DEWS2003 2-P-05

日本のウェブアーカイブにおけるウェブコミュニティ発展過程の詳細分析

豊田　正史† 　喜連川　優†

† 東京大学生産技術研究所
〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †{toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　ハイパーリンクの構造解析を用いて同じトピックを持つウェブページの集合を抽出する手法は現在までに多数提案されており，この集合はウェブコミュニティと呼ばれている．本論文では，1999 年から 2002 年の間に 4 回収集した日本のウェブアーカイブからコミュニティの発展過程を詳細に分析した結果を示す．我々の手法はまず各アーカイブから主要なコミュニティをすべて抽出し，アーカイブ間でコミュニティの比較を行うことで時系列的変化を調査する．発展の度合を理解するため，成長率，新規率などのメトリックスを導入しており，これらの値の分布や時系列的変化を示す．分析の結果，すべての発展過程がべき乗則に従って起こり，分裂と合併の間，および成長と縮小の間に，対称性があることが分かった．

キーワード　ウェブ，リンク解析，ウェブコミュニティ，発展過程

Analyzing Evolution of Web Communities using a Series of Japanese Web Archives

Masashi TOYODA† and Masaru KITSUREGAWA†

† Institute of Industrial Science, University of Tokyo
4–6–1 Komaba Meguro-ku, Tokyo, 153–8505 Japan
E-mail: †{toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　In this paper, we analyze evolution of web communities using a series of Japanese web archives. A web community is a set of web pages created by individuals or associations with a common interest on a topic. So far various techniques have been developed to extract web communities by link analysis. We examined evolution of web communities by comparing Japanese web archives crawled four times from 1999 to 2002. Several metrics are introduced for measuring the degree of web community evolution, such as growth rate, novelty, and stability. Statistics of these archives and the evolution metrics are shown, and the global behavior of evolution is described. We found that most of changes in communities follow the power-law, and the changes are symmetric between split and merge, and between growth and shrinkage.

Key words　Web, link analysis, web community, evolution

1. はじめに

近年，ウェブは急激に成長を続けてきた．日々多くのページが作成および削除されることで，ウェブの構造も変化し続けている．一方，ストレージの大容量化および低価格化に伴い，定期的に収集したウェブアーカイブをすべて保管することが可能となってきている．既に，指定された URL の内容を過去にさかのぼって見られるサービス [14] も始まっているが，まだ単独のページの変化しか見ることはできない．こうした背景のなかでウェブの発展過程を観測し，重要な情報を発見することが重要な課題となってきている．

本論文では，定期的に収集したウェブアーカイブからウェブコミュニティの発展過程を抽出する手法を提案する．ここで言うウェブコミュニティとは，同じトピックに関心をもつ人々や組織によって作成されたウェブページの集合を指す．ウェブコミュニティの例として，同じ業種に属する会社のホームページの集合や，ある野球チームを応援するホームページの集合などが挙げられる．これまでに，WWW をウェブページとその間に張られたハイパーリンクによるグラフと見なし，グラフ構造を解析することで，ウェブコミュニティを抽出する様々な手法が

提案されてきた [2], [4], [5], [7]〜[12] ．しかし，抽出されたコミュニティの発展過程を実際に調査した研究はほとんど発表されていない．

ウェブコミュニティはあるトピックを表すため，新しいトピックがいつ発生して，どのように発展したかを，コミュニティを単位として理解することができる．例えば，2001 年 9 月 11 日のアメリカでのテロ事件について，関連するページがどの程度作られてきたか，といった事例が挙げられる．このような情報は，次のような状況で有用である．(1) ウェブにおけるトピックの履歴に関する質問に答える．(2) ある分野に関連する新たなコミュニティの発生を観察する．(3) 実社会における活動に対応するウェブ上の活動を調査する．

上記の情報を抽出する方法を探るために，我々は，1999 年から 2002 年の間，4 回に渡って収集したウェブアーカイブを比較することで，ウェブコミュニティの発展過程を分析した．分析の手順としては，まず各ウェブのスナップショットから，主要なすべてのコミュニティとそれらの間の関連度を抽出する．これには本研究に先立って発表したウェブコミュニティチャート [13] の成果を利用している．その上で，各コミュニティの時間変化を調査する．この際，コミュニティの成長率，新規率，安定率など，興味ある発展過程の抽出に有用なメトリックスを導入した．これらのメトリックスを用いると，最も成長したコミュニティや，最も新しいコミュニティなどを抽出することが可能になる．

