング
2. 関連ページからなる Web コミュニティ発見
3. Web グラフのモデリング
4. Web ページのサンプリング

2.1 重要ページのランキング

Web から情報を得ようとしてサーチエンジンでキーワード検索をするユーザにとっては、キーワードに関連した内容のページが検索結果の上位にランキングされることが望ましい。一方、Web ページ作成者の多くは自分のページが検索結果の上位に来ることを望んでおり、見えない色で単語の羅列を表示するなどの恣意的な操作が行なわれているページもある。そのため、Web ページ上に出現するキーワードに基づいたランキングでは、ユーザにとって好ましくない結果となる可能性がある。

ハイパーリンクの多くは、その参照先のページ内容に対する支持を表すものであると考えられる。ハイパーリンクのグラフ構造に基づいた Web ページのランキングにおいては、あるページの重要度を、そのページとハイパーリンクで結合している周囲のページの重要度を使って決定する。従って、上述のような操作の影響を受けにくく、サーチエンジンにおける妥当なランキングを実現する上で好ましいものである。そのようなランキングの代表的な例として、特定のトピックのページに関するランキングアルゴリズムである HITS [Kleinberg 98]や、トピックに依存しないランキングアルゴリズムである PageRank [Page 98]がある。

HITS においては Web ページの有用性を表す評価基準として、特定のトピックに関する情報の豊富さを表す authority と、authority へのリンクの豊富さを表す hub を導入している。authority として価値の高いページへリンクを張っているページは hub としての価値が高く、また hub として価値の高いページからリンクを張られているページは authority としての価値が高いと言えるため、両者は相互再帰的に定義することができる。ハイパーリンクによるページ間の結合関係は隣接行列によって表される。隣接行列とは、ページ i からページ j へのハイパーリンクが存在する場合に(i,j)の成分が 1 で、それ以外は 0 であるような n ×n の正方行列である。個々のページの authority 値と hub 値の計算は、隣接行列とその転置行列の積の主固有ベクトルを求める計算に帰着する。

一方、PageRank は「多くの良質なページからリンクされているページはやはり良質なページである」という考えに基づいた、トピックに依存しないランキングアルゴリズムである。あるページ

のPageRank値は、そこにハイパーリンクを張っている他のページのPageRank値によって決まる。PageRank値の高い有用なページからハイパーリンクを張られていたり、多くのページからハイパーリンクを張られていたりすると、そのページのPageRank値が高くなる。各ページのPageRank値はランダムにハイパーリンクをたどる閲覧者がページを訪れる確率に対応している。具体的なPageRankの計算は、隣接行列を転置し各列を非零要素数で割った推移確率行列の主固有ベクトルを求める計算に帰着される。PageRankによるランキングは、サーチエンジンGoogleにおけるランキングの一部として利用されている。

2.2 関連ページからなるWebコミュニティ発見

Webコミュニティという単語は様々な文脈において用いられるが、本稿ではハイパーリンクによって密に結合した関連Webページ集合という意味で用いることにする。Webコミュニティを発見する手法としては、固定したグラフ構造を探索する手法と、密な部分グラフ構造を抽出する手法の二つに大まかに分けることができる。それぞれについて以下で説明する。

