

Web ページ間の関連性の伝播を用いた

Web コミュニティ抽出手法の評価

飯村 卓也[†] 平手 勇宇^{††} 山名 早人^{‡,‡‡}

[†] 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

^{††} 早稲田大学メディアネットワークセンター 〒169-8050 東京都新宿区戸塚町 1-104

[‡] 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

^{‡‡} 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: {iimura, hirate, yamana}@yama.info.waseda.ac.jp

あらまし ユーザが求める情報を効率よく獲得するために数多くの Web コミュニティ抽出手法が提案されている。しかし、既存研究ではある Web ページを Web コミュニティに含めるかどうかの条件を厳しくすることにより、本来 Web コミュニティに含まれるべき Web ページが含まれなくなることがあり、再現率が低下していると考えられる。そこで我々はこれまでに、適合率を低下させることなく、再現率を向上させる Web コミュニティ抽出手法を提案した。本稿では、我々の手法と既存手法との再現率における比較実験を行った。比較実験の結果、我々が提案してきた手法は既存手法と同等の適合率を保ちつつ、再現率を向上させることができることが確認できた。

キーワード リンク解析, Web コミュニティ

Evaluation for Web Community Extraction Method with Web Pages' Relevance Forwarding

Takuya IIMURA[†] Yu HIRATE^{††} and Hayato YAMANA^{‡,‡‡}

[†] Graduate School of Fundamental Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan

^{††} Media Network Center, Waseda University 1-104 Totsuka-cho, Shinjuku-ku, Tokyo, 169-8050, Japan

[‡] Faculty of Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan

^{‡‡} National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

E-mail: {iimura, hirate, yamana}@yama.info.waseda.ac.jp

Abstract To find information from a large collection of Web-pages, several methods for extracting Web communities have been proposed. In the past studies, they succeed in improving precision score by making a rule whether or not to include a certain Web page into a Web community strictly. However, recall score might worsen because the Web page that should be included in the Web community is not included. Therefore, we have proposed the Web community extraction method that can improve recall score without decreasing precision score. The method adds Web pages that have many links from/to the Web pages in a same Web community. In this paper, we conduct comparative experiments and evaluate recall score. Experiments shows that our proposed method can improve recall score keeping precision score equal with existing methods.

Keyword Link Structure Analysis, Web Community

1. はじめに

近年、ユーザが求める情報を効率よく獲得するために、Webコミュニティを発見する研究が活発に行われている。Webコミュニティとは意味的に関連のあるWebページの集合のことであり、任意のWebページをシードページとして抽出されたWebコミュニティを利用法として以下のような利用法が挙げられる。

- (1) ユーザが求める情報への効率的なナビゲーションや新たな知識発見[1].
- (2) Web上での話題の抽出や、ある話題に関するWeb

上での発展過程の分析[2].

- (3) スパムサイトの発見[3].

これまでに提案されている Web コミュニティ抽出手法には Web ページのドキュメントの類似度を計算する手法と Web ページ間のリンク構造を利用した手法がある。本稿ではリンク構造を利用した Web コミュニティ抽出手法について述べる。

リンク構造を利用した手法として、Web ページ間のリンク構造を有向グラフとみなしたグラフ（以後、Web グラフ）から密な 2 部グラフ（DBG: Dense Bipartite

Graph)を形成する部分グラフを抽出する手法が挙げられる。しかし、単に DBG を Web グラフから抽出するだけではトピックを共有しない Web ページも Web コミュニティに含まれてしまうことがある。そこで、DBG を用いた手法の既存研究[8][9]では、ある Web ページを Web コミュニティに含めるかどうかを決定する際の条件（以後、抽出条件）を厳しくすることにより適合率の向上を図っている。

齊田らは、トピックを共有しない Web ページを判別する手法として PlusDBG を提案している[8]。PlusDBG はある Web ページと Web コミュニティとの関連の強さを示す尺度として「距離量」を定義し、Web コミュニティとの距離量が一定以上 Web ページを排除することにより適合率の向上を図った。

沈らは、ある Web ページと Web コミュニティとの関連の強さを示す尺度として「連結度」を定義している[9]。連結度とは、Web コミュニティ中の任意の 2 つの Center(DBG におけるリンク先ノード)に対して定義される値であり、その 2 つの Center に同時にリンクしている Fun(DBG におけるリンク元ノード)の数を指す。ここで、同時にリンクしている Fun を連結子という。沈らは、Web コミュニティ中の任意の 2 つの Center が N 以上の連結度を持つような 2 部グラフを抽出することにより、齊田らの手法よりも高い適合率を有する Web コミュニティを抽出することに成功している。

沈らは適合率に関する評価実験としてディレクトリサービスを用いた評価実験と TF-IDF を用いた評価実験を行い、齊田らの手法によって抽出される Web コミュニティよりも高い適合率を有する Web コミュニティを抽出できることを示した。しかし、沈らの研究では適合率は向上しているが再現率は低下していると言える。

これらの研究に対し、我々は、これでまでに、Web コミュニティの適合率を下げることなく再現率を向上させることを目的として、新たな Web コミュニティ抽出手法の提案を行った[10]。本稿では、我々が提案してきた手法と沈らの手法との再現率における比較実験を行う。

以後、2 節では関連研究について述べる。3 節では提案する Web コミュニティ抽出手法について説明する。4 節では我々が提案してきた手法と沈らの手法の比較実験と評価を行う。5 節でまとめを行う。

2. 関連研究

リンク構造を利用した代表的な Web コミュニティ抽出手法として、Web グラフから MaxFlow アルゴリズムを用いて最小カットを抽出する手法[4]と、2 部グラフを用いる手法[5][6][7][8][9]が挙げられる。本節では、これらの研究について説明する。

