DEWS2006 3B-i8

リンク構造解析による不要Webコミュニティの判別

斉田　直幸† 　山名　早人†,††

† 早稲田大学大学院理工学研究科
†† 国立情報学研究所
E-mail: †nao@yama.info.waseda.ac.jp

あらまし　Webコミュニティ抽出は，Webのリンク構造から特徴的な部分構造を抽出することで，WWWを意味的に分類する手法である．Webコミュニティ抽出は，Web利用者への適切なナビゲーションや，Webの意味構造の解析に対して効果が期待されている．しかしながら，Webコミュニティをリンク構造の定義のみから導き出す手法では，抽出されるWebコミュニティが，Webの利用者にとって意味的に正確に分類されていない場合も多く存在する．そこで，本研究では，Webコミュニティの精度の向上を目的として，Webコミュニティ抽出手法 PlusDBG [8] によって抽出されたWebコミュニティの解析と考察を行なう．さらに，ノイズとして抽出されるWebコミュニティを機械的に判別し，これを削除する手法について検討する．

キーワード　データマイニング, Webとインターネット, 知識発見, Webコミュニティ抽出，リンク構造解析

Noisy Web Community Detection based on Link Structure

Naoyuki SAIDA† and Hayato YAMANA†,††

† Graduate School of Science & Engineering，WASEDA University
†† National Institute of Informatics
E-mail: †nao@yama.info.waseda.ac.jp

Abstract　Web Mining aims to discover valuable informations from the huge WWW. Web community extraction is contained in a division of Web Mining. Web community extraction clusters web pages into the semantic sets by extracting characteristic link substructure from the link structure of WWW. Web community extraction is expected for web navigation and analysis of semantic structure of WWW. However, web community extraction only based on the definition of the link structure extracts the web communities which is semantically incorrect for web users. Then, this paper analyze the web communities extracted by web community extraction scheme called PlusDBG [8] to improve the precision of web communities. Moreover, we discuss about noisy web community detection scheme.

Key words　Data Mining, Web and Internet, Knowledge Discovery, Web Community Extraction, Link Analysis

1. はじめに

近年のWWWの普及に伴って，WWW上のデータ量も急速に増大している．Netcraft社の調査によると，2005年12月の時点でのWebサーバ数は，74,353,258サーバが確認されており，2004年12月の時点でのWebサーバ数である，56,923,737から大幅に増加した[1]．WWW全体でのWebページ数に関しては，正確な数はわかっていないが，1999年2月の時点で8億ページ以上存在すると推定されており[2]，Webサーバ数の増加とともに，大幅な増加を見せていると考えられる．また，Gulliらは，2005年1月の時点で，様々な検索エンジンにてインデクシングされているWebページ数が，115億ページを超えると推測している[3]．

WWW上のデータは，様々な目的や意図を持った無数のユーザの手によって作成され，データが扱う内容も，データの質も幅広いものとなっている．加えて，WWW上に存在するデータの形式についても，テキスト情報やタグ情報，画像や動画などのマルチメディア情報など，様々な形式が存在し，WWW上のデータを統一的に扱うことは困難である．

このように，WWWには，実際に収集され，インデックス化されているものだけでも無数のデータが存在し，その内容も，質も統一されていない．そのため，現在では，WWW上の膨大なデータから価値のある情報を発見するWebマイニング技術が，WWWを有効に活用するための必要不可欠な技術となっている．

Webマイニングにおける一分野として，Webコミュニティ

抽出が存在する．Web コミュニティとは，“共通のトピックを持った Web ページの集合”と定義される．Web コミュニティ抽出では，Web のリンク構造から特徴的な部分構造を抽出することで，Web コミュニティを抽出し，WWW を意味的に分類・整理する．Web コミュニティ抽出には，WWW を意味的に分類することによって，以下の 2 つの点が期待される．

（1） 特定のトピックに対して興味を持つユーザに対しての，適切で価値のある Web ページの提示による，効率的なナビゲーションの提供

（2） トピック単位で分類された Web を利用してのマイニングや，WWW の意味構造の解析．

従来，Web ページをトピックごとに分類する作業は，主に Yahoo! [4] や Open Directory Project [5] のような Web ディレクトリサービスにおいて，人の手によって行われてきた．しかし，WWW に対して，人手による分類作業には限界があり，膨大で，頻繁に更新されている Web ページの全てを適切に分類することは不可能に近い．そこで，大規模な Web データから，Web ページを機械的に分類・整理する手法として，Web コミュニティ抽出手法が注目されている．

従来の Web コミュニティ抽出に関する研究には，1999 年に Kumar らが提案した trawling [6] や，2000 年に Flake らが提案した MaxFlow アルゴリズムによる Web コミュニティ抽出手法 [7] などがあげられる．これらの研究では，主に Web コミュニティのリンク構造を定義することが研究の焦点となっている．Kumar らは，“完全二部グラフを含む，十分に密で，十分に大きな二部グラフ” を，Web コミュニティとして定義している．Flake らは，“Web コミュニティのメンバとのリンク数が，それ以外の Web ページとのリンク数よりも多い Web ページ”を Web コミュニティとして定義している．しかしながら，Web コミュニティをリンク構造の定義のみから導き出す手法では，抽出される Web コミュニティが，Web の利用者にとって意味的に正確に分類されていない場合も多く存在する．

そこで，本研究では，Web コミュニティの精度の向上を目的として，PlusDBG [8] によって抽出された Web コミュニティの解析と考察を行なう．PlusDBG は，高い精度と網羅性を持った Web コミュニティを抽出することを目的として，2005 年に著者らが提案した手法である．さらに，ノイズとして抽出される Web コミュニティを機械的に判別し，これを事前に削除する手法について検討する．