固定したグラフ構造の探索

ハイパーリンクが特定のグラフ構造を構成するようなWebページ集合が意味的なまとまりをもつならば、Webのスナップショットデータからそのグラフ構造を探索することによってWebコミュニティの発見を行なうことができる。そのようなアプローチの代表的な例として、Kumarらによるtrawlingの研究[Kumar 99]がある。この研究においては、二つの頂点集合FとCから構成され、Fの各頂点uからCの各頂点vへの有向辺が存在するような完全二部グラフを、興味を共有するページからなるWebコミュニティであるとしている。Kumarらは約2億ページの大規模スナップショットデータから、$|F|=i$, $|C|=j$となるサイズ(i,j)の完全二部グラフを高速に探索するための枝刈りの手法を提案し、個数の分布を実験結果によって示している。$i_1<i_2$のとき、サイズ(i_2,j)の完全二部グラフの部分集合(i_1,j)も完全二部グラフであることから、数え上げにはアプリオリアルゴリズムを用いることができる。jを固定してまず(1,j)を見出し、その結果を利用して順に(2,j)、(3,j)、…を見出すことで数え上げを行なっている。またKumarらは得られたWebコミュニティの質を評価するために、ランダムに選択した400個の二部グラフを人手で調べた結果、共通性のないWebページ集合によって偶然に形成されたものが4%あり、データ収集から18ヶ月後のWeb上においては既に存在しないものが約30%あったが、それ以外のものは実際に関連性のあるWebページ集合であった。

密な部分グラフ構造の抽出

上述のアプローチとは別に、与えられたWebデータのグラフ構造を分割するなどして、密な部分グラフであるWebコミュニティを見出すアプローチもある。Flake[Flake 02]らは、全頂点集合Vの部分集合Cにおいて、Cの各要素vがV-Cの頂点よりもCの頂点とより多くリンクしていることをWebコミュニティの定義としている。このようなWebコミュニティを見出すことは一般にはNP完全のグラフ分割問題であることから、ネットワーク理論における最大流問題の枠組みでこの問題を捉え直す。種となる頂点集合を想定し、それを含むようなWebコミュニティを効率的に見出す問題を考える。

グラフにおける辺を水、頂点を接合点とみなし、各辺には正の容量が付与されているとする。頂点集合s(source)とt(sink)が与えられたとき、最大流問題は各辺の容量を越えることなくsからtへの最大流を求める問題であり、これはsとtを分離する最小容量カット問題と同値であることが知られている。

2.3 Webグラフのモデリング

近年、多くの実ネットワークにおいて、次数kの分布P(k)が$k^{-\gamma}$に比例することがBarabasiらによって指摘されてきている[Barabasi 02]。このような次数分布のベキ則は、特徴的なスケール(縮尺)が無いという意味でスケールフリーと呼ばれる。スケールフリー・ネットワークとしては、ハ

イパーリンクによるWebの他に、物理的なインターネットや、映画俳優の共演者のネットワーク、タンパク質の反応のネットワークなどが挙げられている。このようなスケールフリー・ネットワークを生成するモデルとして、成長と優先的選択の二つのルールを持つBAモデルや、一般化ランダム・グラフ、コンフィギュレーションモデルなどが提案されている[増田 05]。またWebのグラフ構造の理解とそれを利用したアルゴリズムのためのワークショップ(WAW)が開催されている[Leonardi 04]。

2.4 Webページのサンプリング

Webは膨大であり、全ページを収集するのは困難である。Webページにおける統計的な調査を行なうためには、サンプリングを行なう必要がある。サンプリングの手法としては、ハイパーリンクをたどるランダムウォークによるサンプリングや、IPアドレスによるサンプリングがある。前者のランダムウォークにおいては、PageRank値の高いページほど訪れやすいため、訪れたページをPageRank値に反比例した確率でサンプリングすることでランダムサンプリングを行う手法が提案されている。Lawrence らはランダムサンプリングによって、Webにおけるドメイン(com, eduなど)の割合や、内容(Scientific, Governmentなど)の割合を推定している[Lawrence 99][Henzinger 04]。

3. 二部グラフによるWebコミュニティの発見

膨大なWebページの中からの情報獲得を支援するための試みとして、著者はハイパーリンクによるグラフ構造に基づいてWebページの関連性を見出す研究を行なっている。ハイパーリンクによるco-citationを基にWebページ間の関連性を見出し、関連するページが近接するようなグラフを生成する視覚化システム[Murata 99]や、Webのグラフ構造中の完全二部グラフをコミュニティとみなし、サーチエンジンから得られるbacklink情報を基にWebコミュニティを発見するシステム[村田 01]を構築している。

