

ウェブコミュニティ抽出アルゴリズムの改良

Improvement of Web Community Extraction Algorithm

沈 垣甫[▼] 田浦 健次郎[◆] 近山 隆[▲]

Wonbo SHIM Kenjiro TAURA
Takashi CHIKAYAMA

密な二部グラフをウェブコミュニティとみなして抽出する手法が提案されている。そこでは密な二部グラフは一般に、多数のファンからリンクされているセンター、多数のセンターをリンクしているファンから成るグラフとして定義されている。以前より、そのようなグラフを抽出するための手法として、DBGやPlusDBGなどのアルゴリズムが提案されてきた。これらの手法は、シードページから二部グラフ構造を拡張して、その中から二部グラフの条件に合った構造を取り出すことにより、二部グラフを抽出している。しかしこれらの手法で抽出されたグラフは、連結でない複数の二部グラフを含んでいたり、共通のファンを持たないセンターを含んでいる可能性があり、これは本来の定義と異なったグラフを抽出する可能性がある。本論文では、2つのセンターは必ず N 以上のファンにより結ばれている二部グラフを抽出するアルゴリズムを提案し、従来の手法が持っていた問題を解決する。我々はクローラを用いて集めたウェブページを使って実験を行い、その結果を従来の手法と比較する。

Several methods to find Web communities by extracting dense bipartite graph structure from the Web graph are proposed. A dense bipartite graph is a graph which has two groups of vertices, fans and centers, and of which all fan links several centers and all center is linked by several fans. The DBG and the PlusDBG are an example of algorithms to extract such bipartite graphs. Brief of existing algorithms is first to expand a bipartite graph and second to extract a DBG from it. However, this process possibly causes extracting two unconnected bipartite graphs as one Web community or extracting a DBG of which centers are unrelated, and such graphs possibly does not match to the original idea of Web community. In this paper, we propose a DBG structure of which two centers are linked by more than one centers as Web community topology. Then, we conduct an experiment and compare the result with existing methods.

[▼] 東京大学大学院新領域創成科学研究科修士課程
eddieh@logos.ic.i.u-tokyo.ac.jp

[◆] 東京大学大学院情報理工学系研究科
tau@logos.ic.i.u-tokyo.ac.jp

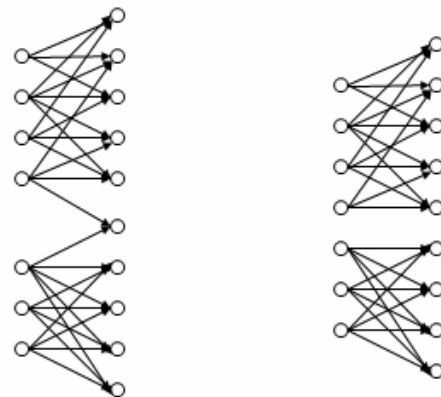
[▲] 東京大学大学院新領域創成科学研究科
chik@logos.k.u-tokyo.ac.jp

1. はじめに

ウェブコミュニティとは、トピックを共有するウェブページ集合のことである。ウェブコミュニティを抽出することは、ウェブページクラスタリングに有効に利用できることと、ウェブ上で新しいトピックを見つげられることが期待され、活発に研究が行なわれている。そのウェブコミュニティを抽出する方法は、大きく二つに分けられる。その一つは一般的なドキュメントクラスタリング手法を使って分類する方法である。もう一つの方法はウェブページ間のリンク構造を用いた手法である。本稿では主にリンク構造を用いたウェブコミュニティ抽出手法について述べる。

