ウェブコミュニティ抽出アルゴリズムの改良

沈　垣甫† 　田浦　健次朗†† 　近山　隆†

† 東京大学大学院新領域創成科学研究科
†† 東京大学大学院情報理工学系研究科
E-mail: †{eddieh,tau,chik}@logos.ic.i.u-tokyo.ac.jp

あらまし　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と、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。
キーワード　ウェブコミュニティ、リンク構造解析、データマイニング、知識発見

Improvement of Web Community Extraction Algorithm

Shim WONBO†, Kenjiro TAURA††, and Takashi CHIKAYAMA†

† Dep. of Frontier Science, University of Tokyo
†† Dep. of Information Science and Technology, University of Tokyo
E-mail: †{eddieh,tau,chik}@logos.ic.i.u-tokyo.ac.jp

Abstract　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 commutnity topology. Then, we conduct an experiment and compare the result with existing methods.
Key words　Web community, Link analysis, Data mining, Knowledge discovery

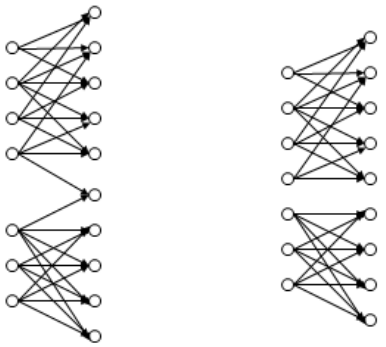
1. 初　め　に

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け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すものとされ

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 [6] [7] と、斉田らの [8] [9] 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 2. 章でも述べるように、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）。

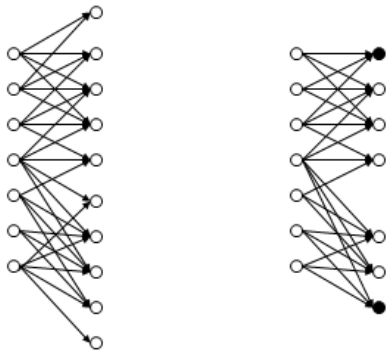


(a) 拡張された二部グラフ　(b) 抽出される密な二部グラフ

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

斉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。[9] 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、



(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。（$p_{th} = 3$, $q_{th} = 3$, $Dis_{th} = 1.0$）

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、このようなグラフが抽出されないようにする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。

以上のような考えに基づき、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. 章で関連研究について述べ、3. 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 4. 章で実験と結果について説明し、5. 章で結論と今後の課題について述べる。

2. 関 連 研 究

ウェブコミュニティに関する研究は、主にウェブグラフ [3] から特徴的なリンク構造を抽出する手法を提案している。特徴的なリンク構造としては、[4] [5] で提案しているような「外側へのリンクの数より内側へのリンクの数が多い部分グラフ」と [2] [6] [7] [8] [9] で提案されているような「密な二部グラフ」の 2 つを挙げることができる。以下、これらの研究について説明する。

Flake らは [4] で、ウェブグラフ上でエッジをまたがった閉境界を設けたとき、境界の内側にあるノード間のエッジの数が境界上にあるエッジの数より多いとき、それをウェブコミュニティとしている。そして、そのようなウェブコミュニティを見つけるために、ウェブグラフに仮想の始端と終端を設け、「$s-t$ 最大フロー・最小カット」アルゴリズムを用いてこのグラフを二等分する、という手法を提案している。

しかし、Flake らの手法には抽出するウェブコミュニティのグラフ構造にあいまいさが残っている。すなわち、ウェブコミュニティのメンバーとなるべきノードが一意に定まらない。Ino らは [5] で、Flake らの手法が持つあいまいさを指摘し、それを改善するためのコミュニティ抽出手法を提案している。

一方で Kumar らは [2] で、ウェブコミュニティを「完全な二

部グラフを内包する密な二部グラフ」と定義している。[2] では、ウェブコミュニティを抽出するためにインリンクとアウトリンクの数を元にファンとセンターの候補を選び、それらが形成する二部グラフを枝狩りすることでウェブコミュニティを抽出している。