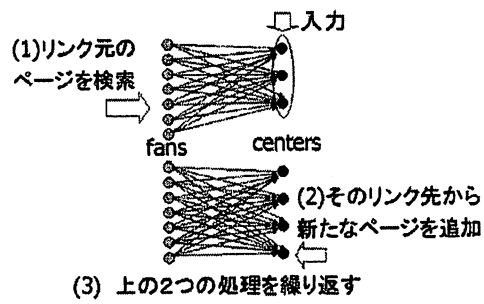

図1:Webコミュニティ発見手法

著者が提案したWebコミュニティ発見手法[村田 01]においては、ハイパーリンクによるWebのネットワークから局所的な情報をもとに完全2部グラフを見出す。入力URLをcentersとし、その全てのURLに対してハイパーリンクを張っているようなWebページ集合をサーチエンジンのbacklink探索によって獲得してfansとする。次に、そのページからのリンク先をすべて抽出し、最も多くのfansが参照しているURLをcentersに追加し、そのcentersを用いて上述の処理を繰り返すというものである。KumarらはWebのスナップショットデータを用いて二部グラフの探索を行ったが、大規模なWebデータを収集し維持することは困難である。サーチエンジンは比較的新しいWebデータを保持していると考えられることから、本手法ではWebデータ獲得のためのいわば実験としてサーチエンジンでの検索を行っている。このアイディアを発展させて、トピックに関するより主要なページからなるコミュニティを得るための洗練やその視覚化なども行っている[村田 02] [Murata 03] [Murata 05]。

4. おわりに

本稿では、ハイパーリンクのグラフ構造に注目したWeb構造マイニングの紹介を行うとともに、局所的な二部グラフ探索による筆者のWebコミュニティ発見手法について述べた。Webコミュニティ発見手法については、筆者のものも含めてZhangの本に数多く紹介されている[Zhang 06]。

Web構造マイニングに関する最新の研究の会

議やワークショップとしては、WWW conference [WWW 05]や、WebKDD [Nasraoui 05]、LinkKDD [LinkKDD 05]などがあげられる。またWebに限らず構造をもったデータのマイニングはLink Mining[Getoor 05]と呼ばれている。Getoorは一般的なLink Miningのタスクとして、以下のような分類を行っている。

1. Object-Related Tasks
   (a) Link-Based Object Ranking
   (b) Link-Based Object Classification
   (c) Object Clustering (Group Detection)
   (d)Object Identification (Entity Resolution)
2. Link-Related Tasks
   (a) Link Prediction
3. Graph-Related Tasks
   (a) Subgraph Discovery
   (b) Graph Classification
   (c) Generative Models for Graphs

この分類では、HITSやPageRankなどのランキングは1(a)、コミュニティ発見は1(c)、Web生成モデルは3(c)に属すると言える。Webデータを含めた構造データを扱う手法は現在盛んに研究されており、今後も幅広い応用分野を持つものとして期待される。

謝辞

本研究の一部は、科学研究費補助金若手研究(A)「Webのハイパーリンク構造のモデル化に関する研究」によるものである。

参考文献

[Barabasi 02] A. L. Barabasi: "Linked – The New Science of Networks", Perseus Publishing, 2002.

[Bharat 98] K. Bharat, M. Henzinger: "Improved Algorithms for Topic Distillation in a Hyperlinked Environment", Proc. of the 21st Int'l ACM SIGIR Conf. pp.104-111, 1998.

[Broder 00] A. Broder et. al.: "Graph structure in the Web", Proceedings of the Ninth World Wide Web Conference, 2000.

[Chakrabarti 98] S. Chakrabarti et. al.: "Experiments in Topic Distillation", Proc. of ACM SIGIR workshop on Hypertext Information Retrieval on the Web, 1998.

[Chakrabarti 99] S. Chakrabarti, et. al.: "Mining the Web's Link Structure", IEEE Computer, Vol.32, No.8, pp.60-67, 1999.