MaxFlow アルゴリズムを用いた手法は Flake ら[4]によって提案された。Flake らは Web コミュニティを「同じ Web コミュニティに含まれる Web ページ間のリンク数が、同じ Web コミュニティに含まれない Web ページへのリンク数よりも多い Web ページの集合」と定義している。そして、MaxFlow アルゴリズムを用いて Web グラフからシードページを含む最小カットを求めることによって Flake らが定義しているような Web コミュニティを抽出する手法を提案している。

2 部グラフを用いる手法には 2 部グラフに HITS を

適応する手法と 2 部グラフを形成する部分グラフを抽出する手法がある。

2 部グラフに HITS を適応する手法は豊田[5]によって提案された。豊田はシードページを含む 2 部グラフに HITS を適応し、Authority スコアと Hub スコアが上位の Web ページ集合を Web コミュニティとした。

2 部グラフを形成する部分グラフ抽出する手法は Kumer ら[6]によって提案された。Kumer らは Web グラフに含まれる完全 2 部グラフ(CBG: Complete Bipartite Graph)を形成する部分グラフを Web コミュニティと定義している。Kumer らが提案した trawling により、Web グラフに含まれる全ての完全 2 部グラフ抽出することができる。

Kumer の研究以降、Web グラフから 2 部グラフを構成する部分グラフを見つけることにより Web コミュニティを発見する手法が数多く提案され、現在までに様々な改良が行われてきた[7][8][9]。

Reddy ら[7]は、Kumer らの手法の抽出条件を緩め、DBG を Web コミュニティとして抽出している。ここで、DBG とは以下のような条件を満たした有向グラフを指す。

- (1) ノードを各エッジ始点（以後、Fun）と終点（以後、Center）の 2 つのグループに分けたとき、同グループ中のノード間にエッジが存在しない。
- (2) 各 Fun が自然数 p 以上のエッジを持ち、かつ、各 Center が自然数 q 以上のエッジを持つ。

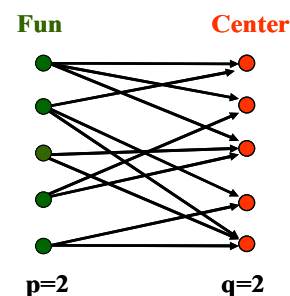


図 1. DBG ($p=2, q=2$)

Reddy らの手法により、Kumer らの手法よりも Web コミュニティに含まれる Web ページ数（以後、コミュニティサイズ）の大きな Web コミュニティを抽出することができるようになった。

齊田ら[8]は PlusDBG を提案している。PlusDBG では、まず、距離量を用いて Web コミュニティとトピックを共有しない Web ページを判別し、Web コミュニティのシードとの距離量が一定以内にある Web ページのみを、Web コミュニティの候補とする。そして、候補となった Web ページ集合から Reddy らの手法を用いて DBG を抽出する。

沈ら[9]は、Center に含まれる Web ページ間の関連の強さを示す連結度を定義し、Center 中の任意の 2 つの Web ページが一定以上の連結度を持つ Web ページ集合を Web コミュニティとして抽出する手法を提案した。沈らは Fun によって強く結ばれた Center のみを Web コミュニティとして抽出することによって、齊田らの手法によって抽出される Web コミュニティよりも高い適合率を有する Web コミュニティを抽出できるこ

とを示した．以下に Center 集合を T ，Fun 集合を S とし、沈らの手法の手順を示す．

- (1) シードページ t を選び， $T=\{t\}$ ， $S=\varnothing$ とする．
- (2) T に含まれる Web ページにリンクしている Web ページ集合を S' とする．
- (3) S' がリンクしている Web ページ集合から T に含まれる Web ページを除いたものを T' とする．
- (4) $t' \in T'$ において， t' が T に含まれる全ての Web ページと N 以上の連結度を持つかを判定．
 - (a) N 以上の連結度を持つならば， $T=\{t'\} \cup T$ ， $S=S \cup \{\text{連結子}\}$ とし，(2)へ．
 - (b) N 以上の連結度を持たないならば，他の t' を選び(4)へ．
- (5) $|S|>p$ ， $|T|>q$ ならば Web コミュニティとして Web 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．

ここで，Step. (5)の p は各 Fun が持つエッジの数の閾値であり， q は各 Center が持つエッジの数の閾値である．

沈らの手法では，抽出条件を厳しくすることにより適合率を向上させることに成功している．しかし，再現率の低下を考慮していないため，本来 Web コミュニティに含まれるべき Web ページを Web コミュニティから排除してしまっている可能性がある．実際に，沈らの手法によって抽出される Web コミュニティは，既存手法によって抽出される Web コミュニティと比べて高い適合率を有しているが，コミュニティサイズを比較してみると既存手法のコミュニティサイズよりも小さくなっている．Web コミュニティの精度を決定する尺度として再現率も重要な意味を持つことを考えると，適合率を低下させることなく再現率を向上させる必要があると考えられる．そこで我々は，[10]において，Web コミュニティの適合率を下げることなく再現率を向上させることを目的として，新たな Web コミュニティ抽出手法の提案を行った．

3. Web ページ間の関連性の伝播を用いた Web コミュニティ抽出

3.1. 定義

我々が提案してきた手法[10]は，Web グラフを無向グラフとして扱い，以下のようなアプローチで Web コミュニティを抽出する．まず，シードページと直接リンクで繋がれている Web ページを Web コミュニティに加える．以後，Web コミュニティのメンバ（以後，コミュニティメンバ）である Web ページから 1 ホップで到達できる Web ページ集合のうち， L 個以上のコミュニティメンバと繋がれている Web ページを再帰的にコミュニティメンバに加えていくことにより，Web コミュニティを抽出する．