本論文では，まず上記 4 回分のウェブアーカイブおよびアーカイブから抽出したウェブコミュニティチャートの詳細な分析結果を示す．その上で，コミュニティ発展過程における全体的な挙動を示し，発展の詳細をメトリックスの値の分布やその変化で示す．また，我々が開発したウェブコミュニティの発展過程ビューアを用いて，幾つかの発展過程の例を示す．

本論文の構成は以下の通りである．第 2. 節では，ウェブコミュニティチャートの概要を述べる．第 3. 節では，コミュニティの発展過程の詳細について解説し，発展のメトリックスを導入する．第 4. 節では，実験に用いたアーカイブと，抽出したコミュニティチャートについて説明する．第 5. 節では，発展過程ビューアとそれを用いた発展過程抽出の例を示し，第 6. 節で，発展過程の詳細分析の結果を示す．第 7. 節で，まとめと今後の課題を述べる．

2. ウェブコミュニティチャート

本節では，ウェブコミュニティチャート [13] の概要について説明する．コミュニティチャートは，ウェブコミュニティをノードとし，関連するコミュニティの間に重み付のエッジを張ったグラフである．エッジの重みは，コミュニティの関連度を表す．図 1 に，我々が作成したコミュニティチャートの一部を示す．中央に大手コンピュータメーカのコミュニティがあり，その周りに関連するコミュニティとして，ソフトウェア，周辺機器，デジタルカメラなど関連業種の会社のコミュニティが抽出されている．以下に，コミュニティチャートを作成する方法を簡単に述べる．詳細については [13] を参照されたい．

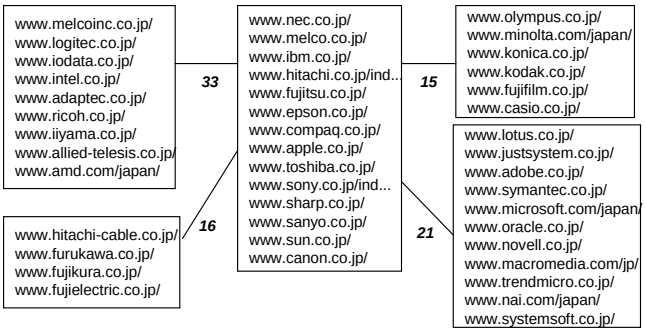

図 1　コミュニティチャートの一部
Fig. 1　A part of our web community chart

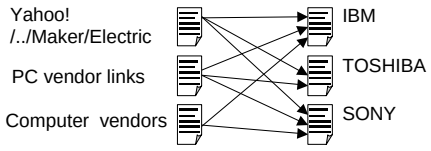

図 2　Authority および hub からなる典型的なグラフ構造
Fig. 2　Typical graph structure of hubs and authorities

コミュニティチャートの作成のために，我々は関連ページアルゴリズム [6], [13] を利用している．関連ページアルゴリズムは，(1)1 つのシードページを入力として与えると，(2) シードページの近傍のウェブグラフから，良い authority ページおよび良い hub ページを抽出し，(3) 上位の authority ページを関連ページとして出力するアルゴリズムである．ここで良い authority とは，多くの良い hub からハイパーリンクを張られている著名なページを表す．良い hub とは，リンク集およびブックマークなど，多くの良い authority へハイパーリンクを張っているページを表す．この循環した定義により，密に結合した hub と authority が抽出され，それらがよく関連したページを表すことが [6], [13] で示されている．

図 2 に典型的な authority と hub のグラフ構造の 1 例を示す．このグラフの右側には，IBM，TOSHIBA，および SONY といった大手のコンピュータ関連会社が authority としてあり，それらに密にリンクを張っているリンク集が左側に hub としてある．このようなグラフ構造は，ウェブ上に多々見られるものである．関連ページアルゴリズムは，図 2 のように密に結合された authority と hub を抽出するものである．IBM，TOSHIBA，SONY のどれかひとつをシードとして与えると，これらの会社のリストが結果として出力されることになる．