2 節では，Web コミュニティ抽出に関する従来の研究について述べる．3 節では，本論文の Web コミュニティ解析において，Web コミュニティ抽出に用いた PlusDBG について説明する．4 節では，PlusDBG によって抽出した Web コミュニティの解析を行い，Web コミュニティ抽出の改善点を考察する続いて，5 節で，ノイズとして抽出される Web コミュニティを事前に判別する手法について検討する．最後に，6 節において，本論文のまとめを行う．

2. 関連研究

WWW は，Web ページを頂点とし，リンクを辺とする巨大な有向グラフとしてとらえることができる [9]．従来の Web コミュニティ抽出に関する研究では，このような Web グラフから，Web コミュニティに特徴的なリンク構造を定義し，その構造を抽出する手法を提案している．本節では，Web コミュニティ抽出手法に関する代表的な研究として，trawling [6], Max Flow アルゴリズムによる Web コミュニティ抽出手法 [7], DBG 抽出手法 [10] について説明する．

Kumar らの提案した trawling [6] では，Web コミュニティを，“十分に大きく，十分に密な 2 部グラフ”として定義している．また，このような Web コミュニティの多くは，そのコアとして完全 2 部グラフ (CBG: Complete Bipartite Graph) を構成する Web ページ集合を，1 つ以上含んでいると仮定し，WWW データセットから全ての完全 2 部グラフを抽出する手法を提案している．

また，村田は，Web 検索エンジンを利用して，与えられたシードセットから完全 2 部グラフを Web コミュニティとして抽出する手法を提案している [11]．

Flake らは，“Web コミュニティのメンバとのリンク数が，それ以外の Web ページとのリンク数よりも多い Web ページ”を Web コミュニティのメンバとして定義している [7]．そして，Web グラフに対して $s-t$ 最大フロー問題 [12] を解くことで，Web コミュニティとそれ以外とを切り離す辺の集合 (最小カット) を求めることができると示している．

また，伊野らは，Flake らの Web コミュニティの定義を厳密にした IKN-community を定義し，IKN-community を抽出するためのアルゴリズムについて提案している [13]．

Reddy らは，Kumar らの定義を継承し，Web コミュニティを密な二部グラフ（DBG: Dense Bipartite Graph）と定義している [10]．Reddy らの定義する DBG は，完全 2 部グラフよりも条件が緩やかであり，コアを含まない Web コミュニティも抽出が可能である．DBG は，Fan と $Center$ から成り，“各 Fan が，少なくとも q 以上の $Center$ にリンクしている”，“各 $Center$ が，少なくとも p 以上の Fan からリンクされている”の 2 点を満たす二部グラフとして定義される．

trawling, Max Flow，DBG 抽出のいずれの研究も，Web コミュニティを定義し，その定義を満たす Web 構造を Web グラフから抽出することに焦点が当てられている．しかしながら，Web コミュニティを定義のみから導き出す手法では，抽出される Web コミュニティが，Web の利用者にとって意味的に正確に分類されていない場合も多く存在する．

3. PlusDBG

著者らが 2005 年 3 月に提案した PlusDBG [8] は，DBG 抽出手法 [10] に，“距離量”の概念を加えることで，精度と網羅性を向上させた手法である．“距離量”とは，Web コミュニティ中のメンバが持つトピックが，シードとして入力された Web ページの持つトピックから，どれだけ離れているかを表す値である．PlusDBG は，シードを入力として，シードから一定の“距離量”内にある Web ページを，共参照関係 (Web ページ a, b が共通のページに対してリンクしている場合の a, b 間の関係)

を繰り返し辿ることによって探し出し，そこから DBG (Dense Bipartite Graph) を抽出する．

従来手法では，Web コミュニティ抽出時に，複数のトピックを持った Web ページが存在する場合，その Web ページを経由して，異なるトピックを持った Web ページが Web コミュニティに含まれてしまう．PlusDBG は，“距離量”を用いることで，本来のトピックとは異なったトピックを持つ Web ページが，Web コミュニティのメンバとなることを防いでいる．

3.1 PlusDBG の “距離量”

PlusDBG において，シードページ s を起点として，共参照関係を n 回辿った時点で発見された Web ページ p までの “距離量”は，式 (1), (2) において定義される．

$$Dis(p, F_{n-1}(s)) = (1 - Sim(p, F_{n-1}(s))) + Dis(q, F_{n-2}(s)) \quad \text{for } n \geqq 2 \tag{1}$$

$$Dis(p, F_0(s)) = (1 - Sim(p, \{s\})) \text{ for } n = 1 \tag{2}$$

ここで，$F_n(s)$ は，s をシードとして，Web ページの共参照関係を n 回辿った時点で発見された Web ページ集合を指す．$F_0(s)$ はシード s のみの集合であり，$F_1(s)$ は，$F_0(s)$ に，s と共参照関係にある Web ページ中で距離の条件を満たす Web ページを加えた集合である．また，q は，F 中に存在する，p がリンクしている Web ページに対して最も多く共通にリンクしている Web ページとする．

さらに，式 (1), (2) における $Sim(p, F)$ を，式 (3) で定義する．

$$Sim(p, F) = \frac{|children(p) \cap children(F)|}{|children(p)|} \tag{3}$$

ここで，$children(p)$ とは，ページ p からリンクされている Web ページの集合を指す．また $children(F)$ はページ集合 F 中の Web ページからリンクされている Web ページの集合を指す．

つまり，$Sim(p, F)$ は，Web ページ p が，F の持つトピックに対してリンクを張っている割合を示している．複数のトピックを持つ Web ページの場合，一つのトピックに対するリンクの割合は，単一のトピックしか持たない Web ページに比べて低くなる．そのため，Sim の利用によって，複数トピックの Web ページに対して，単一トピックの Web ページより大きな “距離量”を与えることが可能となる．