Reddy ら [6] は Kumar らの定義を緩め、密な二部グラフをウェブコミュニティとした。Reddy らの密な二部グラフは、ファンはセンターに対して p_{th} 以上のリンクを持ち、センターはファンから q_{th} 以上のリンクを持つ二部グラフ（以下、DBG と称する）を意味する。DBG を抽出するアルゴリズムは、シードページからリンクをたどり DBG の候補を抽出し、そこから DBG を抽出することである。具体的には、まずシードページ s を選んで $S = \{s\}$、$T = \emptyset$ とし、S からリンクされているページを T に加えて再び T をリンクしているページを S に加えることを繰り返し行い、密な二部グラフの候補とする。そして $s \in S$ に対し、T に対する s のリンクの数が一定値以下なら S から抜き、$t \in T$ に対しても S からのリンクの数が一定値以下なら T から抜く、ということを繰り返す。Reddy らは DBG によるウェブコミュニティの抽象化をウェブコミュニティ同士に適用することにより、関連するウェブコミュニティを抽出している [7]。

斉田らは [8] で、DBG を抽出するアルゴリズムとして PlusDBG を提案している。PlusDBG はウェブページ間の共参照関係を用いて距離量を定義する。この距離量を用いて、PlusDBG はシードページから一定の距離量以内にあるページを繰り返したどり、密な二部グラフの候補とする。そして、候補となったグラフから Reddy らの手法と同様な密な二部グラフを抽出する。この手法を用いて、斉田らは [9] で距離量の閾値が違うウェブコミュニティを抽出して比較することより、不要なウェブコミュニティを判別した。

密な二部グラフによるウェブコミュニティの抽象化は、センターがファンによってつながることにより同じトピックを持つ、という考えに基づいている。しかし前章にも述べたように、Reddy らや斉田らによって提案された手法はセンターがファンによって結ばれないような二部グラフを抽出する可能性がある。これは本来目的としたウェブコミュニティの構造とは異なっており、抽出アルゴリズムに改善が必要であると考えられる。

3. ウェブコミュニティ

本研究も [6] [8] と同様に、密な二部グラフによるウェブコミュニティの抽象化を行う。ただし条件として、どの二つのセンターも複数のファンによってリンクされていることを与える。本章では、我々が提案するウェブコミュニティに関する定義と、その抽出手法について述べる。

3.1 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ぶ。

［**定義** 2］（連結子 (Connecter)） 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。

また、連結可能な 2 つのセンターの連結度を連結子の数を用いて次のように定義する。

［**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 $Connectivity(t_1, t_2)$ は N である。

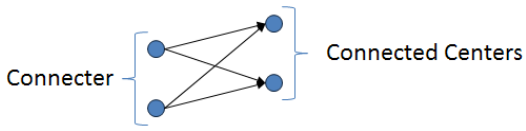


図 3　連結子と連結可能なセンター

連結子と連結可能なセンターの関係の例を図 3 に表す。

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。

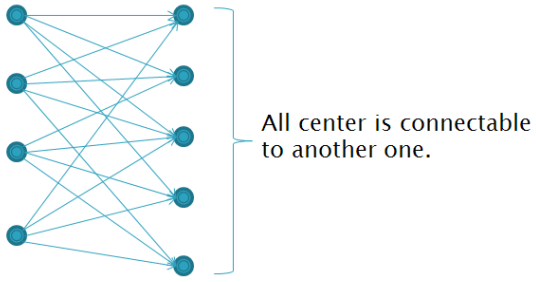


図 4　抽出される二部グラフ構造

連結度 2 の二部グラフ構造の例を図 4 に表す。図 4 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。

3.2 抽出アルゴリズム

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 $T = \{t'\} \cup T$、$S = S \cup \{$ 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。