ここで， L は最低リンク数と呼ばれる閾値であり，抽出過程で増加するコミュニティサイズ $|C|$ と，定数 M によって以下のように定義される．

$$L = \frac{|C|}{M} + 1 \quad (1)$$

定義式(1)中の M は，ユーザが Web コミュニティの適合率と再現率のバランスを決定するための定数である．コミュニティサイズ $|C|$ が小さい場合は最低リンク数 L も小さな値になるため，新規ページがコミュニティメンバになりやすく，逆に，コミュニティサイズ $|C|$ の増加にともなって最低リンク数 L も増加するため，新規ページがコミュニティメンバになりづらくなる．定数 M を変化させて比較実験を行った結果， M の値が小さいほど再現率が低く，適合率が高い Web コミュニティが抽出されることが確認できている．4 節で述べる比較実験では， $M=30$ における実験結果を用いる．

最低リンク数を設けた理由は，コミュニティメンバの増加にともない以下に挙げるような弊害が発生するからである．

- (1) Web コミュニティ中にコミュニティメンバとトピックを共有しない Web ページが含まれる可能性が高くなる．
- (2) コミュニティメンバが持つリンクの総数がコミュニティサイズに比例して増える．

コミュニティサイズに比例して抽出条件を厳しくすることによって，コミュニティメンバとトピックを共有しない Web ページを Web コミュニティから排除することができ，さらに，コミュニティサイズの爆発を防げる．

我々が提案してきた手法は「互いに関連を持つ Web ページ集合中の複数の Web ページとリンクで繋がれている Web ページは Web ページ集合と深い関連を持つ」という仮定に基づいている．ここで，リンクの向きを考慮しないのは，リンクで繋がれている Web ページ同士はリンクの向きに関わらず何らかの関連を持っていると考えたからである．つまり，我々の手法は「コミュニティメンバとどの程度リンクで繋がれているか」ということのみを抽出条件としている．

図 2 は我々が提案してきた手法により抽出されるグラフ構造を描画した図ある．図 2 を有向グラフで描画したのは，2 部グラフを用いた手法によって抽出される Web コミュニティのグラフ構造との違いを説明するためである．

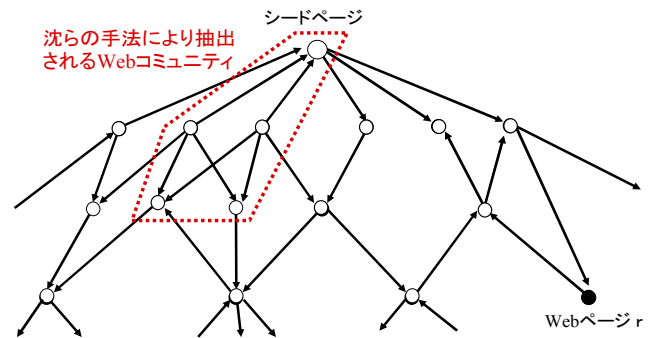


図 2. 我々の手法で抽出されるグラフの構造

図 2 中の点線で囲まれた部分グラフは沈ら手法によって抽出させる Web コミュニティのグラフ構造を示している．我々が提案してきた手法によって抽出される Web コミュニティは，シードページが持つリンク数が最低リンク数の定義式(1)中の M 未満である場合は，図 2 のように，沈らの手法によって抽出されるコミュ

ニティメンバを 100 パーセント包含する。

また、2 部グラフを抽出する手法では抽出できなかった Web ページもコミュニティメンバに含めることができる。例えば、図 2 中の Web ページ r は 2 部グラフを形成しないため 2 部グラフを用いた既存手法ではコミュニティメンバになりえなかった。しかし、Web ページ r もコミュニティメンバとトピックを共有している可能性があり、我々が提案してきた手法では、沈らの手法で抽出される Web ページ以外に Web ページ r のような Web ページを抽出することにより、再現率の向上を図っている。

さらに、2 部グラフを構成するような部分グラフは密な部分グラフであるため、2 部グラフを用いた既存手法では、シードページが持つリンク数（以後、次数）が小さい場合は Web コミュニティを抽出することができなかった。一方、我々が提案してきた手法では、次数の小さい Web ページがシードページとして与えられた場合でも Web コミュニティを抽出することが可能であり、この点の改善も図っている。

3.2. アルゴリズム

我々が提案してきた手法は、リンクデータより無向の Web グラフを作成し、Web グラフ中の任意のノードをシードノードとして、再帰的にコミュニティメンバを増やしていくことにより Web コミュニティを抽出する。以下に手順を示す。

- (1) シードノード t を選ぶ。
- (2) t から 1 ホップで到達できるノード集合を T_t とし Web コミュニティ C を $C = t \cup T_t$ とする。
- (3) 最低リンク数 L を $L = |C|/M + 1$ とする。
- (4) C 中の各 c から 1 ホップで到達できるノード集合のうち、 L 個以上のコミュニティメンバと繋がれているノード集合を C_{new} とし、 $C = C \cup C_{new}$ とする。
- (5) $|C_{new}| > 0$ であれば(3)へ。
- (6) C を Web コミュニティとして出力する。

Step. (3)において、最低リンク数 L を設けたのは、コミュニティサイズが大きくなるごとに関連の低い Web ページが Web コミュニティに含まれる可能性が高くなるためである。コミュニティメンバ数 $|C|$ の増加に比例して抽出条件を厳しくすることにより、コミュニティメンバとのリンクが少ない Web ページを排除することができる。次節の比較実験では、最低リンク数の定義式(1)中の M の値を 30 として実験を行った。

