1F1-03

Webコミュニティの中心性
Centrality of Web Communities

村田剛志
Tsuyoshi Murata

国立情報学研究所
National Institute of Informatics

科学技術振興事業団
Japan Science and Technology Corporation

Several attempts have been made for Web structure mining whose goals are to discover or to rank related Web pages based on the graph structure of hyperlinks. Discovery of Web communities, groups of related Web pages, is important for assisting users' information retrieval from the Web. As is the case with human communities, Web communities are not uniform and are overlapped with one another. This makes the identification of Web communities difficult. Discovery of the centrality of Web communities is closely related with the discovery of their boundaries. This paper proposes a method for identifying Web communities from some positive and negative examples. Since the boundary of a Web community is hard to define only from positive examples, negative examples are used for limiting its boundary from outer side. Experimental results show that our new method is effective for discovering Web communities' boundaries in some cases.

1. はじめに

ハイパーリンクのグラフ構造に注目する Web 構造マイニングにおいて、興味を共有する Web ページ集合(Web コミュニティ)を発見するための研究が行なわれてきている。ハイパーリンクによって形成されるグラフ構造に基づいてページ間のまとまりを見出すことは、類似ページの推薦などを行なう上で有用である。

Web コミュニティを発見する研究としては Kumar らの trawling[Kumar 99]や Flake らの手法[Flake 02]をはじめ、様々な試みがなされてきている。前者は二部グラフ構造に注目してマイニングを行なうものであり、後者はネットワーク理論における定理を適用したものである。一般に、一つの Web ページが複数のトピックに関わっていることは少なくないため、Web コミュニティは相互に重なりあっていると考えられる。したがって、ハイパーリンクのグラフ構造だけから Web コミュニティを完全に客観的に切り出すことは困難である。Web コミュニティにおける関連性やまとまりをとらえる観点によって、最終的に得られる Web コミュニティの範囲は大きく異なってくると考えられる。例えば、自動車という大きなトピックには関係のある Web ページでも、○○会社の自動車などといったより細分化されたトピックにおいては関係がなくなったりするように、Web コミュニティの内と外を判断するための観点なり関連性なりを考慮する必要がある。従来の Web コミュニティ発見手法において得られる Web コミュニティにおいては、そのような点について検討が十分でないと考えられる。

本稿では、Web コミュニティを見出すにあたり、そのコミュニティに属するページ集合(正例)とともに、その Web コミュニティには属さないような Web ページ集合、いわば負例を明示的に与えて発見を行なう手法について検討する。このような条件設定により、Web コミュニティにおけるページ間の内容的な関連性をユーザ側からの入力でインタラクティブに決めていくようなシステムを実現することが可能になり、正例と負例の決め方に対応して Web ページのグラフ構造上での両者の関連性の度合いを理解しやすいものにすることができる。

連絡先:〒101-8430 東京都千代田区一ツ橋 2-1-2 国立情報学研究所情報学基礎研究系 村田剛志, tmurata@nii.ac.jp
http://research.nii.ac.jp/~tmurata/

2. Webコミュニティの関連研究

2.1 Webコミュニティの発見

Web コミュニティ発見の研究としては、大きく分けて二つのアプローチがある。固定したグラフ構造を Web のスナップショットデータから探索する手法と、与えられたグラフから密な部分グラフを抽出する手法である。前者のアプローチとしては、Kumar らの trawling が代表的なものである。Kumar らは、興味を共有するページ集合のハイパーリンクは完全二部グラフを構成することに注目し、Web の大規模スナップショットデータからサイズを固定した完全二部グラフ構造を高速に探索する手法を提案して実際に実験を行なっている。得られた Web コミュニティをランダムサンプリングして、その内容を人手で分析した結果、そのようにして得られた Web コミュニティの大部分が関連性のあるページ集合であることが確認できている。

また、後者のアプローチとして、Flake らはネットワーク理論における最大流最小カット定理を Web のグラフに適用することによって、Web コミュニティの内側と外側とを分ける境界を見出している。このアプローチにおいては、Web コミュニティの種として与える頂点の個数や流量の係数等を調整するなどの工夫が必要になっている。また、Girvan らは、グラフ構造における辺 betweenness に注目して密な部分グラフ構造を切り出すアプローチを提案し[Girvan 01]、Tyler らはそのアプローチを電子メールのやり取りから得られる人間間のグラフ構造に適用してコミュニティを見出している[Tyler 03]。現在のところ、異なるアプローチによって得られる Web コミュニティの質などを比較するなどの試みはなく、様々なやり方が試みられている段階である。