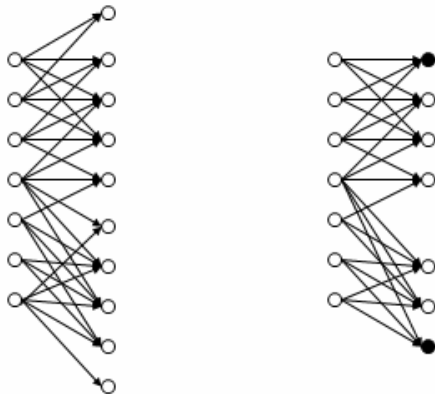
ウェブページの中には、多数のページからリンクされているページが存在し、このようなページを共通参照しているページも存在する。[1][2]では、このように連結されたページ群を見つけると、共通の主題を持っていることがあると主張し、そのページ群をウェブコミュニティと称した。それは、多数のページからリンクされているページは興味をもたらずものとされていて、またそのページを他のページが結んでいるため互いに共通した内容を含んでいると考えられるからである。それで、ウェブからこのような構造をしたページ群を見つけることにより、ウェブコミュニティを発見するためのアルゴリズムが提案されている。

ウェブコミュニティ抽出アルゴリズムとして、P. K. Reddyらの[3][4]と、斉田らの[5][6]が挙げられる。これらの研究では、「多数の共通リンクを含むページ」をファン、「多数のファンからリンクされているページ」をセンターと呼び、ファンとセンターで構成された密な二部グラフをウェブグラフから抽出するアルゴリズムを提案している。二部グラフとは、ノード群を2つに分割したとき、一つのノード群からもう一つのノード群へのエッジしか存在しないグラフを意味する。密な二部グラフ (Dense Bipartite Graph, DBG) は、ある条件を満たす二部グラフのことである。Reddyらはシードページを二部グラフに拡張し、その中からファンのアウトリンク数が p_{th} 以上、センターのインリンク数が q_{th} 以上となるような二部グラフを抽出し、抽出されたウェブページ群をウェブコミュニティとした。しかし、この手法では、密な二部グラフを抽出する際、ノードのインリンク数とアウトリンク数を数え条件に合わないノードを消去しているため、共通参照ノードの消去により連結でない二部グラフを抽出する可能性がある(図1)。



(a) 拡張された二部グラフ (b) 抽出される密な二部グラフ

図1 連結でない二部グラフが抽出される例 ($p_{th} = 3, q_{th} = 3$)
Fig.1 Example of unconnected DBG



(a) 拡張された二部グラフ (b) 抽出される密な二部グラフ

図2 共通のファンを持たないセンターを含んだ密な二部グラフが抽出される例。(b)の黒い二つのセンターは共通のファンを持たない ($p_{th} = 3, q_{th} = 3, Dis_{th} = 1.0$)

Fig. 2 Example of DBG of which two centers have no common fan

齊田らは共参照リンクを用いて距離量を定義し、二部グラフを抽出する際に一定距離以内にあるページのみを二部グラフに入れていき、その中から密な二部グラフを抽出している。この手法では距離量を変化させることより連結でない二部グラフが抽出されないようにすることが可能である。しかしこの手法は、共通のファンを持たないセンターを含んだ二部グラフを抽出する可能性がある(図2)。

二部グラフを用いたウェブコミュニティの抽象化は、本来興味の対象であるセンターを、興味を持つ主体であるファンが結んでいるという考えに基づいている [6]では、PlusDBGの距離量の閾値を小さくするとPlusDBGの精度が向上すると報告しているが、これもファンによってより強くセンターが結ばれることにより得られた結果であると考えられる。したがって、連結でない二部グラフとセンターが共通のファンを持たない二部グラフは、目的とするウェブコミュニティを反映していないと言える。逆にいうと、グラフを抽出する段階でこのようなグラフを除くことによってコミュニティの精度を上げられると考えられる。

以上の考えに基づき、本稿ではウェブコミュニティを「2つのセンターは必ず N 以上のファンを持つ二部グラフ」と定義し、そのようなウェブコミュニティを抽出するアルゴリズムを提案する。以降は2.章で提案するウェブコミュニティの定義と抽出アルゴリズムについて述べる。そして3.章で実験と結果について説明し、4.章で結論と今後の課題について述べる。

2. ウェブコミュニティ

2.1 定義

我々は二部グラフのうち、リンクを張っている側をファンと呼び、その要素を s_i 、集合を S とする。また、リンクの受け側をセンターと呼び、その要素を t_i 、集合を T とする。このとき、2つのセンターの連結可能な関係を次のように定義する。