最終的に，PlusDBG は，閾値以下の “距離量”を持つ Web ページのみを利用して DBG を抽出する．

3.2 PlusDBG アルゴリズム

PlusDBG による DBG 抽出では，入力としてシードページ s を受け取り，s と同じトピックを持つページの集合である Web コミュニティ C_s を出力する．アルゴリズムの流れは，Reddy らの DBG 抽出手法 [10] と同様であり，以下の 2 つのステップで構成される．

（1） s と同じトピックを持つ Web ページ集合を，Web コミュニティ候補として探索

（2） ステップ 1 で得られた Web ページ集合から，DBG を構成する Web ページを抽出

3.2.1 STEP1: Web コミュニティ候補の探索

PlusDBG では，まず，シードページ s から，s と関連性のある Web ページを共参照性を利用した繰り返し作業で探索していく．その手順は以下のようになる．

（1） “距離量”の閾値 dis_border を指定する．

（2） Web ページ集合 F を定義し，シードページ s を F に追加する．

（3） F に新たに追加された Web ページによる集合を F_{new} とし，$F_{new} = F$ とする．

（4） $|F_{new}| > 0$ である間，以下の手順 (a)，(b)，(c) を繰り返す．

（a） F_{new} と共参照関係にある，すなわち $|children(F_{new}) \cap children(p)| > 0$ 且つ $p \notin F$ を満たす Web ページの集合 $\{p\}$ を探し出す．

（b） 各 $p \in \{p\}$ において，F との距離量 $Dis(p, F)$ を計算する．

（c） $F_{new} = \{p | p \in \{p\}, Dis(p, F) \leqq dis\_border\}$，$F = F \cup F_{new}$ とする．

（5） 得られた F を Web コミュニティ候補として出力する．

3.2.2 STEP2: DBG 抽出

STEP1 で Web コミュニティ候補 F を得たあと，Reddy らの手法と同様の手順で DBG を構成する Web ページ集合を見つけ出し，これを最終的な Web コミュニティとする．DBG の抽出手順は以下のようになる．

（1） Fan の持つアウトリンク数の閾値 u と，$Center$ の持つインリンク数の閾値 v を与える．

（2） STEP1 により得られた Web ページ集合 F を Fan とし，F からリンクされている Web ページ集合を $Center$ とする．

（3） Fan と $Center$ を構成する Web ページ数が共に収束するまで，以下の手順 (a)，(b) を繰り返し，DBG の定義を満たす Web ページのみを抽出する．

（a） $Center$ に含まれる Web ページに対して，u 未満のリンクしか持たない Fan 中の Web ページを削除する．

（b） Fan に含まれる Web ページから，v 未満のリンクしか受けていない $Center$ 中の Web ページを削除する．

（4） 最終的に残された Fan と $Center$ を Web コミュニティとして出力する．

4. Web コミュニティ解析

本節では，Web コミュニティ抽出の精度を向上させることを目的として，PlusDBG によって抽出された Web コミュニティを解析する．[8] において，PlusDBG と従来の DBG 抽出手法 [10] の比較により，PlusDBG が，従来手法よりも高い精度と網羅性を持って Web コミュニティを抽出できることが分かった．そこで，本論文では，[8] において 1 に固定していた PlusDBG の “距離量”の閾値を変化させることで，Web コミュニティの性質に変化が現れないか比較する．ここで，精度は，Web コミュニティ中に存在する，Web コミュニティとトピックが一致している Web ページの割合とする．

4.1 データセット

NTCIR-4 データセット [14] から，PlusDBG を用いて可能な限り全ての Web コミュニティを抽出する．

データセットは，複製ページのマージと，リンク数による枝刈りによって前処理を行なう．複製ページのマージでは，以下の 2 つの条件を満たす Web ページ a, b を複製ページであると定義する．

（1） a, b が共に 10 リンク以上のアウトリンクを持っている。

（2） $|children(a) \cap children(b)|$ が，a, b それぞれの総アウトリンク数のうち 90 %以上を占めている．

複製ページと判断された Web ページ a, b はマージし，1 つの Web ページとして扱う．リンク数による枝刈りでは，各 Web ページのアウトリンク数とインリンク数に閾値を設け，いずれかの閾値を超えた Web ページを削除する．本研究では，アウトリンク数，インリンク数がそれぞれ 3 以下の Web ページを削除する．また，インリンク数が 50 以上の Web ページも削除する．

データセット中の，アウトリンクを持つ全ての Web ページをシードとし，“距離量”の閾値を 0.8 から 1.2 として，各シードから Web コミュニティを抽出した (P0.8，P0.9，P1.0，P1.1，P1.2)．また，データセット中で，Yahoo!Japan ディレクトリに登録されている Web ページのみをシードとして，同様に Web コミュニティを抽出した (YP0.8，YP0.9，YP1.0，YP1.1，YP1.2)．各 Web コミュニティにおける，Fan と $Center$ の閾値は，それぞれ 3 とする．

4.2 “距離量”による変化

まず，図 1 に，“距離量”の閾値を変化させた場合に抽出される Web コミュニティのサイズの分布を示す．本研究では，Web コミュニティのサイズを，DBG 中の $Center$ に含まれる Web ページ数と定義する．P0.8 と P1.0 を比較すると，ほぼ全てのサイズで，P1.0 の Web コミュニティ数が P0.8 よりも増加している．しかし，P1.0 と P1.2 の比較では，P1.2 は，大きなサイズの Web コミュニティ数が増加している反面，小さなサイズの Web コミュニティ数は全く増加していない．さらに P1.0 から P1.2 では，全 Web コミュニティのサイズの合計値が 2 倍以上に増加しているのに対して，総 Web コミュニティ数自体はほとんど増加していない．つまり，P1.0 は，P0.8 では抽出されなかった Web コミュニティを多く抽出しているが，P1.2 では，P1.0 で抽出された Web コミュニティを拡大しているだけだと考えられる．