2.2 Webコミュニティにおける中心と境界

現実の Web コミュニティは、二部グラフに限らず様々な構造を持っていると考えられる。しかしながら、グラフ構造とその意味的なまとまりとの対応が十分に明らかになっているのは現在のところ二部グラフだけである。従って、固定したグラフ構造を探索することによって Web コミュニティを見出すという Kumar らのアプローチだけでは、現実の Web コミュニティを見出すことは困難である。

また、個人のリンク集の多くは、その人が興味を持つ様々なトピックの Web ページへのリンクが共存することが多いため、グラフ構造としての Web コミュニティは互いに重なり合っていると考えられる。内容を全く考慮せず、ハイパーリンクのグラフ構造だけから Web コミュニティの内側と外側とを分ける客観的な境界を見出すことは一般には困難である。ユーザが Web コミュニティをどのような観点でとらえ、どのような用途でそれを使用するかによって、その境界は変化し得るものである。

Web structure mining の研究でも、コンテンツのトピック間の関連性を見出すなど、より粒度の細かい構造解明に向けての研究がなされつつある[Chakrabarti02]。本稿では、Web コミュニティに含まれる Web ページの例だけでなく、含まれない例を明示的に与えることによって、境界を定める作業をインタラクティブに決定することが可能になる。また、正例と負例を決定する戦略を変えることによって、最終的に得られる Web コミュニティの性質を変化させることができる。

3. 正例と負例を利用した Web コミュニティ発見

一般に一つの Web ページは複数のコミュニティに属していると考えられる。例えば、阪神タイガースの Web ページは、セントラルリーグ、日本のプロ野球、阪神グループなどの複数のコミュニティに属しており、それらが階層的な関係になっているとも限らない。Web コミュニティとそうでないものとを分けるような境界を見出すための指標として、Web コミュニティの外側にある負例を与えて、正例の類似ページ集合と負例の類似集合をそれぞれ見出して追加していき、両者の共通要素が出てきたら終了することで境界を見出すアプローチを考える。

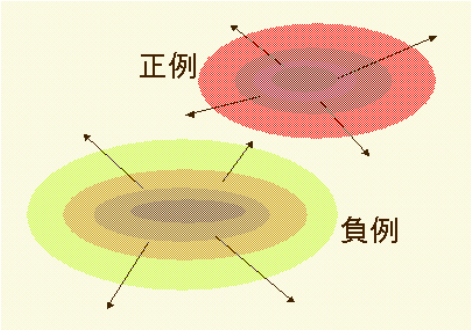

図 1 正例と負例からの Web コミュニティ発見

3.1 類似ページの発見

与えられた正例と負例それぞれの類似ページを発見する実際の処理手順としては、基本的には筆者による発見手法[村田01]を用いている。

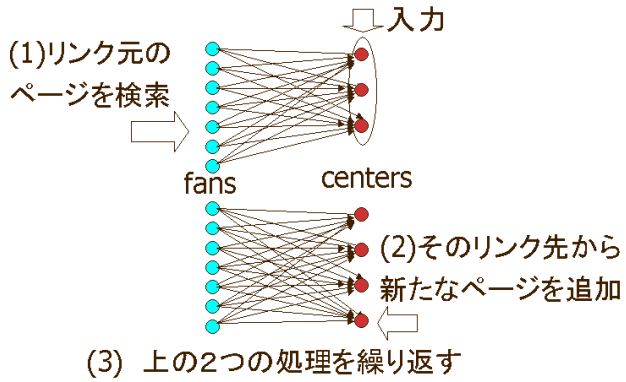

図 2 backlink を用いた類似ページの発見

この手法は、「類似したページへのリンクは共起する場合が多い」という仮定に基づいている。図 2 に示すように、入力された正例または負例(centers)について、その全てを参照しているようなページ集合(fans)を、サーチエンジンの backlink 検索を用いて獲得する。次に、その fans の HTML ファイルからリンク先を調べ、最も多くの fans が参照しているページ(図 2 下図の centers の最も下)を centers に追加する処理を繰り返すことによって類似ページを発見する。

3.2 負例と Web コミュニティの境界

前節において説明した類似ページの発見を、入力された正例と負例のそれぞれについて反復して行なう。正例・負例それぞれの類似ページを増やしていき、両者に重なりが出るか、類似ページを増やせなくなった時点で処理を打ち切る。このような終了判定を導入することにより、最終的に得られる Web コミュニティをユーザが特徴づけることが比較的容易になると期待できる。