4. 比較実験と評価

4.1. データセット

データセットとして NTCIR-4 WEB task test collection [11]を用いた。NTCIR-4 WEB task test collection は 2001 年に収集された jp ドメインの Web ページと Web ページ間のリンクのデータであり、Web ページ数は 77,260,751、リンク数は 941,818,016 である。なお、実験はページ単位ではなくサイト単位で行った。ここで、サイトとは Web ページをドメイン名ごとにまとめた Web ページの集合を指し、ある 2 つのサイト間に複数のリンクがある場合はそれらのリンクを 1 つのリンクとして扱った。サイト単位で実験を行った理由

としては、Web ページよりもサイトの方が扱うトピックが明確であるということが挙げられる[12]。また、Yahoo!ディレクトリ[13]を用いた比較実験において、トップディレクトリから深さが 3 までのディレクトリしか用いていないことが挙げられる。つまり、比較するトピックが細かい分類ではないため、Web ページ単位ではなく、より大きな分類のトピックを扱うサイト単位で実験を行ったということである。なお、Yahoo!ディレクトリを用いた比較実験の詳細については次節で説明する。

本研究では、前処理として以下のようなサイトを削除した。まず、既存研究[6][7][8][9]に倣い、有名なサイトを削除する。本研究では有名なサイトとして、インリンク数がデータ全体の上位約 1%に相当する 200 以上のインリンクを有するサイトを削除した。Kumerらは文献[6]にて、非常に多くのインリンクを有する Web ページはリンク元の Web ページ集合が扱うトピックの種類が多くなると述べている。さらに、非常に多くのインリンクを有する Web ページを Web グラフから削除しないで Web コミュニティ抽出を行うと適切な Web コミュニティが抽出されないと述べている。1 節で述べた Web コミュニティの使用法を想定した場合においても、有名なサイトを Web コミュニティのメンバとして抽出する利点は少ないと考えられる。以上のようなことを考慮し、本研究では有名なサイトの削除を行った。

次に、リンク先が複数のトピックを扱うサイトとして、リンク集を削除した。本研究ではリンク集として、アウトリンク数がデータ全体の上位約 1%に相当する 200 以上のアウトリンクを有するサイトを削除した。有名なサイトと同様の理由から、リンク集を削除しないで Web コミュニティ抽出を行うと適切な Web コミュニティが抽出されないと考えられる。

以上のような前処理を行った結果、対象とするデータセットに含まれるサイト数は約 26.7 万、リンク数は約 494 万となった。

4.2. 適合率の評価

本研究では、沈らの研究と同様に、抽出された Web コミュニティ内の各サイトが Yahoo! ディレクトリの同一のディレクトリに登録されている割合を調べることより適合率の評価を行う。実験は、データセットの全サイト中 Yahoo!ディレクトリに登録されている 91,197 サイトをシードサイトとして行った。以下に、評価手法の詳細を示す。

まず、任意サイト a 、 b に対して $score(a,b)$ を以下のように定義する。

$$score(a,b)=1, \text{ サイト } a, b \text{ が同一のディレクトリに登録されている} \quad (2)$$

$$score(a,b)=0, \text{ サイト } a, b \text{ が同一のディレクトリに登録されていない} \quad (3)$$

ここで同一のディレクトリとは、トップディレクトリから深さ D までのディレクトリが共通していることを指す。例えば、 $D=2$ の場合、Yahoo!ディレクトリの「トップ/趣味とスポーツ/ゲーム/オンラインゲーム」に登録されているサイトと、「トップ/趣味とスポーツ/ゲーム/パソコンゲーム」に登録されているサイトは、

トップディレクトリからの深さが 2 である「ゲーム」までが共通しているので $score$ は 1 となる。

次に $score$ を用いて、Web コミュニティ C の適合率 $P(C)$ を以下のように定義する。

- (1) Web コミュニティ C 中のサイトで Yahoo!ディレクトリに登録されているサイトの集合を YC とした場合、 YC に含まれるサイト数 $|YC|$ が 3 未満の場合は評価を行わない。
- (2) YC 中のサイト r に対して、 r のスコア $Pscore(r)$ を以下のように定義する。

$$Pscore(r) = \frac{\sum_{x \in (YC-r)} score(r, x)}{|YC-r|} \quad (4)$$

- (3) Web コミュニティ C の適合率 $P(C)$ を以下のように定義する。

$$P(C) = \frac{\sum_{r \in YC} Pscore(r)}{|YC|} \quad (5)$$

図3に $D=2$ 、図4に $D=3$ の場合の適合率を示す。また、表1に各手法で抽出されたWebコミュニティの $P(C)$ 値の平均を比較した表を示す。

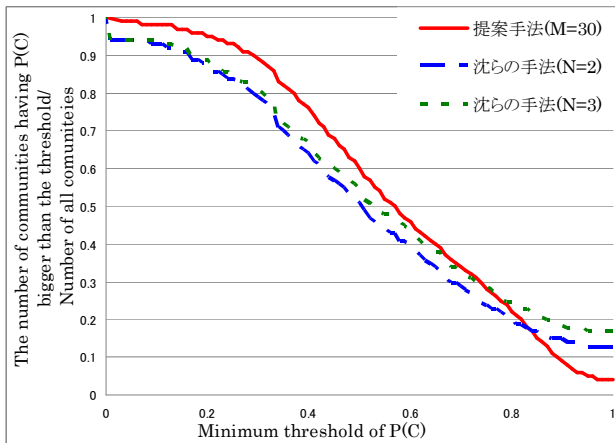


図 3. 適合率ごとの Web コミュニティの割合 ($D=2$)