PlusDBG における “距離量”の計算では，3. 節における式 (3) で定義されるように，$Sim(p, F)$ のとりうる値は [0, 1) の間となる．つまり，“距離量”の閾値を 1.0 に設定することで，シードページと直接の共参照関係にある Web ページは全て，Web コミュニティ候補に含まれるようになる．P1.0 においては，シードページと直接の共参照関係にあれば，関連性の低い Web ページであっても必ず Web コミュニティ候補となるため，P0.8 に比べて Web コミュニティを抽出しやすい，と考えることができる．逆に，P1.2 では，P1.0 と比較した場合に，関連性の高い Web ページしか Web コミュニティ候補に追加できないため，Web コミュニティ数が増加しにくいのではないかと考えられる．

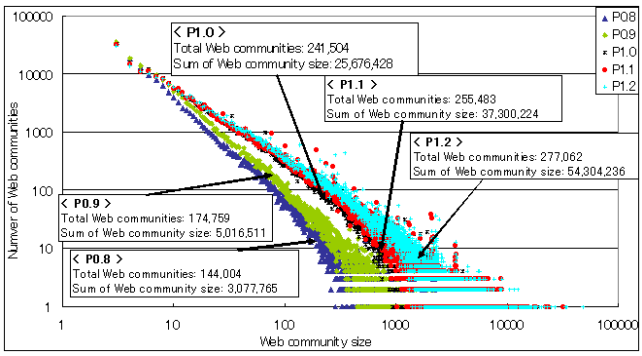

図 1 “距離量”の変化による Web コミュニティサイズの分布

表 1 で，従来の DBG 抽出手法 [10] によって抽出された Web コミュニティ（DBG1, DBG2）と，P0.8，P1.0，P1.2 の比較を行なう. [10] の Web コミュニティ抽出手法は，3. 節で説明した PlusDBG の手法と同様の手順を取り，Web コミュニティ候補の探索ステップのみが異なる．DBG 抽出手法 [10] の Web コミュニティ候補の探索ステップは，以下の通りとする．

（1） $F = \{s\}$ を定義する．

（2） 以下の (a)，(b) を “与えられた繰り返し回数”n 回繰り返す.

（a） $P = \{p | p \notin F, |children(F) \cap children(p)| > 0\}$ となる P を見つける．

（b） $F = F \cup P$.

（3） F を Web コミュニティ候補とする．

DBG1 は “与えられた繰り返し回数”を 1 とし，DBG2 は 2 として，それぞれ，NTCIR-4 データセット [14] から，全ての Web ページをシードページとして，Web コミュニティを抽出する．

表 1 より，DBG2 は，Web コミュニティサイズの合計値は P1.2 よりも大きいが，総 Web コミュニティ数自体は P1.0 よりも少ないことが分かる．つまり，PlusDBG は，従来手法の DBG 抽出に比べて，シードページと関連した Web ページを収集する能力が高いと考えられる．

表 1 Web コミュニティ数と Web コミュニティサイズ

	総 Web コミュニティ数	Web コミュニティサイズの合計値
DBG1	190,293	55,372,697
DBG2	223,791	6,981,184,986
P0.8	143,627	4,958,343
P1.0	342,106	89,552,263
P1.2	360,325	186,690,367

図 2 に，Web コミュニティのサイズを増加させた際の，抽出される Web ページのデータセットに対するカバー率を示す．図 2 に示されるとおり，Web ページのカバー率は，“距離量”に関係なく，Web コミュニティのサイズが 600 を超えるあたりで飽和している．つまり，サイズの大きな Web コミュニティは，サイズの小さな Web コミュニティでは未知の Web ページを発

見しているのではなく，サイズの小さな Web コミュニティを結合していると考えられる．

また，“距離量”の変化に注目すると，P0.8 から P1.0 の間で大幅にカバー率が上昇していることが分かる．しかし，P1.0 から P1.2 の間では，カバー率の増加は確認できない．図 1 の結果を見ると，P1.0 から P1.2 の間で，Web コミュニティのサイズは大幅に増加していることが分かる．つまり，P0.8 から P1.0 では，Web コミュニティは，他のいずれの Web コミュニティにも含まれていない Web ページを含めることで，Web ページのカバー率を向上させていると考えられる．しかし，P1.0 から P1.2 では，Web コミュニティは，他の Web コミュニティとの重複部分を拡大することで，カバー率を向上させずに，Web コミュニティのサイズのみを増加させている．

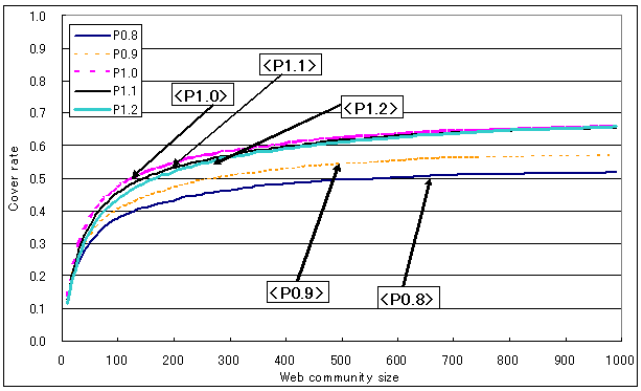

図 2　Web ページのカバー率

図 3 に，各 Web ページが属している Web コミュニティ数を示す．また，比較として，Yahoo! Japan (Yahoo!) と Open Directory Project (ODP) において，各 Web ページが属しているディレクトリ数を示す．図 3 において，x 軸は，各 Web ページが属する Web コミュニティの数を示し，y 軸は Web ページ数を示す．図 3 では，P0.8 から P1.0 の間で，各 Web ページが属する Web コミュニティの数は大幅に増加している．P1.0 から P1.2 では，図 2 において示した通り，カバー率に差が見られないため，分布に大きな変化は見られない．しかし，多くの Web コミュニティに同時に属している Web ページの数が増加しており，少数の Web コミュニティに同時に属している Web ページの数は減少している．つまり，図 2 で推測した通り，P1.0 から P1.2 では，Web コミュニティは，互いの重複部分を拡大することで，Web コミュニティのサイズを拡大していると分かる．そのため，P1.2 では，Web コミュニティ同士が密に重なり合っていると考えられる．