4. ウェブコミュニティの解析

4.1 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(注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(注2)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、データセットをリンクのみで表現した。リンクの表現には 32 ビットのハッシュ値を用いた。そして、このリンクのうち、データセットの外側へのリンクを削除した。これによってリンクの数は約 1586 万になった。

次に、データセットから複製ページを削除した。複製ページは次の条件を満たす 2 つのページとした。

（1） $children(a) > 10, children(b) > 10$

（2） $0.9 \mid children(b) \mid \leqq \mid children(a) \mid \leqq 1.1 \mid children(b) \mid$, $0.9 \mid children(a) \mid \leqq \mid children(b) \mid \leqq 1.1 \mid children(a) \mid$

（3） $\frac{|children(a) \cap children(b)|}{|children(a) \cup children(b)|} \geqq 0.9$

ここで、$children(a)$ は a がリンクしているページの集合である。複製と判断された場合は a と b のどれかを削除し、削除したページを参照しているページのリンクを複製ページに入れ替えた。

最後に、アウトリンク数とインリンク数に閾値を設け、閾値に達しないページを削除した。まず、アウトリンク数が 3 以下のページを削除し、インリンク数が 50 以上のページを削除した。

複製ページの削除と枝狩りを行った結果、残ったウェブページの数は 145 万、リンクの数は 509 万となった。そして、ホストの数は約 37.4 万であった。

4.2 ウェブコミュニティ抽出の結果

本節では、提案手法により抽出されたウェブコミュニティついて述べる。比較のために従来の $PlusDBG(p_{th} = 3, q_{th} = 3, Dist = 0.8)$ を用いて抽出したウェブコミュニティの結果を併記する。評価として、まず抽出されたウェブコミュニティのサイズについて述べ、次に ODP への再現率を示す。そして、一般的なドキュメントクラスタリングによく用いられる *TF-IDF* 空間上でのウェブコミュニティの分布について説明する。

4.2.1 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ウェブコミュニティサイズの合計

抽出手法	ウェブコミュニティ数	サイズの合計	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	568,939	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。

そして、図 5 にウェブコミュニティのサイズの分布を示す。ここでは比較とために、PlusDBG(0.8) を併記した。図 5 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

(注1)：Shim Crawler, thtp://www.logos.ic.i.u-tokyo.ac.jp/crawler/
(注2)：http://chardet.feedparser.org/

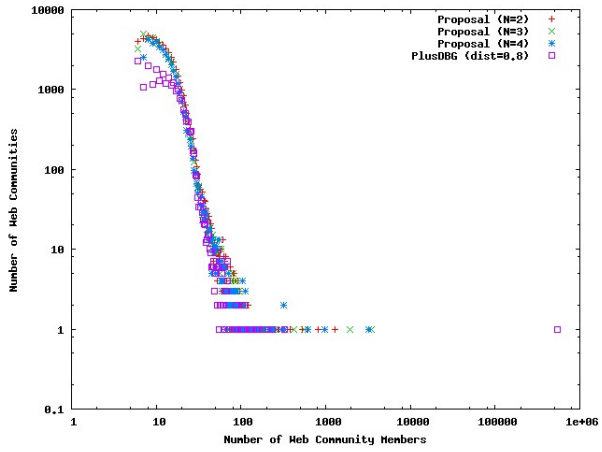


図 5　ウェブコミュニティのサイズの分布

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、比較的に小さい (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。

4.2.2　ODP(注3) 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 Open Directory Project で作成されたウェブディレクトリの比較を行った。ODP は、人手によって構成したウェブディレクトリをサービスするものである。