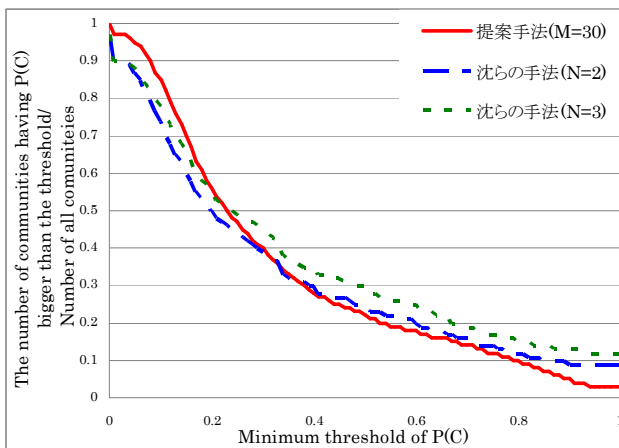


図 4. 適合率ごとの Web コミュニティの割合 ($D=3$)

表 1. Web コミュニティの $P(C)$ 値の平均

	D=2	D=3
提案手法(M=30)	0.58	0.33
沈らの手法(N=2)	0.57	0.36
沈らの手法(N=3)	0.59	0.39

図 3、図 4 より高い適合率を有する Web コミュニティの割合は沈らの手法の方が高いが、 $P(C) < 0.8$ の場合は提案手法でも沈らの手法と同等の適合率を有する Web コミュニティを抽出できるということが確認できる。また、表 1 より両手法の平均適合率に大きな差はないことが確認できる。

4. 3. 再現率の評価

表 2 に、Yahoo!ディレクトリに登録されている 91,197 個のサイトを用いた比較実験における、提案手法と沈らの手法の Web コミュニティ数、平均サイト数、抽出されたのベサイト数を示す。ここで、抽出されたのベサイト数の値は Web コミュニティ数と平均サイト数を乗算した値であり、複数の Web コミュニティに属するサイトは属する Web コミュニティの数だけ重複して数えられている。表 2 中の各手法における括弧内の値は、沈らの手法では連結度を、提案手法では最低リンク数の定義式(1)中の定数 M を示している。

表 2. Web コミュニティ数、平均サイト数、抽出されたのベ Web サイト数

抽出手法	Webコミュニティ数	平均サイト数	コミュニティサイズの合計
提案手法(M=30)	63,859	171	10,919,889
沈らの手法(N=2)	47,423	91	4,315,493
沈らの手法(N=3)	34,060	66	2,247,960

表 2 に示した 1 つの Web コミュニティ当たりの平均サイト数を比較すると、提案手法の方が沈らの手法よりも大きいことが分かる。これは、提案手法が沈らの手法では抽出することができなかった DBG を構成しないサイト抽出しているためである。適合率の比較実験により、提案手法によって抽出される DBG を構成しないサイトを含んだ Web コミュニティの適合率は沈らの手法によって抽出される Web コミュニティの適合率と同等であることが確認できた。これにより、DBG を構成しないサイトにもシードサイトとトピックを共有するサイトが存在することが示された。そして、提案手法を用いることにより、シードサイトとトピックを共有し、かつ、DBG を構成しないサイト集合を抽出できることが示された。以上のことから、提案手法を用いることにより、沈らの手法では Web コミュニティから排除されていたシードサイトとトピックを共有するサイトもコミュニティメンバとして抽出できることが確認できた。

また、Web コミュニティ数に関しても、提案手法の方が沈らの手法よりも多くのコミュニティを抽出していることが分かる。これは、提案手法は次数の少ないサイトがシードサイトとして与えられた場合でも、シードサイトとトピック共有するサイト集合を Web コミュニティとして抽出できるからだと考えられる。

4.4. シードサイトの次数を限定した評価

提案手法が次数の少ないサイトがシードサイトとして与えられた場合でも、シードサイトとトピック共有するサイト集合を Web コミュニティとして抽出できることを確かめるために、シードサイトの次数を 10 以下、5 以下と限定させて実験を行った。図 4～図 16 に実験結果を示す。

4.4.1. シードサイトの次数を限定した場合の Web コミュニティ数の比較

図 5～図 7 は $D=2$, 図 8～図 10 は $D=3$ の場合の適合率ごとの Web コミュニティの積算の数を示した図である。ここで、図 6, 図 7 及び、図 9, 図 10 はシードサイトの次数を 10 以下、5 以下と限定させて実験をした場合の実験結果を示した図である。図 5～図 10 の横軸は $P(C)$ の値の閾値であり、縦軸はその閾値以上の $P(C)$ を有する Web コミュニティの数である。また、図 5～図 10 の各手法における括弧内の値は、沈らの手法では連結度を、提案手法では最低リンク数の定義式(1)中の定数 M を示している。

図 6, 図 7 及び、図 9, 図 10 より、シードサイトの次数が小さいほど再現率の差は大きくなっていることが分かる。実際、シードサイトの次数の制限を 4 以下～15 以下と変化させて実験を行った場合、次数が小さいほど再現率の差は大きくなるという結果が得られた。これにより、我々の手法を用いることで沈らの手法よりも多くサイトをシードサイトとして Web コミュニティ抽出を行うことができ、シードサイトの次数が小さい場合は、沈らの手法よりも我々の手法の方が高い再現率を有する Web コミュニティを抽出できることが確認できた。

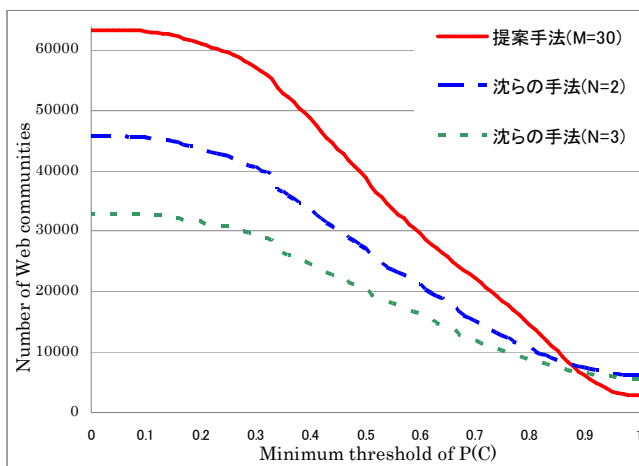


図 5. 適合率ごとの Web コミュニティの積算 ($D=2$)