[定義1] 連結可能(Connectable)

センター t_1 とセンター t_2 が連結可能であるとは、 t_1 と t_2 を同時に参照しているファンが存在することを意味する

そして、2つのセンターが連結可能であるとき、その二つのセンターをリンクしているファンを連結子と呼び、連結子の数を用いて連結度を定義する。

[定義2] 連結子(Connector)

センター t_1 とセンター t_2 が連結可能であるとき、 t_1 と t_2 を同時に参照しているファン s を連結子とする

[定義3] 連結度(Connectivity)

センター t_1 とセンター t_2 が N 個の連結子によって連結されているとき、連結度 $Connectivity(t1, t2)$ は N である

これらの関係を用いてすべてのセンターが連結可能な関係にあり、またファンがセンターの連結子であるような密な二部グラフDBGが存在すれば、我々はそのようなDBGをウェブコミュニティとする。

[定義4] ウェブコミュニティ

密な二部グラフ $DBG(S, T)$ において、 T のすべての要素が連結度 N 以上で連結可能であり、 S がその連結子の集合であるとき、 $DBG(S, T)$ をウェブコミュニティとする

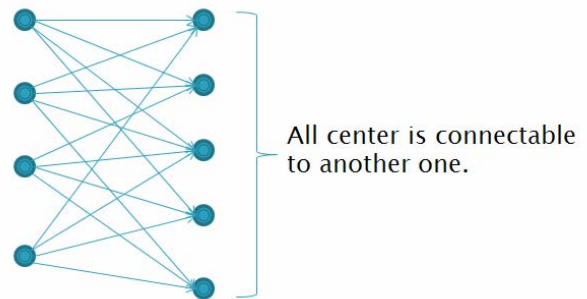


図3 提案するウェブコミュニティ

Fig. 3 Proposed Web Community Structure

連結度2の二部グラフ構造の例を図3に表す。図3で、すべてのセンターは2個以上のファンによってリンクされていることがわかる。本稿ではこのような二部グラフ構造を抽出し、それをウェブコミュニティとする。

2.2 抽出アルゴリズム

我々は定義4のDBGを抽出するために、最初にシードページとして一つのセンターを選び、センター集合とする。そして、センター集合のすべてのメンバーと連結可能なセンターを一つ探し、DBGを拡張することを繰り返す。その手順は以下とおりである。

- (1) シードノード t を選び、 $T = \{t\}$, $S = \emptyset$ とする
- (2) T のノードにリンクをしているノードの集合を S' とする
- (3) S がリンクしているすべてのノードから T のメンバーを除いたものを T' とする
- (4) $t' \in T'$ において、 t' が T のすべてのメンバーと連結可能であるかを判定
 - (a) 連結可能なら $T = \{t'\} \cup T$, $S = S \cup \{\text{連結子}\}$ とし、(2)へ
 - (b) 連結可能でなければ他の t' を選び4.へ
- (5) $|S| > p, |T| > q$ ならウェブコミュニティとし、ウェブグラフから削除する
- (6) ウェブグラフにノードが残っていれば(1)へ

本稿ではDBGが抽出されたときそのDBGを全体のグラフから取り除くことを行っているが、これは便宜のためであり

深い意味を持つことはない。また、Step. (4) で t' の選び方によって抽出されるウェブコミュニティが変わる可能性がある。これに関する議論は今後の課題とする。

3. ウェブコミュニティの解析

3.1 データセットと前処理

データセットとしては、我々のクローラ¹を用いて集めたページを利用した。このウェブページは、日本語のページを保有する約 15 万のホストから、日本語のページのみを深さ 2 までたどって集めたものである。日本語のページは、chardet の python モジュールを使い、EUC-JP, ShiftJIS, ISO-2022-JP (JIS) の文字を含むページとした。これによって集まったページは約 235 万、その中に含まれているリンクの数は約 6129 万である。

