大規模Webデータからのコミュニティ抽出

梅沢 晃　　山名 早人
早稲田大学理工学研究科

1 はじめに

急速に普及する WWW に対して，情報検索サービス等で用いられる Web ディレクトリにおける Web ページの分類コストは増大し，人手による分類作業には限界がきている．そこでコミュニティと呼ばれる同一トピックに関連した Web ページの集合を抽出することにより，Web ページの分類作業を自動化，または半自動化することが考えられている．本稿では，2002 年 8 月から 2002 年 10 月までの 3 ヶ月間に WWW 上から収集した約 2 億の Web ページを対象に，被リンク数を利用した Web サイト単位の決定方法を示し，キーワード等を用いずにリンク構造のみを利用した新たなコミュニティ抽出手法の提案を行う．

2 関連研究

WWW 上から全てのコミュニティを抽出する研究には，全てのコミュニティの列挙を，全ての完全二部グラフの抽出に置き換えた Trawling[1]や，稠密二部グラフの抽出によって，コミュニティの抽出を行う方法がある[2].

本稿では[2]の方法を用いて，収集した Web ページよりコミュニティを抽出する．そして，抽出されたコミュニティ同士の重なりから，コミュニティの関連を示す.

3 解析単位の決定

コミュニティ抽出の前準備として，同一作者により制作された Web ページ群の範囲を決定する．解析単位を同一作者により制作された Web ページ群とすることで，inlink が集中している入り口となる Web ページと，outlink が集中している出口となる Web ページを同一の解析対象 Web ページ群と見なすことができる.

本稿では解析対象 Web ページ群の最小単位をディレクトリとする．異なるサーバの上の解析対象 Web ページ群から閾値以上の inlink があるディレクトリを解析対象 Web ページ群とする．inlink 数が閾値未満のディレクトリは，上位のディレクトリが所属する解析対象 Web ページ群の一部であると考える(図 1)．その際，inlink は上位のディレクトリへの inlink，outlink は上位のディレクトリからの outlink と置き換える．上記の作業を解析対象 Web ページ群の数が収束するまで繰り返す.

本稿の実験では，閾値を 5 とし，異なるサーバからの inlink 数が 5 以上のディレクトリを解析対象 Web ページ群とする．表 1 に閾値を 5 としたときのリンク数，解析対象 Web ページ群あたりのリンク数を示す.

また，YAHOO!のように極端に人気のある Web ページがコミュニティに含まれると，トピックドリフトが起こるので，inlink 数が 50 を超える解析対象 Web ページ群は除去する.

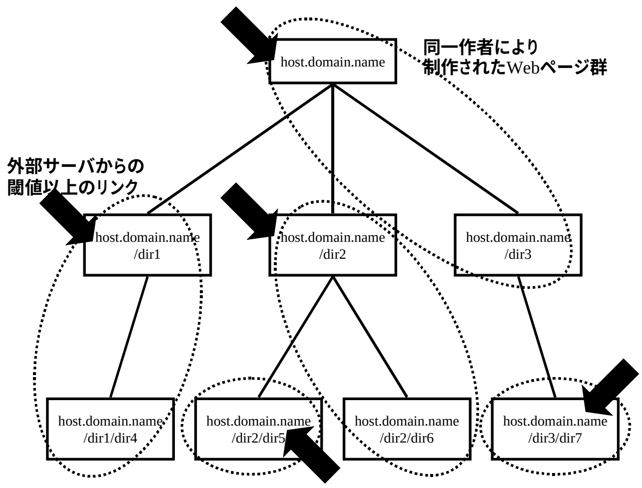

図 1:解析対象 Web ページ群

表 1:対象データのリンク数

inlink 数の閾値	0	5
総リンク数	162,432,620	41,990,000
Outlink≧1 の解析対象 Web ページ群	13,960,284	480,236
リンク数/解析対象	11.64	87.44
Inlink≧1 の解析対象 Web ページ群	25,970,887	2,096,835
リンク数/解析対象	6.25	20.03

Community Extraction from Large-scale Web Data
Akira Umezawa, Hayato Yamana
Graduate School of Science and Engineering, Waseda University

4 コミュニティの抽出

本稿で行うコミュニティの抽出方法の手順を示す．シードページには outlink を持つ全ての解析対象 Web ページ群を用いる．

(1) シードページの outlink 先と同じ解析対象 Web ページ群への outlink を持つ解析対象 Web ページ群をコミュニティ候補として抽出する．