我々のチャート作成アルゴリズムは，分類したいシードページの集合を入力として受取り，チャートを結果として出力する．シードページとしてはウェブ上で著名なページを抽出して使用する．判断基準は，外部のサーバから IN 本以上リンクが来ていることとした．IN は，チャートのサイズを決めるパラメタとなる．

シードセットを受け取ると，各シードページについて別々に，上記の関連ページアルゴリズムを適用し，各シードが他のシードをどのように関連ページとして導出するかを調べる．この際，

関連ページアルゴリズムの結果のうち上位 N 個を使用する．N はコミュニティの粒度を決めるパラメタとなる．我々は，シード a がシード b を関連ページとして導出し，かつその逆も成り立つという対称関係に注目し，この関係で密に結合されたシード同士は，しばしば同じレベルのトピックを共有することを [13] で示した．これに従って，対称関係で密に結合されたシード同士をコミュニティとして抽出する．さらに 2 つのコミュニティのメンバ間に導出関係がある場合には，その間にエッジを張ることでコミュニティのグラフ (チャート) を作成する．

3. コミュニティの発展過程

本節では，ウェブコミュニティの発展過程について説明し，成長率，新規率など，発展の度合いを測るメトリックスを導入する．まず，本節で用いる記号を以下に示す．

$t_1, t_2, ..., t_n$: 各ウェブアーカイブが収集された時間．現在は 1 月を単位時間として使用している．

$W(t_k)$: 時間 t_k に収集されたウェブアーカイブ．

$C(t_k)$: $W(t_k)$ から作成されたウェブコミュニティチャート．

$c(t_k), d(t_k), e(t_k), ...$: $C(t_k)$ に含まれるコミュニティ．

3.1 発展の種類

コミュニティの発展過程は，定期的に収集されたウェブのスナップショット ($W(t_1), W(t_2), ..., W(t_n)$) を基に以下のように観察する．(1) すべてのスナップショットについて，ウェブコミュニティチャート ($C(t_1), C(t_2), ..., C(t_n)$) を作成する．(2) 隣接する時間におけるコミュニティチャートの差を比較調査する．

時間 t_k におけるコミュニティチャート $C(t_k)$ の変化は，前向き，後ろ向きの 2 通りに調べられる．簡単のため，ここでは後ろ向きの調べ方のみを述べる．すなわち，$C(t_k)$ と $C(t_{k-1})$ とを比較して，t_{k-1} から t_k までの間にコミュニティがどう発展したかを調べる．前向きの調べ方も，同様に行うことが可能である．以下では，コミュニティの発展過程の種類を列記し，発展のメトリックスを導入する．

発生: コミュニティ $c(t_k)$ が，$C(t_{k-1})$ におけるどのコミュニティとも URL を共有していないとき，$c(t_k)$ は，$C(t_k)$ において発生したとみなす．$c(t_k)$ 内のいくつかの URL は $W(t_{k-1})$ に存在し，コミュニティに含まれる程の結合度を持っていない可能性があることに注意されたい．

解散: コミュニティ $c(t_{k-1})$ が，$C(t_k)$ におけるどのコミュニティともページを共有していないとき，$c(t_{k-1})$ は，解散したとみなす．$c(t_{k-1})$ 内のいくつかの URL は，結合度を失ったものの $W(t_k)$ に残っている可能性があることに注意されたい．

成長および縮小: $C(t_{k-1})$ 中の $c(t_{k-1})$ が，$C(t_k)$ 中のただひとつの $c(t_k)$ と URL を共有しており，かつその逆も成り立つときは，成長か縮小の 2 通りの変化しか起こり得ない．新たな URL が出現すれば成長し，URL が消失すれば縮小する．出現した URL 数が消失した URL 数より多ければコミュニティは最終的には成長したことになり，その逆の場合，縮小したことになる．

分裂: コミュニティ $c(t_{k-1})$ が，$C(t_k)$ における複数のコミュニティと URL を共有するとき，コミュニティは複数のコミュニティに分裂したとみなす．コミュニティは分裂する前後に成長および縮小する可能性がある．また，分裂したコミュニティが，別なコミュニティと合併することもありうる．