このデータセットに前処理を行い、実験に用いるデータセットを作成した。前処理は、リンクの中でデータセットに含まれるウェブページを向いていないリンクの削除、複製ページの削除、有名あるいは無名なページの削除の3段階で構成されている。まず我々はデータセットの中を向いていないリンクを削除した。次に、90%以上のリンクが同じページを複製ページとし、一方を削除し、リンク先が削除されたリンクに対してはリンク先を複製ページに変更した。最後に、インリンクの数が 50 以上、アウトリンクの数が 3 以下のページを削除した。

前処理を行った結果、データセットのウェブページ数は 145 万、リンク数は 509 万となった、そして、ホストの数は約 37.4 万であった。

3.2 ウェブコミュニティ抽出の結果

3.2.1 ウェブコミュニティのサイズ

表 1 ウェブコミュニティ数とコミュニティサイズの合計
Table.1 Number of Web communities and community size

抽出手法	ウェブコミュニティ数	サイズ合計	平均
PlusDBG(1.2)	7,527	923,100	122
PlusDBG(1.0)	8,077	922,053	114
PlusDBG(0.8)	22,902	865,945	37.8
提案手法(N=2)	50,065	648,626	12.9
提案手法(N=3)	45,027	568,932	12.6
提案手法(N=4)	37,234	501,329	13.5

表 1 に抽出されたウェブコミュニティの数とウェブコミュニティとして含まれるウェブページの数を示す。表 1 より、提案手法では連結度の閾値を上げることにより抽出されるウェブコミュニティの数は減るが、抽出されたコミュニティの平均サイズはほとんど変化していないことがわかる。これは、提案手法で抽出されたウェブコミュニティは連結度の閾値が上がればウェブコミュニティでないと判定されるものの、コミュニティが分離されることは少ないからであると考えられる。一方で PlusDBG では距離量の閾値を小さくすればするほど、たくさんのウェブコミュニティが抽出され、またコミュニティの平均サイズは小さくなった。これは、PlusDBG では距離量を変化させることにより、一つのウェブコミュニティが複数に分けられていると考えられる。この点は提案手法と PlusDBG の大きな違いである。

¹ Shim-Crawler, <http://www.logos.ic.i.u-tokyo.ac.jp/crawler/>

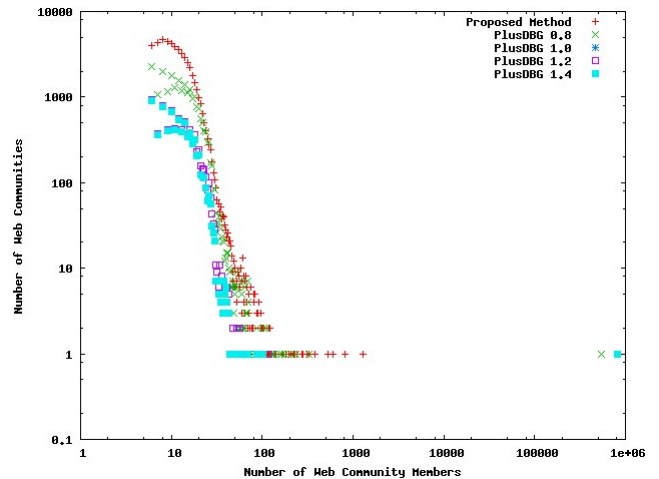


図 4 ウェブコミュニティサイズの分布 (x 軸: ウェブコミュニティメンバーの数, y 軸: コミュニティの数)

Fig.4 Distribution of Web community size

そして、図 4 にウェブコミュニティのサイズの分布を示す。ここでは比較のために、PlusDBG(0.8)を併記した。図 4 では、連結度を変化させてもコミュニティサイズの分布はあまり変化しないことがわかる。また、PlusDBG と比較すると、我々の提案手法では抽出されなかった大きな二部グラフ(約 60 万ノード)が PlusDBG では抽出されていて、比較的の小さい(10~20)サイズのウェブコミュニティが我々の手法ではたくさん抽出されていることである。これは、図 2 に示されたような、共参照されていないセンターが PlusDBG ではコミュニティとして抽出されており、我々の手法では別々の小さなウェブコミュニティとして抽出されているからであると考えられる。