(2) コミュニティ候補の解析対象 Web ページ群をノード，コミュニティ候補間で張られているリンクをエッジとした二部グラフを作成する．

(3) 二部グラフから inlink 数，outlink 数が閾値に満たないノードを除去し，残ったノードからリンク元の解析対象 Web ページ群をコミュニティとして抽出する(図 2)．

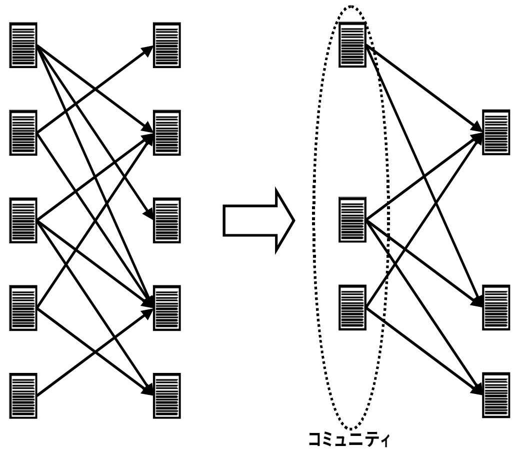

図 2:2 部グラフによるコミュニティの抽出

5 コミュニティの解析

inlink 数，outlink 数の閾値を共に 5 としてコミュニティを抽出したときのコミュニティサイズの分布を図 3 に示す．また，抽出コミュニティ数とコミュニティサイズの平均を表 2 に示す．表 2 より，抽出コミュニティ数とコミュニティサイズの平均の積が，元のデータに存在する解析対象 Web ページ群の数より大きくなっていることがわかり，多くのコミュニティで解析対象 Web ページ群の重複が起こっていることが考えられる．

表 2:抽出したコミュニティ数

Link 数の閾値	inlink≧5, outlink≧5
抽出コミュニティ数	42,819
平均コミュニティサイズ	568.74

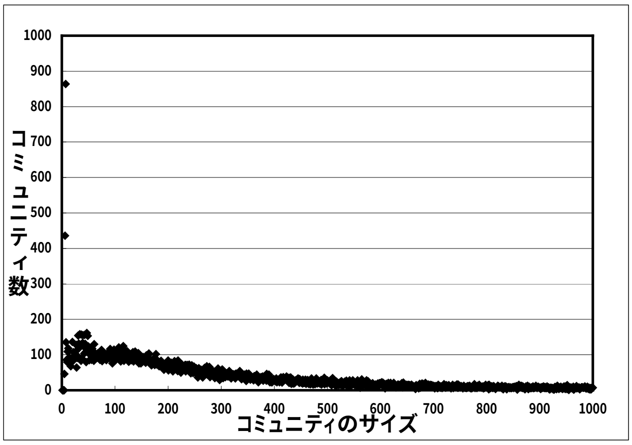

図 3:コミュニティサイズの分布

重複した解析対象 Web ページ群を持つコミュニティは，それぞれのコミュニティが持つ解析対象 Web ページ群をマージして 1 つの巨大なコミュニティであると考えることができる．また，解析対象 Web ページ群の重複数により，コミュニティ同士の関連性の強さを示すものと考えることもできる．

6 おわりに

本稿では，2 億の Web ページを対象にコミュニティの抽出を行った．収集した Web データ全体から全てのコミュニティを抽出しようとすると，多くのコミュニティで解析対象 Web ページ群の重複が起きる．

コミュニティが持つ重複の割合により，コミュニティ同士を同一のコミュニティであると考えたり，関連性の高いコミュニティであると考えることができる．

本稿で示したような膨大な量のコミュニティを実際に見て判断することは不可能であり，今後はコミュニティの評価方法について，検討する必要がある．

参考文献

[1] S. Ravi Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins: "Trawling the web for emerging cyber-communities", WWW8 / Computer Networks 31 (1999), No. 11-16, pp.1481-1493 (1999)

[2] P.Krishna Reddy and Masaru Kitsuregawa: "An approach to relate the web communities through bipartite graphs", Proceedings of the 2nd International Conference on Web Information Systems Engineering, IEEE Computer Society (2001.12), pp. 301-310