合併: コミュニティ $c(t_k)$ が，$C(t_{k-1})$ における複数のコミュニティと URL を共有するとき，コミュニティは合併したとみなす．コミュニティは合併の前後に成長および縮小する可能性がある．

3.2 発展のメトリックス

発展のメトリックスは，コミュニティ $c(t_k)$ がどの程度発展してきたかを，測るものである．例えば，コミュニティがどの程度急速に発生したか，どの程度成長したかなどが計測できる．このメトリックスを用いると最も急速に発生したコミュニティや，最も成長したコミュニティなどを検索することが可能になる．

$c(t_k)$ の変化を測定するためには，まず $c(t_k)$ に対応するコミュニティ $c(t_{k-1})$ を $C(t_{k-1})$ において定めなくてはならない．時間 t_{k-1} における $c(t_k)$ の**対応コミュニティ**は，$C(t_{k-1})$ においてもっとも多くの URL を $c(t_k)$ と共有するコミュニティ $c(t_{k-1})$ と定義する．ただし，同数の URL を共有するコミュニティが複数あった場合，一番 URL 数の多いコミュニティを選択する．

逆に，$c(t_{k-1})$ の時間 t_k における対応コミュニティ $c'(t_k)$ を求めることも出来る．この $c'(t_k)$ が $c(t_k)$ と一致するとき，$(c(t_{k-1}), c'(t_k))$ のペアを**本流**と呼ぶ．一致しない場合には**支流**と呼ぶ．本流は，定義にしたがって時間軸方向に伸ばして行くことで，コミュニティの列に拡張できる．本流に含まれるコミュニティは，自己を保っていると考えられ，それが持つ話題の周辺の変化を調べる際の良い出発点となり得る．

発展のメトリックスは，$c(t_k)$ と $c(t_{k-1})$ の差分によって定義される．メトリックスの定義には，以下の記号を用いる．

$N(c(t_k))$: $c(t_k)$ に含まれるページ数

$N_{sh}(c(t_{k-1}), c(t_k))$: $c(t_{k-1})$ と $c(t_k)$ に共有されるページの数

$N_{dis}(c(t_{k-1}))$: $c(t_{k-1})$ から消失したページの数

$N_{sp}(c(t_{k-1}), c(t_k))$: $c(t_{k-1})$ から $c(t_k)$ 以外へ分裂したページの数

$N_{ap}(c(t_k))$: $c(t_k)$ に出現したページの数

$N_{mg}(c(t_{t-1}), c(t_k))$: $c(t_{k-1})$ 以外のコミュニティから $c(t_k)$ へ合併したページの数

ここで，発展のメトリックスは以下のように定義される．以下のメトリックスは，アーカイブの取得時期が不定期でも使用できるように，単位時間当たりに増加したり減少したりした URL 数を用いている．

成長率: 単位時間当たりの URL の増加数を表す．コミュニティを成長率でソートすることで最も成長または縮小したコミュニティを検索できる．成長率は以下のように定義される．もし $c(t_{k-1})$ が存在しない場合は，$N(c(t_{k-1}))$ として 0 を用いることに注意されたい．

$$R_{grow}(c(t_{k-1}), c(t_k)) = \frac{N(c(t_k)) - N(c(t_{k-1}))}{t_k - t_{k-1}}$$

安定率: コミュニティ内の URL がどの程度安定して保たれて

いるかを示す．単位時間当たりに，消滅，出現，合併，分裂した URL の総数で表される．URL の入れ替わりがなければ安定率は 0 となり，入れ替わりが激しい程値は大きくなる．成長率が 0 であっても，安定率が 0 とはならない場合があることに注意されたい．これは，URL が入れ替わっている可能性があるからである．安定したコミュニティは，周辺のコミュニティの変化を調べるための良い出発点となる．安定率は以下のように定義される．

$$R_{stability}(c(t_{k-1}), c(t_k)) =$$

$$\frac{N(c(t_{k-1})) + N(c(t_k)) - 2N_{sh}(c(t_{k-1}), c(t_k))}{t_k - t_{k-1}}$$