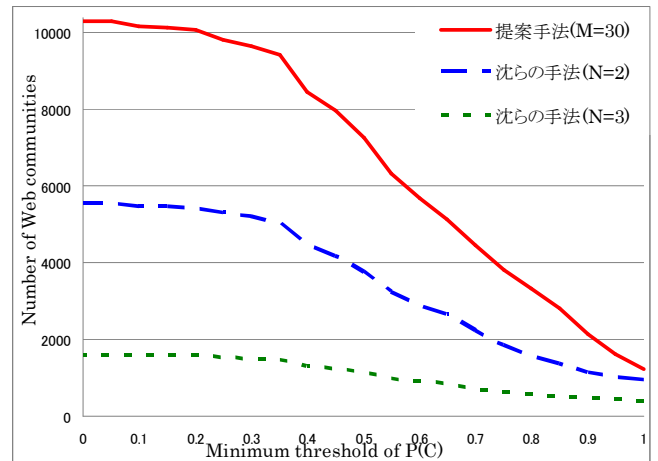


図 6. 適合率ごとの Web コミュニティの積算 ($D=2$, シードサイトの次数が 10 以下)

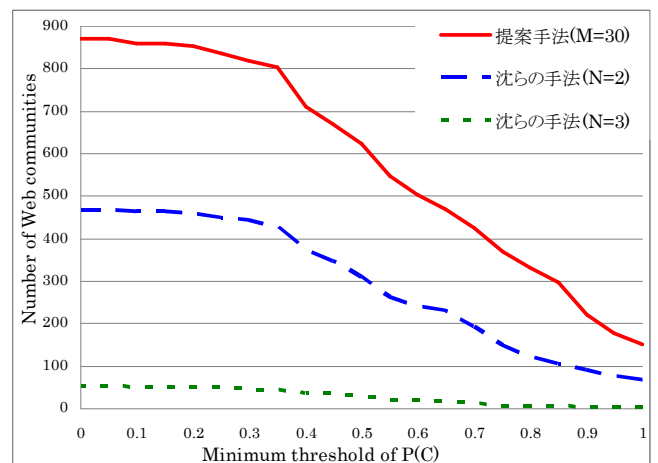


図 7. 適合率ごとの Web コミュニティの積算 ($D=2$, シードサイトの次数が 5 以下)

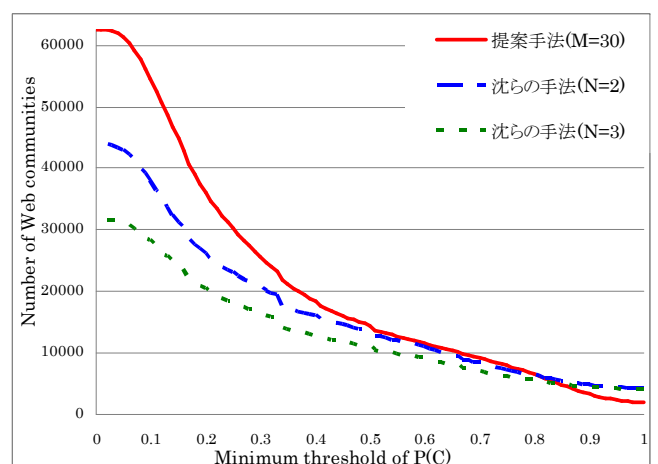


図 8. 適合率ごとの Web コミュニティの積算 ($D=3$)

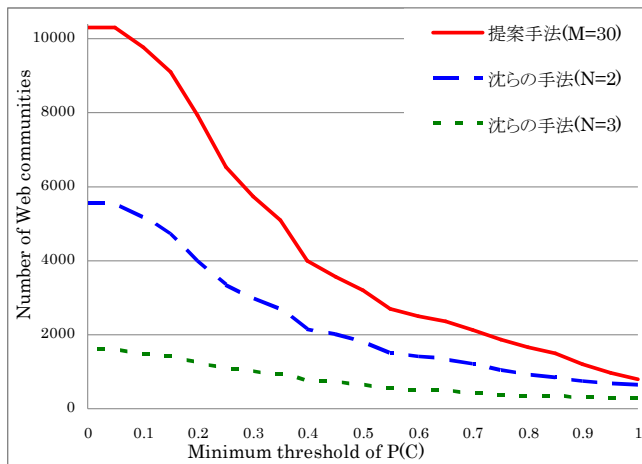


図 9. 適合率ごとの Web コミュニティの積算
($D=3$, シードサイトの次数が 10 以下)

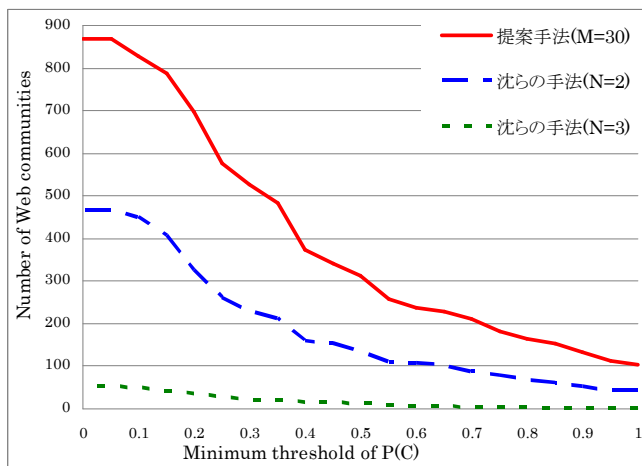


図 10. 適合率ごとの Web コミュニティの積算
($D=3$, シードサイトの次数が 5 以下)