ウェブコミュニティの精度評価には、[8] で提案された手法を用いる。この手法では、まず 2 つのウェブページ p、q に対して、次のような score を与える。

$$score(p,q) = 1, \text{ 同ディレクトリに存在} \quad (1)$$

$$score(p,q) = 0, \text{ 違うディレクトリに存在} \quad (2)$$

ここでは、ODP のツリー構造中深さ M までのディレクトリが同じ場合、同ディレクトリに存在するとする。たとえば $M = 2$ のとき、ディレクトリ「*Top/Arts/Movie/Film Archives*」に存在するウェブページと、「*Top/Arts/Movie/Filmmaking*」に存在するウェブページは同じディレクトリの「*Top/Arts/Movie*」にあるものとされ *score* は 1 となる。ただし、ディレクトリ中の「*Top/World*」と「*Top/Regional*」に関しては、深さ $(M+1)$ までのディレクトリが同じ場合 *score* を 1 とする。

このような *score* を用いて、次のように精度を定義する。

（1）　ウェブコミュニティ C の中に、ODP に存在するウェブページを OC とする

（2）　$|\,OC\,| < 3$ の場合は評価を行わない。

（3）　OC 中のページ r に対して、r のスコアを次のように定義する。

$$Pscore(r) = \frac{\sum_{x \in (OC-r)} score(r,x)}{|\,OC - r\,|} \quad (3)$$

（4）　これを用いて、C の精度 $P(C)$ を次のように定義する。

$$P(C) = \sqrt{\frac{\sum_{r \in OC} Pscore(r)}{|\,OC\,|}} \quad (4)$$

$Pscore$ はコミュニティの中で一つのウェブページと共通するディレクトリに含まれるウェブページの割合を表す。そして、P は $Pscore$ の平均値と考えることができる。すなわち、P は一つのウェブコミュニティの平均再現率と言うことができる。

本研究では、以上のような精度の定義を用いて、抽出されたウェブコミュニティと ODP ツリーの深さ $M = 2$ と $M = 3$ との比較を行った。その結果を以下で述べる。

まず、データセット、PlusDBG によってウェブコミュニティのメンバーとして抽出されたウェブページ全体、提案手法によってウェブコミュニティのメンバーとして抽出されたウェブページ全体それぞれと、ODP に同時に含まれたウェブページの数を表 2 に示す（参考までに、ODP には約 438 万のウェブページが登録されている）。そして、ODP とウェブページを共有しているウェブコミュニティのうち式 4 の条件を満たし、評価の対象としたウェブコミュニティの数と、表 2 に表すウェブページが含まれる ODP のディレクトリの数を表 3 と表 4 に表す。

表 2 から、提案手法は PlusDBG がウェブコミュニティとして抽出した ODP ページの約 55% をウェブコミュニティとして抽出していることがわかる。また、連結度の閾値が上がるに

表 2　ODP と共通しているウェブページの数

ウェブページ集合	ウェブページ数
データセット全体	29153
PlusDBG(0.8)	23,287
提案手法 (N=2)	12,406
提案手法 (N=3)	10,202
提案手法 (N=4)	8,560

表 3　表 2 を持つウェブコミュニティ数と ODP のディレクトリ数 $(M = 2)$

抽出手法	ウェブコミュニティ数	ODP ディレクトリ数
PlusDBG(0.8)	459	426
提案手法 (N=2)	4811	337
提案手法 (N=3)	4028	307
提案手法 (N=4)	3282	282

表 4　表 2 を持つウェブコミュニティ数と ODP のディレクトリ数 $(M = 3)$

抽出手法	ウェブコミュニティ数	ODP ディレクトリ数
PlusDBG(0.8)	459	1186
提案手法 (N=2)	4808	925
提案手法 (N=3)	4027	820
提案手法 (N=4)	3281	752

(注3)：Open Directory Project, http://dmoz.org/

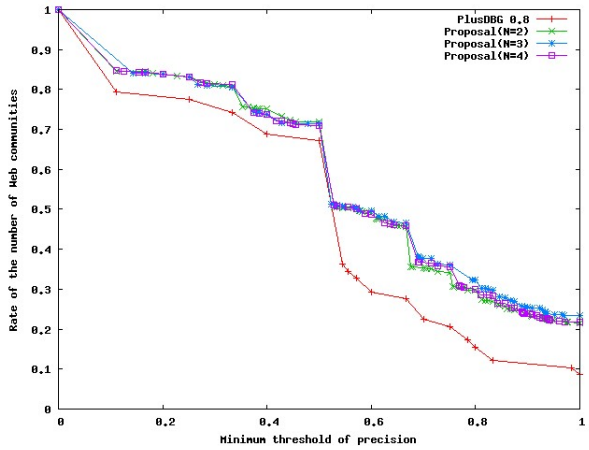


図 6　ウェブコミュニティの精度ごとのウェブコミュニティの割合 ($M = 2$)