新規率: 単位時間当たりに，コミュニティに新たに出現した URL 数を表す．新規率でコミュニティをソートすることで新たに発生したコミュニティを検索できる．新規率は以下のように定義される．

$$R_{novelty}(c(t_{k-1}), c(t_k)) = \frac{N_{ap}(c(t_k))}{t_k - t_{k-1}}$$

消失率: 単位時間当たりに，コミュニティから消失した URL 数を表す．以下のように定義される．

$$R_{disappear}(c(t_{k-1}), c(t_k)) = \frac{N_{dis}(c(t_{k-1}))}{t_k - t_{k-1}}$$

分裂率: 分裂して他のコミュニティへ移動した URL 数を単位時間当たりにならしたもの．大きく分裂したコミュニティを検索できる．分裂率は以下のように定義される．

$$R_{split}(c(t_{k-1}), c(t_k)) = \frac{N_{sp}(c(t_{k-1}), c(t_k))}{t_k - t_{k-1}}$$

合併率: 他のコミュニティから合併により吸収した URL 数を単位時間当たりにならしたもの．主に合併により成長したコミュニティを検索するのに利用できる．合併率は以下のように定義される．

$$R_{merge}(c(t_{k-1}), c(t_k)) = \frac{N_{mg}(c(t_{k-1}), c(t_k))}{t_k - t_{k-1}}$$

上記のメトリックスを組み合わせることで，以下のようにもう少し複雑な発展過程を表すこともできる．ここでは，成長に関してのみ例を挙げるが，縮小に関しても同様に組み合わせることができる．

安定成長: コミュニティは，正の成長率を持ち，消失率および分裂率が小さいとき，安定成長している．

出現による安定成長: コミュニティが安定成長しており，かつ新規率が高い場合，コミュニティは主に新規に出現したページにより安定成長していることになる．

合併による安定成長: 同様に，コミュニティが安定成長しており，かつ合併率が高い場合，コミュニティは主に合併により安定成長していることになる．

また，単位時間当よりも長い期間に渡ってのメトリックスは，コミュニティの本流を利用して計算することが出来る．例えば，$(c(t_i), c(t_{i+1}), ..., c(t_j))$ という本流があったとき，この新規率は以下のように計算する．

$$R_{novelty}(c(t_i), c(t_j)) = \frac{\sum_{k=i}^{j} N_{ap}(c(t_k))}{t_j - t_i}$$

他のメトリックスについても同様に計算が可能である．

Year	Period	Crawled pages	Total URLs	Links
1999	Jul. to Aug.	17M	34M	120M
2000	Jun. to Aug.	17M	32M	112M
2001	Early Oct.	40M	76M	331M
2002	Early Feb.	45M	84M	375M

表 1　ウェブアーカイブの詳細
Table 1　Details of our web archives

4. ウェブアーカイブおよびウェブコミュニティチャートの詳細

4.1 ウェブアーカイブとウェブグラフ

実験には，我々が 1999 年から 2002 年の間，4 回に渡って収集した日本のウェブアーカイブを使用した．この節では，これらのアーカイブの変化の詳細を示す．各アーカイブは幅優先探索でページを収集するウェブクローラを用いて，jp ドメイン内のページを大規模に収集したものである．表 1 にその詳細を示す．1999 年と 2000 年には同じクローラを使用して約 1700 万ページずつを収集した．2001 年，2002 年にはクローラを大幅に改良し，4000 万以上のページを収集した．

収集した各アーカイブから，URL とリンクからなるウェブグラフを抽出し，リンク解析に使用するデータベースを作成した．このウェブグラフにはアーカイブ内のページの URL のみではなく，それらのページからリンクされているアーカイブの外側の URL も含まれる．結果として com や edu ドメインなどの URL もグラフに含まれることになる．表 1 には，グラフに含まれる URL の総数とリンクの総数も示されている．

リンク解析を効率的に行うためこれらのウェブグラフは，指定された URL から隣接 URL を検索できるメインメモリデータベースとして実装した．実装方式は，connectivity server [1] と同様である．全てのシステムは，Sun Enterprise Server 6500 で実現されており，現在の実装では，2002 年のアーカイブからグラフデータベースを作成するには約 1 日が必要である．