図 4 に，P0.8 と P1.0 における，互いに 30 %以上重複している Web コミュニティのペア数を示す．重複率の調査では，P0.8 と P1.0 のそれぞれにおいて，全ての Web コミュニティのペアに対して重複率を計算する．Web コミュニティ A, B の間の重複率を $Dup(A,B) = \frac{|A \cap B|}{|A \cup B|}$ とする．図 4 より，PlusDBG は，多くの重複した Web コミュニティを持つことが分かる．P1.0 においては，100 %の重複率を持つペアだけで，100 万ペア存在しており，30 %以上の重複率を持つペアの総数は，約 4700

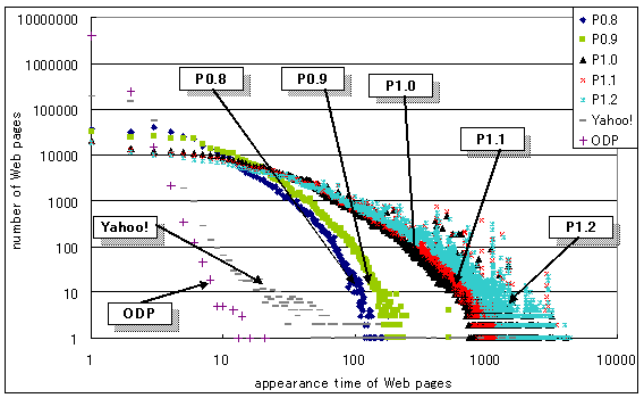

図 3　Web ページが属する Web コミュニティ数の分布

万ペアにのぼる．また，P0.8 と P1.0 の間における，Web コミュニティのペア数の差は，10 倍から 100 倍の差があり，“距離量”の閾値を変化させることで，Web コミュニティ間の重複の度合いが大きく変化することが分かる．

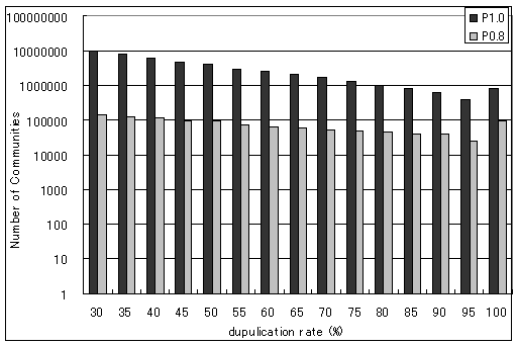

図 4　Web コミュニティの重複率

また，図 3 において，Yahoo! Japan などの人手による Web ディレクトリサービスと，PlusDBG の分布が異なるの原因として，以下の 2 つの場合が考えられる．1 つ目は，PlusDBG がシードページの選別を行なわないために，同じトピックを扱ったコミュニティが重複して抽出されてしまう場合である．また，2 つ目は，複数トピックを持った Web コミュニティが互いに重複した Web ページを含む場合である．これらの Web コミュニティは，Web をトピックによって分類する際のノイズであると考えられる．これらの不要な Web コミュニティを判別して削除することで，Web コミュニティ抽出の精度の向上が期待できる．

4.3　Web ディレクトリサービスとの比較

Web コミュニティの精度を評価するために，PlusDBG によって抽出された Web コミュニティと，Yahoo!Japan ディレクトリとの比較を行なった．

Web ディレクトリサービスとの比較に際して，Web コミュニティの精度を以下に定義し，評価を行なう．まず，2 つの Web ページ p, q が，共通の Yahoo!ディレクトリに出現する場合，$score(p,q) = 1$ を与え，出現しない場合，$score(p,q) = 0$ を与えることにする．

（1） Web コミュニティC 中で，Yahoo!ディレクトリにも存在する Web ページの集合を YC とする．

（2） YC 中のページ r において，r のスコア $Pscore(r) = \frac{\sum_{x \in \{YC-r\}} score(r,x)}{|\{YC-r\}|}$ とする．

（3） C の精度 $P(C) = \sqrt{\frac{\sum_{p \in YC} Pscore(p)}{|YC|}}$

ここで，精度が 0 % の Web コミュニティとは，Web コミュニティ中の全ての Web ページが，Yahoo!ディレクトリ中の別々のカテゴリに出現している状態である．これは，PlusDBG が Yahoo!に存在しないトピックの Web コミュニティも抽出しているためだと考えられる．図 5 に，YP0.8 中での精度 0 %の Web コミュニティの例を示す．図 5 で示される各 URL は，大部分が山口県内の企業や団体の Web ページとなっており，1 つのトピックに関連する Web ページの集合であると考えられる．そこで，3 ページ以上の Web ページが同じカテゴリに属している Web コミュニティのみ，精度評価の対象とする．