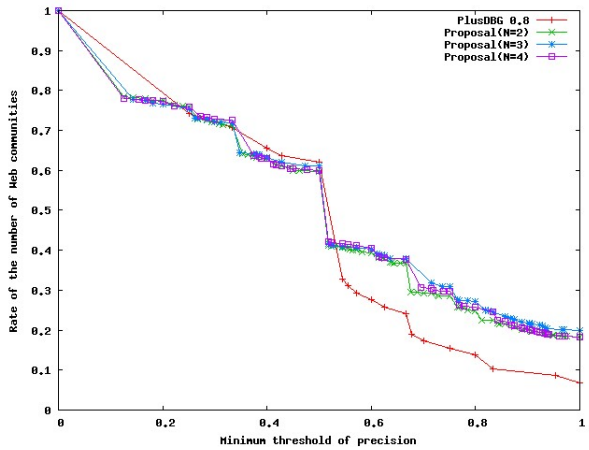


図 7　ウェブコミュニティの精度ごとのウェブコミュニティの割合 ($M = 3$)

つれ、ODP と共有するウェブページの数が減っている。表 1 の結果を同時に考慮に入れると、提案手法が PlusDBG よりカバー率が小さいことに起因すると考えられる。

そして表 3 と表 4 より、提案手法では PlusDBG よりたくさんのコミュニティが ODP のページをメンバーとして含んでいることがわかる。図 5 を考慮すると、これは提案手法によって抽出されたウェブコミュニティがより小さいサイズで、それぞれのコミュニティの ODP のページが部分集合として入っているからであると推測できる。

図 6 と図 7 に、ODP に対する抽出されたウェブコミュニティの再現率を表す。図 6 は ODP ツリーの深さ M を 2、図 7 は 3 としたときの再現率である。それぞれの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。

図 6 と図 7 でわかるように、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。

4.2.3　TF-IDF 空間上での分布

次に、一般的なドキュメントクラスタリングによく用いられる *TF-IDF* 空間におけるウェブコミュニティの性質について述べる。TF-IDF はドキュメントをベクトルとして表す手法で、その表現方法にはさまざまなものがあるが、本研究では以下の定義を用いた。

$$TF - IDF(term) = \|TF\| \times log(IDF) \tag{5}$$

この実験のために、まず単語ベクトルですべてのページを表した。そのときにパーサとして MeCab(注4) を用いた。次に、各単語の IDF を調べ、全体のウェブページ中 90%以上、あるいは 0.1%以下に出現する単語を削除し、上記の TF-IDF で各ウェブページを表した。次に、各ウェブコミュニティの重心を求め、TF-IDF 空間上でウェブコミュニティのメンバーから重心までの距離の平均、そしてその距離の分散を計算した。

以上のことを用いて提案手法 (N=2) によって抽出されたウェブコミュニティの分布と PlusDBG(0.8)、PlusDBG(1.0) によって抽出されたウェブコミュニティを TF-IDF 空間上で比較した結果を図 8 と図 9 に表す。