これら 4 つのグラフを比較した結果，日本のウェブにおけるグラフ構造は非常に大きく変化していることが分かった．約 1 年間でおよそ半数の URL が消滅し，それ以上の数の URL が作成されている (消えた半数の中には場所を移動した URL も含む)．図 3 に，各グラフが他のグラフと URL を共有している数を計算して，URL の推移を示した．各棒は，URL の総数を表しており，URL が存続した期間に従ってブロックに分割されている(注1)．同じ色のブロック同士は，同じ URL を含んでおり，補助線で結ばれているため，幾つの URL が，何時出現し，どのくらいの期間存続したかを見ることが出来る．1999 と 2000

(注1)：図 3 では，断続的に出現する数百万の URL を省いて示している．例えば，1999 年と 2001 年にしか現れない URL は除かれている．

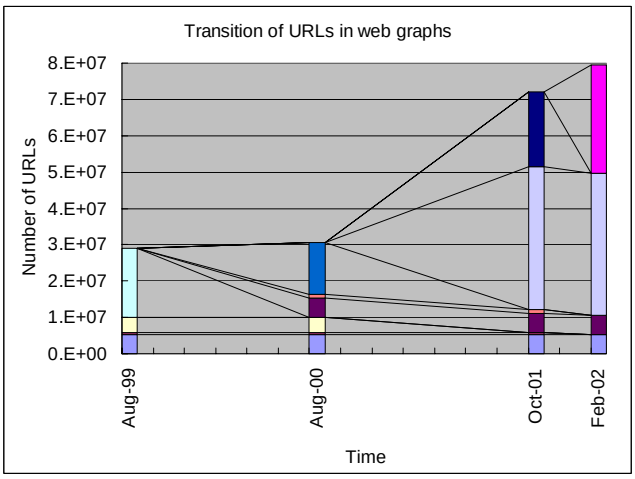

図 3　ウェブグラフにおける URL の推移
Fig. 3　Transition of URLs in web graphs

年のグラフからは，およそ 60%の URL が消滅しており，2001 年 10 月から 2002 年 2 月の 4ヵ月間でおよそ 30%の URL が消滅している．全てのグラフに現れる URL は 5 百万しか存在しない．

また我々は，これらウェブグラフについて，1 つの URL に対してリンクを張っている URL の数 (インリンク数) の分布を調べた．過去の研究 [3], [10], [12] では，i 個のインリンク数を持つ URL が出現する確率の分布は，べき乗則に従うことが示されている．すなわち，正の整数 i の出現確率は，$1/i^k$ に比例する. [3], [10], [12] では，大規模なウェブグラフにおいて，べき指数 k は約 2.1 であることが示されている．我々のウェブグラフにおいては，すべてのグラフについて k は 2.2 の周辺の値を取った．これは過去の研究の結果とほぼ一致している．

4.2　ウェブコミュニティチャート

4.1 節で示した 4 つのウェブグラフそれぞれからウェブコミュニティチャートを作成した．チャートを同じ条件で比較するため，2. で示したチャート作成アルゴリズムにおけるパラメタの値を固定した．シード URL を選択する際の IN としては 3 を使用した．すなわち，異なるサーバからのリンク数が 3 以上の URL をシードとして使用した．リンク数の分布はべき乗則に従うため，これ以上大きな値を取るとシードの数は激減し，小さな値を取るとシードの数が激増する．今回はチャート作成が 1 日以内で終る範囲で IN を決定した．また，関連ページアルゴリズムの上位 N 個を使用する，というパラメタ N については 10 を使用した．これは関連ページアルゴリズムが上位 10 個で良い精度を示しているためである (詳しくは [13] を参照されたい)．表 2 に，各グラフから抽出されたシード URL の総数，およびコミュニティの総数を示す．

5.　発展過程の例

本節では，コミュニティの発展過程の例を，我々が開発した発展過程ビューアを用いて示す．このビューアは，与えられたキーワードや URL によるコミュニティの検索，指定したコミュニティの発展過程の表示，および発展のメトリックスを用いた

Year	Period	Seeds	Communities
1999	Jul. to Aug.	657K	79K
2000	Jun. to Aug.	737K	88K
2001	Early Oct.	1404K	156K
2002	Early Feb.	1511K	170K