以上のことより、本稿での提案手法は既存の手法よりコンパクトなウェブコミュニティを抽出していると言える。

3.2.2 ODP²との比較

抽出されたウェブコミュニティの精度を評価するために、抽出されたウェブコミュニティと ODP で作成されたウェブディレクトリの比較を行った。ここでは精度の定義として、

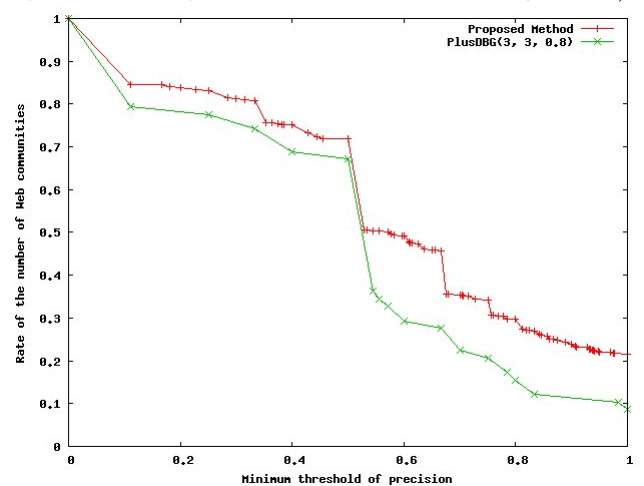


図 5 ウェブコミュニティ精度別ウェブコミュニティの割合
Fig.5 Ratio of Web community to precision threshold

² Open Directory Project, <http://dmoz.org/>

[6]で定義されたものを用いた。その定義とは、一つのコミュニティに属するページが ODP の同じカテゴリに入っている割合を示すものである。ここではツリー状になっている ODP の深さ 3 までを比較し、その結果を図 5 に示した。グラフの横軸は再現率の閾値で、縦軸にはその閾値以上の再現率を持つウェブコミュニティの割合である。

この結果からわかるように、提案手法が PlusDBG より良い精度を見せている。これは、提案手法で抽出されたウェブコミュニティが、PlusDBG より少ない ODP のウェブページを含んでおり、また一つのウェブコミュニティに含まれるウェブページは、ODP による分類に基づき、提案手法が PlusDBG より似ているジャンルのページを同じウェブコミュニティとして抽出しているからと考えられる。また、どのグラフにおいても連結度の変化による再現率の大きな変化は見られなかった。ウェブコミュニティのサイズのことまで考えると、個々のウェブコミュニティの性質に対する連結度の寄与度は大きくないと言える。

3.2.3 TF-IDF 空間上での分布

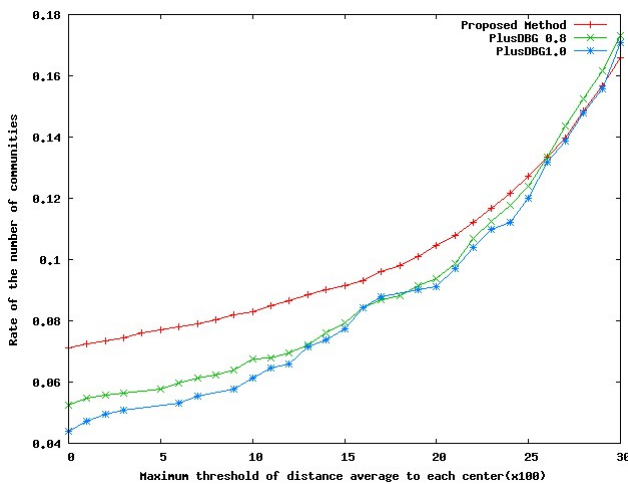


図 6 TF-IDF 空間上で重心までの平均距離の分布

Fig.6 Distribution of average distance in a community

図 6 に各ウェブコミュニティにおける重心とメンバーとの距離の平均値を表す。このグラフでは横軸が距離の平均の閾値を表し、縦軸はその閾値より小さい距離の平均を持つウェブコミュニティの割合を表す。便宜上、距離の平均が 0.30