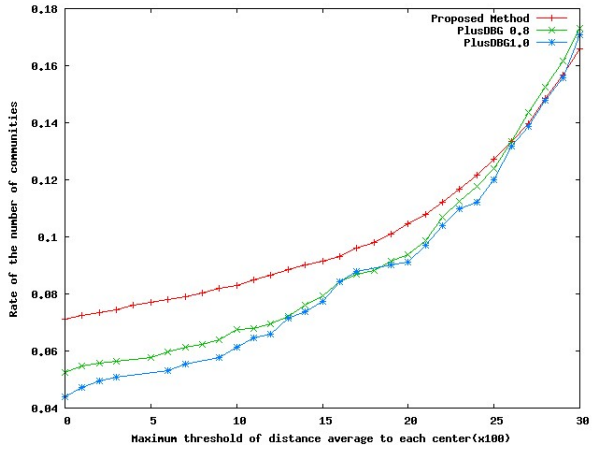


図 8　TF-IDF 空間上で重心までの平均距離の分布

まず図 8 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、距離の平均がが 0.30 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 9 に、各ウェブコミュニティにおける重心とメン

(注4)：MeCab, http://mecab.sourceforge.jp/

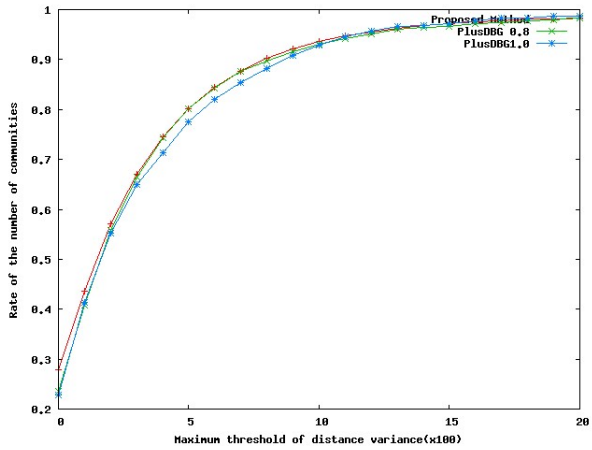


図 9　TF-IDF 空間上で重心までの距離の分散の分布

バーとの距離の分散を表す。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。

5. 終わりに

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より、抽出されたウェブコミュニティが単語空間で以前の研究より意味も持つことを示した。

今後の課題としては、3. で述べたウェブコミュニティアルゴリズムの途中、新しい t を追加する順序を考えることが挙げられる。それは、抽出される DBG の構造は t に依存するため、一意的に定まらないからである。

そして、抽出されたウェブコミュニティを単語空間上でのクラスタリングに活用することが考えられる。単語を用いた一般的なクラスタリングでは、単語空間上に固まっているものを一つのクラスタとみなすことが多いが、ウェブページの場合その数から来る計算量の限界から単語ベクトルをそのまま扱えない問題点がある。しかし本研究で示したようにウェブコミュニティは単語空間上でも意味を持つため、これを利用することで計算量の問題の解決に近づくことができると考えられる。

本論文で提案した手法の一つの応用例としては、本稿で提案した手法を用いて抽出されたウェブコミュニティ中、高い精度を見せるウェブコミュニティに対して、ODP など人手によって構成されたウェブディレクトリに存在しないウェブページをそのウェブディレクトリの一つとして提案することが考えられる。それは高い精度のウェブコミュニティのメンバーは、同じウェブディレクトリに入る可能性が他のディレクトリより高いと思われるからである。

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] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. Graph structure in the web, *Proceedings of 9th International WWW Conference*, 2000.

[4] G. W. Flake, S. Lawrence, C. Lee Giles. Efficient Identification of Web Communities. *ACM SIGKDD*, 2000.

[5] H. Ino, M. Kudo, A. Nakamura. Partitioning of Web Graphs by Community Topology. *Proceedings of The International World Wide Web Conference*, 2005.

[6] 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.

[7] P.Krishna Reddy, Masaru Kitsuregawa. Building a community hierarchy for the Web based on bipartite graphs. *DEWS*, 2002.

[8] Naoyuki Saida, Akira Umezawa, Hayato Yamana. PlusDBG: Web Community Extraction Scheme Improving Both Precision and Pseudo-Recall, *Asia-Pacific Web Conference*, 2005.

[9] 斉田 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