山口県(主に下関市)内の企業・団体のページ	
seedpage:	http://bbs.tip.ne.jp/
	http://djmoko.com/
	http://www.c-fm.co.jp/
	http://www.chijinshokan.co.jp/Books/ISBN4-8052-0609-8.htm
	http://www.chs.nihon-u.ac.jp/ch_dpt/shokai/c-zou.html
	http://www.civic.jp/WWWBoard/Messages/44.html
	http://www.keihouan.co.jp/
	http://www.matsuoka.co.jp/
	http://www.mm-inoue.co.jp/
	http://www.nakashima-uni.gr.jp/
	http://www.nishika.co.jp/
	http://www.ogawauni.co.jp/
	http://www.pawnshop.co.jp/
	http://yoo.to/sarah/
	http://www.sanyototal.co.jp/
	http://www.seibikumiai.or.jp/
	http://www.sunelec.co.jp/
	http://www.tip.ne.jp/stca/
	http://www.tip.ne.jp/wvs/
	http://www.ubematerials.co.jp/
	http://www.uni.or.jp/
	http://www.bigss.com/
	http://www.coara.or.jp/%7ecba/mokof/mokofindex.htm
	http://www.hinanet.ne.jp/%7ek-truck/
	http://www.lifecs.com/
	http://www.mai-d.net/manseidou/index.html
	http://www.move-kando.net/
	http://www.navitown.com/
	http://www.sanwa-co.co.jp/
	http://www.tip.ne.jp/ejima/
	http://www.tip.ne.jp/emission/
	http://www.tip.ne.jp/gosho/
	http://www.tip.ne.jp/hare/
	http://www.tip.ne.jp/kajix/
	http://www.tip.ne.jp/katuyama/
	http://www.tip.ne.jp/kayano/
	http://www.tip.ne.jp/light/
	http://www.tip.ne.jp/ohba/
	http://www.tip.ne.jp/otsuya/
	http://www.tip.ne.jp/palace/
	http://www.tip.ne.jp/pusan/
	http://www.tip.ne.jp/rokan/
	http://www.tip.ne.jp/saiban/
	http://www.tip.ne.jp/sdic/
	http://www.tip.ne.jp/shobido/
	http://www.tip.ne.jp/sjc/
	http://www.tip.ne.jp/sts/
	http://www.tip.ne.jp/sumiyosi/
	http://www.tip.ne.jp/sunnet/
	http://www.tip.ne.jp/takefive/
	http://www.tip.ne.jp/yamaken/
	http://www.tip.ne.jp/youme/
	http://www1.ocn.ne.jp/%7ehowfine/
	http://www2.justnet.ne.jp/%7etakedagon/
	http://www5b.biglobe.ne.jp/%7ek-ko/

図 5　精度 0 の Web コミュニティの例（YP0.8）

図 6 に，各 “距離量”の閾値（YP0.8, YP0.9, YP1.0, YP1.1, YP1.2）における，精度に対する Web コミュニティ数の割合を，精度を下限とした場合の累積値で示す．図 6 において，YP0.8

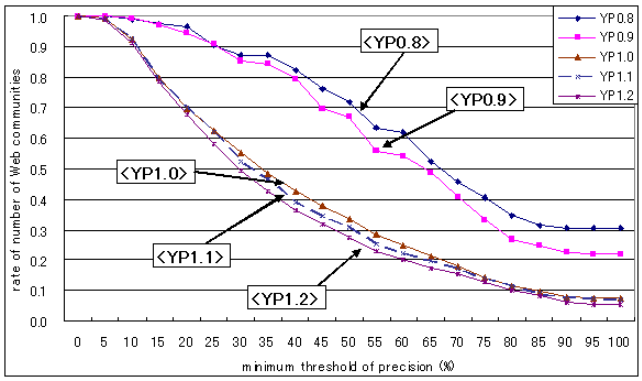

図 6　Web コミュニティの精度ごとの Web コミュニティ数の割合

と YP1.0 の精度を比較すると，YP0.8 のほうが，精度の高い Web コミュニティを高い割合で持っていることが分かる．しかしながら，YP1.0 の精度と YP1.2 の精度とを比べると，ほとんど差が見られない．表 2 に，YP0.8，YP1.0，YP1.2 それぞれの平均精度を示す．表 2 においても，YP1.0 と YP1.2 の精度の間に大きな差は見られない．このことから，YP1.2 では，YP1.0 の Web コミュニティ中に存在する Web ページと，高い関連性を持った Web ページだけが抽出されていると考えることができる．また，YP1.0 と YP0.8 の間で，精度が大きく変化している原因は，図 1, 2 に示される通り，YP1.0 が新しい Web コミュニティや Web ページを数多く抽出しているためだと考えられる．

表 2　Web コミュニティの各 “距離量”における平均精度

	YP0.8	YP0.9	YP1.0	YP1.1	YP1.2
平均精度	0.67	0.63	0.41	0.40	0.37

図 7 に，YP0.8，YP1.0，YP1.2 における，Web コミュニティの精度とサイズの関係を示す．Web コミュニティのサイズが小さい場合，精度は 0 % から 100 % と，幅広く分布している．しかし，Web コミュニティのサイズが 100 前後で，精度の上限が下がり始めており，Web コミュニティのサイズが 1000 前後で，Web コミュニティの精度は完全に 0 % となってしまう．つまり，サイズの大きすぎる Web コミュニティは，抽出する必要性のない Web コミュニティであると考えることができる．

4.4　Web コミュニティ抽出における精度の向上に関する考察

Web コミュニティ解析の結果を基に，不要な Web コミュニティを判別するための要素について考察する．まず，Web コミュニティのサイズによる，不要な Web コミュニティの判別が期待できる．図 7 において，Web コミュニティの精度の上限が，サイズの増加とともに低下しており，サイズが 1000 を超える Web コミュニティでは，精度の上限は完全に 0 になってしまう．また，図 2 から，サイズの大きな Web コミュニティ

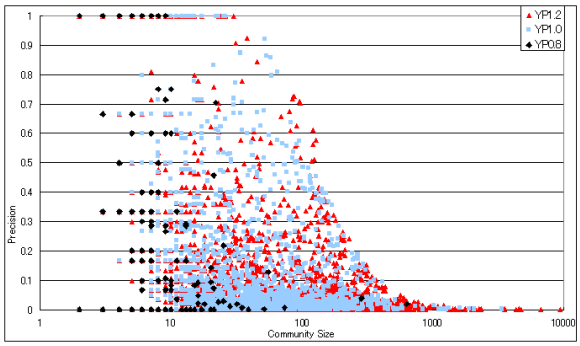

図 7　Web コミュニティの精度とサイズの関係

を削除しても，Web ページのカバー率に大きな変化はないことが分かる．

サイズの大きな Web コミュニティは，図 2 から，小さい Web コミュニティを結合させたものであると予想できる． さらに，この大きな Web コミュニティは，低い精度しか持たないので，同じトピックの Web コミュニティのみを結合しているとは考えにくい．