[Flake 02] G. W. Flake, S. Lawrence, C. L. Giles, G. M. Coetzee: "Self Organization and Identification of Web Communities", IEEE Computer, Vol.35, No.3, pp.66-71, 2002

[Getoor 05] L. Getoor, C. P. Diehl: "Link Mining: A Survey", SIGKDD Explorations, Vol.7, No.2, pp.3-12, 2005.

[Gibson 98] D. Gibson, J. Kleinberg, P. Raghavan: "Inferring Web Communities from Link Topology", Proceedings of the 9th Conference on Hypertext and Hypermedia, 1998.

[Henzinger 01] M. Henzinger: "Hyperlink Analysis for the Web", IEEE Internet Computing, Vol.5, No.1, pp.45-50, 2001.

[Henzinger 04] M. Henzinger, S. Lawrence: "Extracting knowledge from the World Wide Web", Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 101, Suppl. 1, pp.5186-5191, 2004.

[Kleinberg 98] J. Kleinberg et. al.: "The Web as a Graph: Measurements, Models, and Methods", Proc. of COCOON '99, LNCS 1627, pp.1-17, Springer, 1999.

[Kosala 00] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", ACM SIGKDD Explorations, Vol.2, No.1, pp.1-15, 2000.

[Kumar 99] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: "Trawling the Web for Emerging Cyber-Communities", Proceedings of the 8th WWW Conference, pp.403-416, 1999.

[Lawrence 99] S. Lawrence, C. L. Giles: "Accessibility of information on the Web", Nature, Vol.400, No.6740, pp.107-109, 1999.

[Leonardi 04] S. Leonardi (Eds.): "Algorithm

and Models for the Web-Graph", Proceedings of the Third International Workshop (WAW 2004), LNCS 3243, Springer, 2004.
[LinkKDD 05] LinkKDD Organizers: "Link Discovery: Issues, Approaches and Applications", http://www.isi.edu/LinkKDD-05/, 2005.
[Nasraoui 05] O. Nasraoui, O. R. Zaiane, M. Spiliopoulou, B. Mobasher, B. Masand, P. S. Yu: "WebKDD 2005 – Web Mining and Web Usage Analysis Post-Workshop report", IGKDD Explorations, Vol.7, No.2, pp.139-142, 2005.
[Murata 99] T. Murata: "Machine Discovery Based on the Co-occurrence of References in a Search Engine", Proceedings of Discovery Science (DS99), LNAI 1721, pp.220-229, Springer, 1999.
[Murata 03] T. Murata: "Visualizing the Structure of Web Communities Based on Data Acquired from a Search Engine", IEEE Transactions on Industrial Electronics, Vol. 50, No. 5, pp.860-866, 2003.
[Murata 05] T. Murata: Graph Mining Approaches for the Discovery of Web Communities, in T. Washio, J. N. Kok, L. D. Raedt eds., Advances in Mining Graphs, Trees And Sequences (Frontiers in Artificial Intelligence and Applications), pp.199-208, IOS Press, 2005.
[Page 98] L. Page, S. Brin, R. Motwani, T. Winograd: "The PageRank Citation Ranking: Bringing Order to the Web", Online manuscript, http://www-db.stanford.edu/~backrub/pagerank sub.ps, 1998.
[WWW 05] WWW2005 Organizers: "Proceedings of the 14th International World Wide Web Conference", ACM Press, 2005.
[Xhang 06] Y. Zhang, J. X. Yu, J. Hou: "Web Communities – Analysis and Construction", Springer, 2006.
[増田 05] 増田直紀, 今野紀雄: "複雑ネットワークの科学", 産業図書, 2005.
[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", 人工知能学会誌, Vol.16, No.3, pp.316-323, 2001.
[村田 02] 村田剛志: "ハイパーリンクのグラフ構造に基づく Web コミュニティの洗練", 人工知能学会誌, Vol.17, No.3, pp.322-329, 2002.