Web コミュニティに属さないと考えられるページ集合を予め完全に明示的に指定することは、Web コミュニティを発見すること以上に困難なことであると言える。従って、本稿の手法では負例の類似ページ集合も見出して新たな負例としており、それによって早い段階で正例と負例に重なりが出て境界に到達することができると言える。また、目標とする Web コミュニティとの類似性の高い負例、いわばニアミスとなるようなページを入力として与えることによっても、Web コミュニティの境界の発見を早めることができると考えられる。

4. 実験

上述の手法を Java 言語を用いて実装し、Web コミュニティ発見システムを試作した。いくつかの実験結果によって提案手法の振る舞いを示す。

1. 正例：新聞社 (www.asahi.com, www.nikkei.co.jp, www.mainichi.co.jp) 、負例：サーチエンジン (www.google.co.jp, www.yahoo.co.jp, www.goo.ne.jp) を入力した場合、最終的に以下の結果が得られた。
   正例 : www.nikkei.co.jp, www.mainichi.co.jp, www.asahi.com, www.yomiuri.co.jp, www.sankei.co.jp, www.yahoo.co.jp
   負例 : www.goo.ne.jp, www.yahoo.co.jp, www.lyos.co.jp, www.google.co.jp, www.excite.co.jp
   個人のリンク集においては、新聞社等のニュースサイトとサーチエンジンはよく訪れるリンク先として共存することが多い。比較的早い段階で正例と負例の重なりが現われたのは、リンクのグラフ構造上、両者の関連性が高いことを表している。

2. 正例:日本の新聞社(www.asahi.com, www.nikkei.co.jp, www.mainichi.co.jp) 、負例：米国の新聞社 (www.nytimes.com, www.washingtonpost.com, www.usatoday.com)を入力した場合、最終的に以下の結果が得られた。
   正例 : www.yahoo.co.jp, www.lycos.co.jp, www.asahi.com, www.mainichi.co.jp, www.nikkansports.com, www.nikkei.co.jp, www.sanspo.com, www.sponichi.co.jp, www.excite.co.jp, auctions.yahoo.co.jp, wvexnettv.avexnet.or.jp, その他 5 件
   負例:abcnews.go.com, www.cnn.com, www.cnnfn.com, www.csmonitor.com, www.drudgereport.com,

www.latimes.com, www.nytimes.com , www.salon.com, abcradio.com, ajr.newslink.org/gusin.html, その他 5 件

正例として与えたページは先ほどと同じであるが、負例として与えた米国のページとの(グラフ構造上での)関連性が弱く、比較的広い範囲の内容を含む最終結果となっていることがわかる。

実験を行なった際に全般的に見受けられた現象としては、以下のものがあげられる。

- 入力とする正例や負例が極端に少ない(1、2 個)と、関連性のあまりないページが追加される場合が多い。
- 正例と負例との(グラフ構造上の)関連性が弱い場合には収束に時間がかかる。

5. おわりに

本稿では、種となる Web ページ集合(正例)と、それに属さないページ集合(負例)を基に、Web コミュニティを見出す手法について述べ、試作したシステムの実験結果を示した。このようなアプローチは、Web コミュニティの粒度や Web コミュニティ間の関連性を明確にしていく上での第一歩であると考えられる。今後進めていくべき方向性としては、Web コミュニティの境界を見出す上で有効な正例と負例の与え方の検討や、同一の正例を含み、かつ粒度の異なる Web コミュニティの発見のための検討などがあげられる。

参考文献

[Broder 00] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph Structure in the Web: Experiments and models, Proc. of the 9th WWW Conference, pp.309-320, 2000.

[Chakrabarti 02] Chakrabarti, S., Joshi, M.M., Punera, K., Pennock, D.M.: The Structure of Broad Topics on the Web, Proc. of the 11th WWW conferene pp.251-262, 2002.

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

[Girvan 01] Girvan, M., Newman, M.E.J.: Community structure in social and biological networks, http://arxiv.org/abs/cond-mat/0112110/, 2001.

[Kleinberg 99] Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: The Web as a Graph: Measurements, Models, and Methods, Proc. of COCOON'99, LNCS 1627, pp.1-17, 1999.

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

[Rivlin 94] Rivlin, E., Botafogo, R., Shneiderman, B.: Navigating in Hyperspace: Designing a Structure-Based Toolbox, Communications of the ACM, Vol.37, No.2, pp.87-96, 1994.

[Tyler 03] Tyler, J.R., Wilkinson, D.M., Huberman, B.A.: Email as spectroscopy: automated discovery of community structure within organizations, http://xxx.lanl.gov/arXiv:cond-mat/0303264, 2003.

[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.