表 2　シード URL とコミュニティの総数
Table 2　The number of seeds and communities

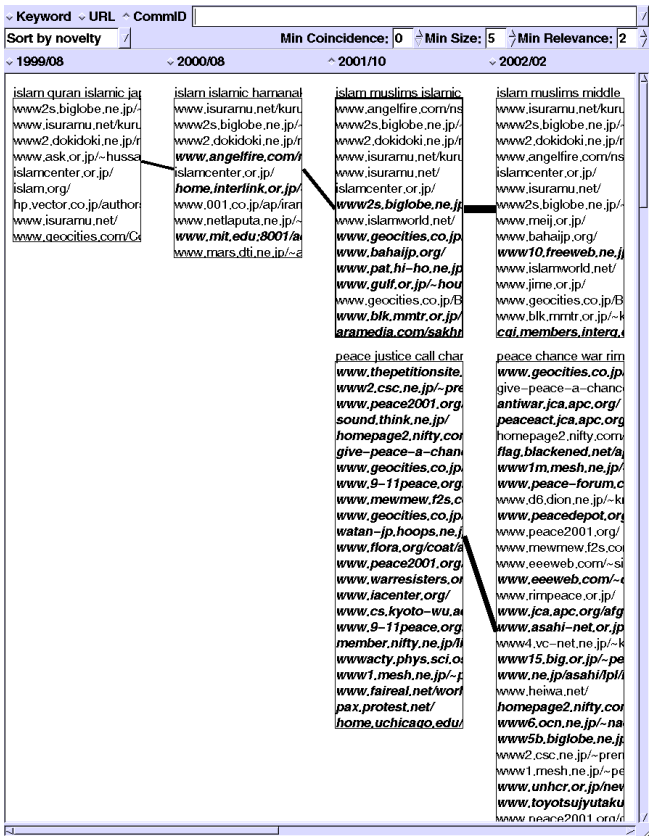

図 4　イスラム関連コミュニティの周辺に発生したコミュニティ
Fig. 4　An emerged community around an Islam information community

コミュニティのソートといった機能を提供し，柔軟な発展過程の閲覧を可能にしている．詳細については [15] を参照されたい．

図 4 は，安定したコミュニティを起点に，その周辺のコミュニティの発生を見たものである．この例では，2001 年 9 月 11 日に起きたアメリカでのテロ事件の後，イスラム教関係のコミュニティの周辺に発生したコミュニティを調べた．まず 2001 年においてイスラム教に関するコミュニティをキーワードにより検索し，安定したコミュニティを得た．このコミュニティの本流が図 4 の上部に，左から右へ時間順に表示されている．コミュニティは URL のリストとして表示され，対応関係を表す線が横方向に引かれている．線の太さは共有する URL 数の多さを表している．また，新しく現れた URL は太字で表されており，このイスラム教のコミュニティが安定して成長していることが分かる．

次に，このイスラム教コミュニティの周辺にある (チャートにおいてエッジが張られている) コミュニティで 2001 年 10 月に発生したものを抽出した．これは新規率で，周辺のコミュニ

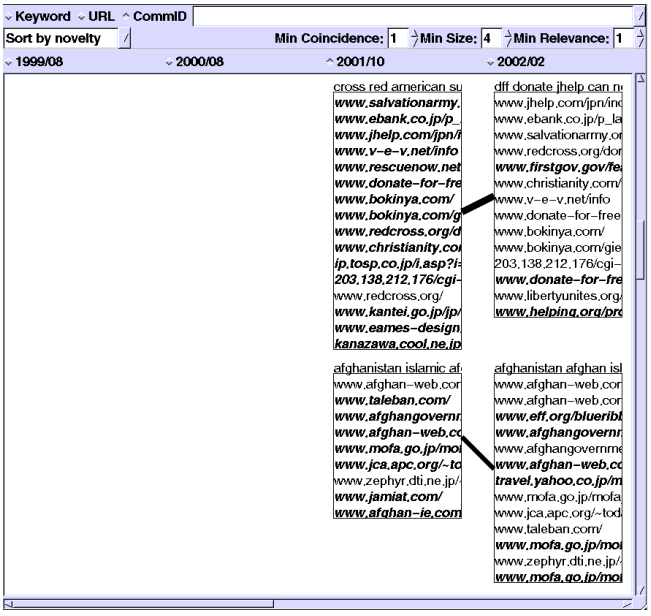

図 5　平和活動コミュニティ周辺に発生したコミュニティ

Fig. 5　Emerged communities around a pacifist community

ティをソートすると得られる．図 4 の下部に，一番新規率の高かったコミュニティが表示されている．このコミュニティは “www.peace2001.org” や “www.9-11peace.org” といった URL を含むことから平和活動に関するコミュニティであることが分かる．