4.4.2. シードサイトの次数を限定した場合の上位の

適合率を有する Web コミュニティ数の比較

図 11～図 13 に $D=2$, 図 14～図 16 に $D=3$ の場合の適合率ごとの Web コミュニティの数を示す。ただし、図 12, 図 13 及び、図 15, 図 16 はシードサイトの次数を 10 以下, 5 以下と限定させて実験をした場合の実験結果を示した図である。図 11～図 16 の横軸は $P(C)$ の値であり、縦軸は横軸の値以下の $P(C)$ を有する Web コミュニティの数である。例えば、横軸の値が 1.0 の場合、 $P(C)$ の値が 0.9 より多く 0.1 以下である Web コミュニティの数が示される。ここで、横軸の値が 0.1 の場合のみ、 $P(C)$ の値が 0 以上 0.1 以下である Web コミュニティの数が示される。また、図 11～図 16 の各手法における括弧内の値は、沈らの手法では連結度を、提案手法では最低リンク数の定義式(1)中の定数 M を示している。

図 12, 図 13 及び、図 15, 図 16 より、0.8 以上の $P(C)$ を有する Web コミュニティの数を比較したとき、シードサイトの次数が小さくなるほど差が大きくなっていることが分かる。また、次数が小さいほど再現率の差は大きくなっていることが分かる。これにより、我々の手法は次数の少ないサイトがシードサイトとして与

えられた場合でも、シードサイトとトピック共有するサイト集合を Web コミュニティとして抽出できることが示され、シードサイトの次数が小さくなるほど抽出される Web コミュニティの適合率が沈らの手法よりも改善されることが確認できた。

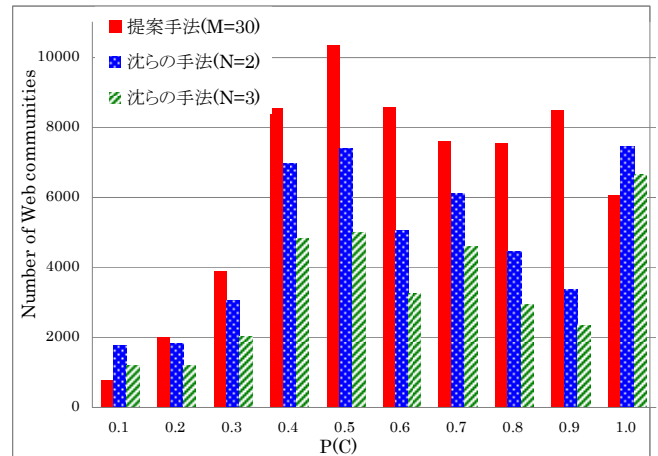


図 11. 適合率ごとの Web コミュニティの数($D=2$)

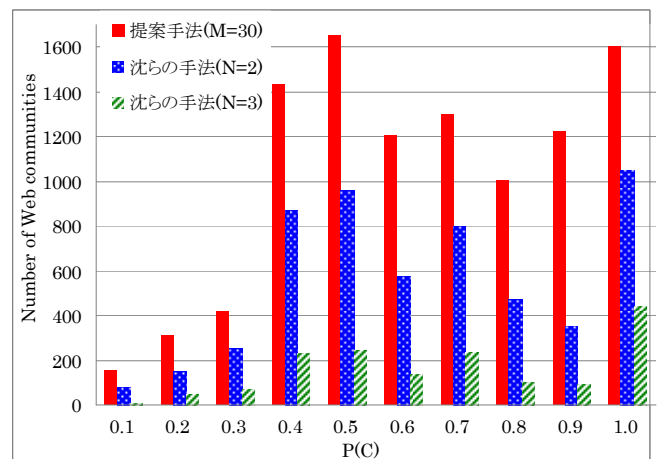


図 12. 適合率ごとの Web コミュニティの数
($D=2$, シードサイトの次数が 10 以下)

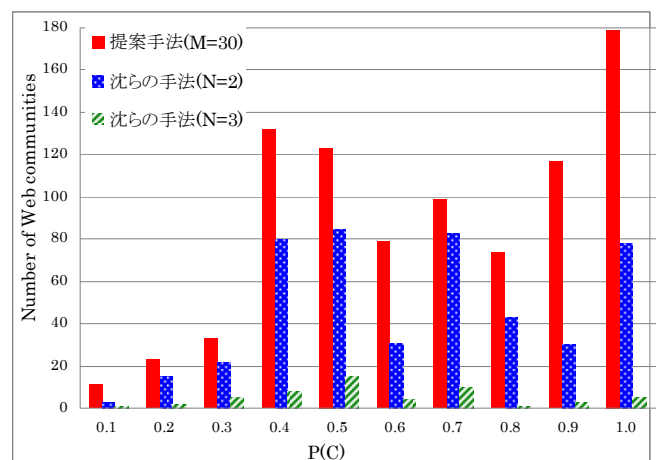


図 13. 適合率ごとの Web コミュニティの数
($D=2$, シードサイトの次数が 5 以下)

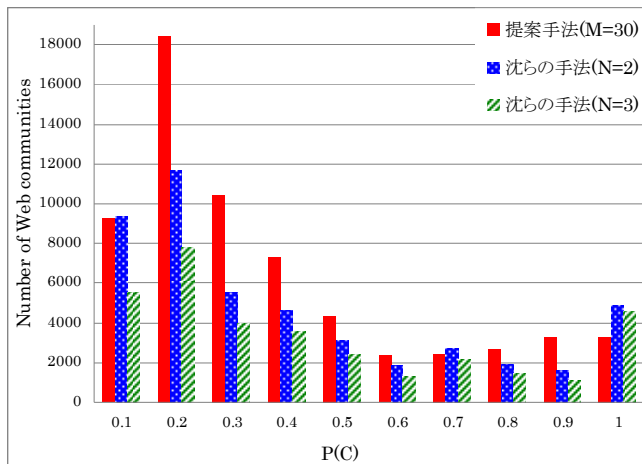


図 14. 適合率ごとの Web コミュニティの数 (D=3)