また，PlusDBG の “距離量”の閾値によって，抽出される Web コミュニティの精度と大きさを細かく調節可能であると分かった．図 6 より，P0.8 の精度が P1.0，P1.2 を上回っていることが分かる．これは，図 1, 2, 3 により示されるとおり，PlusDBG が，P0.8 では，P1.0, P1.2 と異なり，シードページと直接の共参照関係にある Web ページに対しても，“距離量”によってふるいわけを行なっているからである．反対に，P1.0 以降の Web コミュニティでは，総 Web コミュニティ数や抽出される Web ページ数が増加している．これは，P1.0 以降においては，シードページと直接の共参照関係にあれば，関連性の低い Web ページであっても必ず Web コミュニティ候補となることが原因である．図 6 において，P1.0 以降の精度が P0.8 の精度を大幅に下回っているのは，関連性の低い Web ページが Web コミュニティのメンバとして含まれているためだと考えられる．そこで，$Sim(p, F)$ を基にした “距離量”の条件式を改良することで，高い精度を保ったまま，多くの Web コミュニティを抽出することが可能であると考えられる．

また，PlusDBG などの，シードを用いた Web コミュニティ抽出手法が持つ問題点として，WWW 中の Web ページを，可能な限りの Web コミュニティへと分類することを目的とした場合に，シードを選択して指定することができない点があげられる．この問題によって，PlusDBG では，2 種類の不要な Web コミュニティが抽出される．一つ目は，複数のトピックを持った Web ページがシードとなった場合に抽出される，複数トピックを持った Web コミュニティである．二つ目は，互いに関連性の高い Web ページから抽出される，重複した Web コミュニティである．これは，図 4 における，高い重複率のコミュニティペアが当てはまる．Web コミュニティ抽出では，これらの Web ページを予め知ることができないため，不要な Web コミュニティが抽出される原因となっている．

表 3 に，“距離量”の閾値ごとの Web コミュニティの性質をまとめる．表 3 における，飽和カバー率とは，Web コミュニティのサイズが 1000 の時点における，Web ページのカバー率である．また，最大重複数とは，Web ページが属する Web コミュニティ数の中で最大の値とする．表 3 で示すとおり，P0.8 の Web コミュニティは，精度が高く，Web コミュニティの粒度が小さい．加えて，P0.8Web コミュニティは，それぞれの Web コミュニティの重複の度合いが小さく，カバー率も低い．対して，P1.0Web コミュニティは，P0.8 に比べて重複の度合いが大きく，カバー率も大きい．これは，図 1 において考察したように，P1.0Web コミュニティが，Web コミュニティを抽出しにくいシードページからでも，Web コミュニティを抽出可能なためである．逆に，P0.8Web コミュニティは，Web コミュニティの抽出に適さないシードページからは Web コミュニティを抽出せず，精度の高い Web コミュニティの抽出を実現している．つまり，P0.8 では，Web コミュニティの抽出段階で，網羅的に入力されたシードページの選別が行なわれていると言える．

表 3　“距離量”の閾値ごとの Web コミュニティの性質

	P0.8	P1.0	P1.2
精度	0.67	0.41	0.37
総 Web コミュニティ数	144,004	241,504	277,062
コミュニティサイズの合計	307 万	2568 万	5430 万
飽和カバー率	0.52	0.66	0.66
最大重複数	178	4,027	4,319

5. 不要な Web コミュニティの自動判別

本節では，4 節での考察に基づき，不要な Web コミュニティを自動的に判別する手法について考える．4.4 節にて考察したとおり，PlusDBG [8] は，距離量の閾値を小さくとることで，不要な Web コミュニティを自動的に取り除いていると考えられる．そこで本章では，以下に示す手順によって，PlusDBG により Web コミュニティを抽出し，評価を行う．

（1） データセット中の全 Web ページを，シードセット S とする．

（2） “距離量”の閾値を 0.8 として，S 中の全 Web ページから，PlusDBG を用いて Web コミュニティを抽出する．

（3） 手順 2 にて Web コミュニティを抽出できなかったシードページを，シードセット S から取り除く．

（4） “距離量”の閾値を 1.0 として，S 中の全 Web ページから，PlusDBG を用いて Web コミュニティを抽出する．

5.1 提案モデルの評価

提案モデル (P0.8-1.0) によって抽出された Web コミュニティを評価するために，P0.8-1.0 と，“距離量”の閾値をそれぞれ 0.8 (P0.8)，1.0 (P1.0) にした場合に PlusDBG で抽出された Web コミュニティとの比較を行う．Web コミュニティは，4 節と同様に，NTCIR-4 データセット [14] より抽出した．

図 8 に，P0.8，P1.0，P0.8-1.0 における，精度に対する Web コミュニティ数の割合を，精度を下限とした場合の累積値で示す．精度評価においては，4 節と同様の評価手法を用いて評価