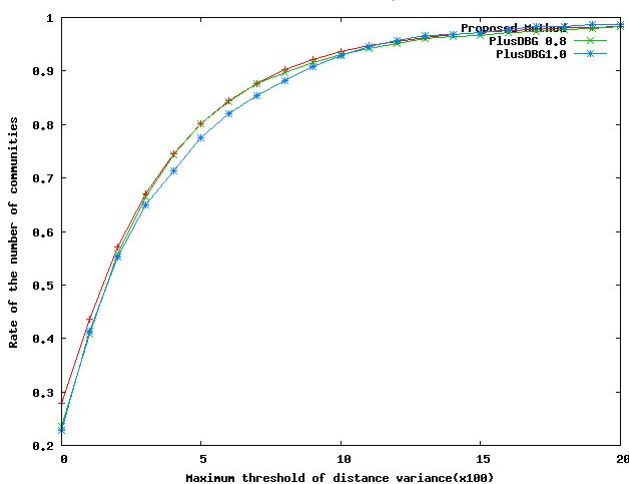


図 7 TF-IDF 空間上で重心までの距離の分散の分布

Fig.7 Distribution of distance variance in a community

以上となるものは省略した。この結果より、若干ではあるが提案手法が既存の手法より重心に近いウェブページ群をウェブコミュニティとして抽出する傾向があることがわかる。

そして図 7 に、各ウェブコミュニティにおける重心とメンバーとの距離の分散を示す。このグラフでは横軸が距離の分散の閾値を表し、縦軸はその閾値より小さい距離の分散を持つウェブコミュニティの割合を表す。この実験からもわかるように、提案手法は既存手法より一様に重心に近いウェブページ群をコミュニティとして抽出することがわかる。

4. まとめ

本論文では、ウェブコミュニティの精度を向上させることを目的とし、従来の二部グラフ抽出手法の改良を行った。本論文では、センター同士は必ず複数のファンによって結ばれるような密な二部グラフをウェブコミュニティとして提案し、その抽出アルゴリズムを述べた。そして提案手法と従来の PlusDBG を用いてウェブコミュニティを抽出し、その性質を比較した。その結果、本論文で提案したウェブコミュニティは PlusDBG よりコンパクトで、ODP との比較をしたとき PlusDBG より良い精度を見せることができた。また、TF-IDF 空間上におけるウェブコミュニティメンバーの分布を調べることで、抽出されたウェブコミュニティが単語空間で意味を持つことを示した。

【文献】

- [1] J. Kleinberg.: Authoritative sources in a hyperlinked environment, *ACM SIAM*, 1998
- [2] R. Kumar, P.Raghavan, S. Rajagopalan, A. Tomkins. : Trawling the Web for Emerging Cyber-Communities. *Computer. Networks*, 1999
- [3] 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.
- [4] P.Krishna Reddy, Masaru Kitsuregawa.: Building a community hierarchy for the Web based on bipartite graphs. *DEWS*, 2002.
- [5] Naoyuki Saida, Akira Umezawa, Hayato Yamana. PlusDBG: Web Community Extraction Scheme Improving Both Precision and Pseudo-Recall, *Asia-Pacific Web Conference*, 2005.
- [6] 齊田直幸, 山名早人. リンク構造解析による不要Web コミュニティの判別, *DEWS*, 2006.

沈 垣甫 Wonbo SHIM

東京大学大学院新領域創成科学研究科修士課程修了。現在韓国 LG Electronics の研究員。

田浦 健次郎 Kenjiro TAURA

東京大学大学院情報理工学系研究科助教授。1997 東京大学大学院にて理学博士号を取得。主に並列処理に関する研究に従事。

近山 隆 Takashi CHIKAYAMA

東京大学大学院新領域創成科学研究科教授。1982 東京大学大学院工学系研究科博士課程修了。同年から第五世代コンピュータプロジェクトに参加の後、1995 より東京大学。プログラミング言語、開発環境、機械学習に関する研究に従事。