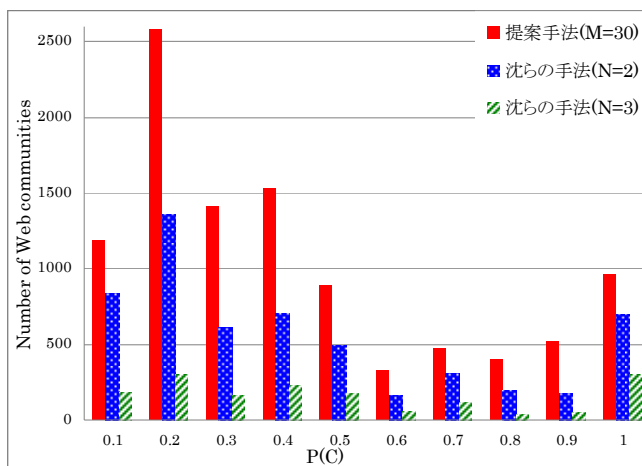


図 15. 適合率ごとの Web コミュニティの数 (D=3, シードサイトの次数が 10 以下)

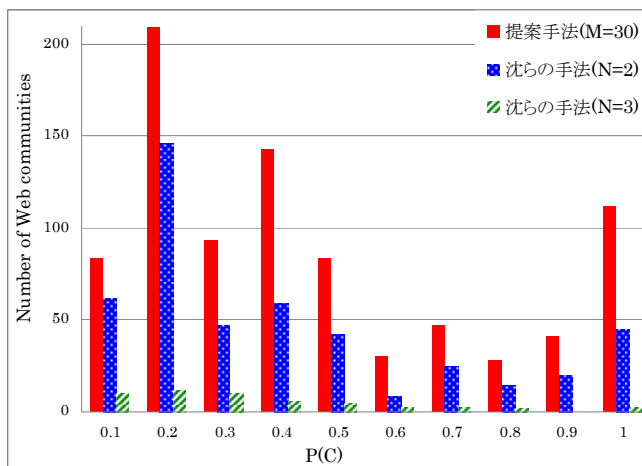


図 16. 適合率ごとの Web コミュニティの数 (D=3, シードサイトの次数が 5 以下)

5. おわりに

本論文では、我々が提案してきた手法と沈らの手法との再現率における比較実験を行った。比較実験の結果、提案手法は既存手法と同等の適合率を保ちつつ再現率を向上させることができることが確認できた。以

下は提案手法の特徴である。

- (1) 提案手法では、沈らの手法よりも多くのサイトをシードサイトにする事ができ、それによって多くの Web コミュニティを抽出できる。
- (2) 提案手法では、沈らの手法では抽出できなかったシードサイトとトピックを共有するサイトをコミュニティメンバとして抽出できる。
- (3) 提案手法では、シードサイトの次数が小さくなるほど抽出される Web コミュニティの適合率が沈らの手法よりも改善される。

今後は、抽出される Web コミュニティの特徴を活かした Web コミュニティ利用法を提案していきたい。

謝辞

本研究の一部は、科学研究費補助金「情報爆発に対応する高度にスケーラブルなモニタリングアーキテクチャ」によるものである。

参 考 文 献

- [1] 高野明彦, 西岡真吾, 今一修, 岩山真, 丹羽芳樹, 久光徹, 藤尾正和, 徳永健伸, 奥村学, 望月源, “汎用連想計算エンジンの開発と大規模文書分析への応用”, 第 19 回 IPA 技術発表会, 2000.
- [2] 谷口智哉, 松尾豊, 石塚満, “Blog コミュニティの抽出と分析”, 人工知能学会 SIG-SWO, 2004.
- [3] 小野拓史, 豊田正史, 喜連川優, “リンク解析を用いたウェブ上のスパム発見手法に関する一考察”, DEWS, 2006.
- [4] G.W.Flake, S.Lawrence, and C.L. Giles, “Efficient identification of Web communities”, Proc. of ACM SIGKDD, pp.150-160, 2000.
- [5] 豊田正史, “WWW における関連コミュニティ群の発見”, DBS 2000.
- [6] S. Ravi Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, “Trawling the Web for Emerging Cyber-Communities”, Proc. of 8th WWW Conference, pp.403-416.1999.
- [7] P.Krishna Reddy and Masaru Kitsuregawa, “An Approach to Relate the Web Communities through Bipartite Graphs”, Proc. of the Second International Conference on Web Information Systems Engineering, Vol.1, pp.301-310, 2001.
- [8] 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.
- [9] 沈垣甫, 田浦健次郎, 近山隆, “Web ページ間の関連性の伝播を用いた Web コミュニティ抽出手法”, DEWS 2007.
- [10] 飯村卓也, 平手勇宇, 山名早人, “Web ページ間の関連性の伝播を用いた Web コミュニティ抽出手法”, iDB 2008.
- [11] NTCIR Project, <http://research.nii.ac.jp/ntcir/>
- [12] Krishna Bharat, Bay-Wei Chang, Monika Henzinger and Matthias Ruhl, “Who links to whom: Mining linkage between web sites”, IEEE International Conference on Data Mining, pp.51-58, 2001.
- [13] Yahoo!カテゴリ, <http://dir.yahoo.co.jp/>