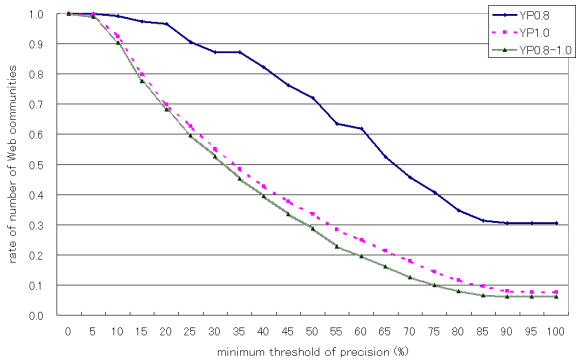

図 8　Web コミュニティの精度

を行った．図 8 から分かるとおり，P0.8-1.0 の精度は，P1.0 の精度とほぼ同等であり，精度の向上と言う期待した結果は得られなかった．また，図 8 では，P0.8 に比べて P0.8-1.0 の精度が低下している．この点から，PlusDBG では，“距離量”の閾値によって，Web コミュニティの持つトピックとは異なる Web ページを取り除いていることが分かる．

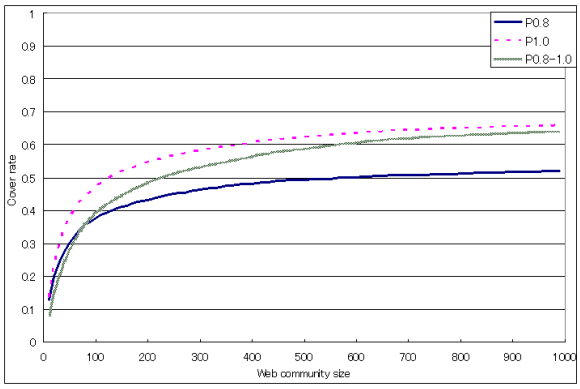

図 9　Web ページのカバー率

図 9 に，P0.8，P1.0，P0.8-1.0 における，Web コミュニティのサイズを増加させた際の，Web コミュニティによるデータセットに対するカバー率を示す．図 9 より，P0.8-1.0 では，P1.0 から Web コミュニティを削減したにもかかわらず，P1.0 と同等のカバー率を達成していることが分かる．また，図 8 において，P0.8-1.0 の Web コミュニティは P1.0 とほぼ等しい精度を持っていることが分かる．従って，P0.8-1.0 の Web コミュニティは，無差別に Web ページを抽出することで，カバー率を増加させているわけではないことが分かる．つまり，P0.8-1.0 では，P1.0 において内容が重複している Web コミュニティを削減していると考えることができる．

4 節において，Web コミュニティ抽出が，データセット全体から可能な限りの Web コミュニティを抽出する場合に，シードページを選択して指定することができない問題を指摘した．また，その結果として，異なるシードページから重複した Web コミュニティが抽出される現象が発生する．P0.8-1.0 の Web コミュニティでは，このような，異なるシードページから抽出される重複した Web コミュニティを，Web コミュニティごとの比較を行わずに，自動的に発見し削除することが可能であると考えられる．つまり，P0.8-1.0 は，Web ページのカバー率を保ったまま，不要な Web コミュニティを削除することで，抽出される Web コミュニティ数の削減に成功していると考えられる．

6. おわりに

本論文では，Web コミュニティ抽出の精度の向上を目的として，PlusDBG [8] によって抽出された Web コミュニティデータの解析を行なった．その結果，Web コミュニティのサイズや，“距離量”のパラメータによって，Web コミュニティの性質が異なることが分かった．さらに，Web コミュニティ抽出によって抽出された Web コミュニティを全体的に眺めることで，Web の意味的な構造を推測することができた．また，PlusDBG における ”距離量”の閾値によって，重複した Web コミュニティを自動的に削除し，Web ページのカバー率を保ったまま，抽出される Web コミュニティ数を削減することができた．

今後は，本論文において考察した Web コミュニティの性質を，Web コミュニティ抽出の精度向上に利用していきたい．また，4.4 において述べた，問題のあるシードページが不要な Web コミュニティを抽出してしまう問題についても，解決していきたい．

謝辞　本研究の一部は，文科省「21 世紀 COE プロダクティブ ICT アカデミア」と「e-Society 基盤ソフトウェアの総合開発」によるものである．

文　献

[1] Netcraft, http://news.netcraft.com/
[2] S. Lawrence and C. L. Giles, “Accessibility of information on the web,” Nature, 400, pp.107-109, 1999.
[3] A. Gulli and A. Signorini, “The Indexable Web is More than 11.5 Billion Pages,” In Proc. of 14th Int. WWW Conf., 2005.
[4] Yahoo!, http://www.yahoo.com/
[5] Open Directory Project, http://dmoz.org/
[6] S. R. Kumar, P. Raphavan, S. Rajagopalan and A. Tomkins, “Trawling the Web for emerging cyber communities,” In Proc. of 8th Int. WWW Conf., 1999.
[7] G. Flake, S. Lawrence and C. Giles, “Efficient identification of Web communities,” In Proc. of 6th ACM SIGKDD Conf., 2000.
[8] N. Saida, A. Umezawa and H. Yamana, “PlusDBG: A Web Community Extraction Scheme Improving both Precision and Pseudo-Recall,” In Proc. of 7th APWeb Conf., 2005.
[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, “Graph structure in the web,” In Proc. of 9th Int. WWW Conf., 2000.
[10] P. K. Reddy and M. Kitsuregawa, “An approach to relate the Web communities through bipartite graphs,” In Proc. of 2nd Int. Conf. on Web Information Systems Engineering, 2001.
[11] 村田剛志, “参照の共起性に基づく Web コミュニティーの発見,” 人工知能学会論文誌, Vol.16, No.3, pp.316-323, 2001.
[12] L. R. Ford Jr and D. R. Fulkerson, “Maximal flow through a network,” Canadian Journal of Mathematics, 8, pp.399-404, 1956.
[13] Hidehiko Ino, Mineichi Kudo and Atsuyoshi Nakamura, “Partitioning of Web Graphs by Community Topology,” In Proc. of 14th Int. WWW Conf., 2005.
[14] NTCIR Project, http://research.nii.ac.jp/ntcir/index-en.html