この平和活動コミュニティの周辺には、他にもさまざまな種類のコミュニティが発生している。図 5 には、平和活動コミュニティに関連のある (チャートにおいて結合されている) コミュニティで、新規に発生したものを 2 つ示している。最初のコミュニティは、アフガニスタンに対する義援金の募集ページの集合である。“www.donate-for-free.com” などの URL が含まれている。2 つめのコミュニティは、アフガニスタンやタリバンに関するページの集合である。これらの例から、テロ事件の直後から事件に関する様々な種類のページに人々の関心が集って hub ページが多数作成され，急速に関連するコミュニティ発生したことが分かる．

6.　ウェブコミュニティ発展過程の詳細分析

本節では，ウェブコミュニティの全体的な発展過程を詳細に分析する．

6.1　サイズの分布

コミュニティのサイズ (含まれる URL の個数) の分布は，べき乗則に従い，べき指数は時期によってほとんど変化しないことが分かった．図 6 に，コミュニティのサイズと，そのサイズのコミュニティの個数を両対数グラフを示す．4 つのコミュニティチャート全てがべき乗則に従っており，べき指数は 2.9 から 3.0 の間であった．

6.2　発展過程の種類

コミュニティのサイズの分布は安定しているが，コミュニティ内部の構造には，時期による変化が多々見られる．図 7 は，何個のコミュニティが t_{k-1} から t_k の間にどのような種類の変化

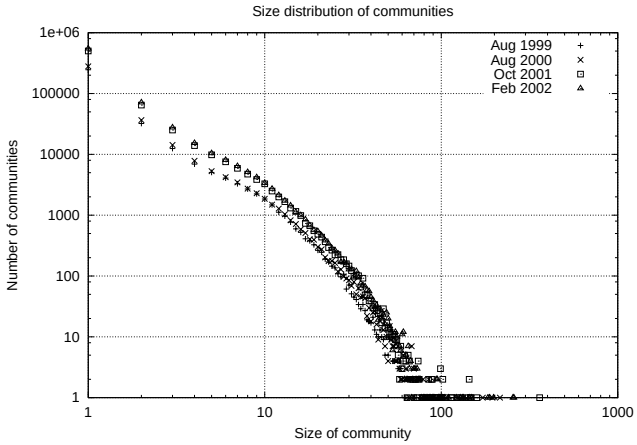

図 6　コミュニティのサイズの分布

Fig. 6　Size distribution of communities

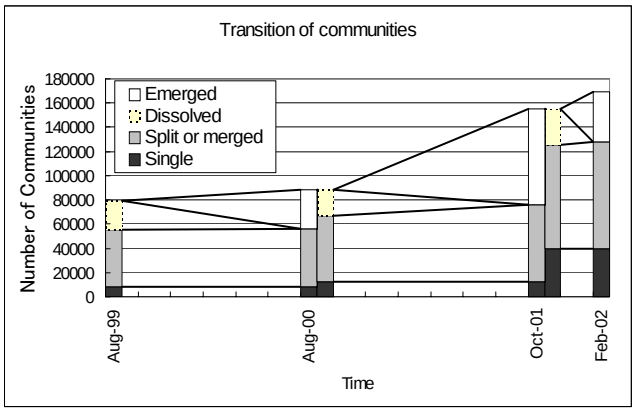

図 7　コミュニティの推移

Fig. 7　Transition of communities

	t_{k-1}: Aug. 99 t_k: Aug. 00	Aug. 00 Oct. 01	Oct. 01 Feb. 02
t_{k-1} での支流の数	28,467	32,490	41,501
本流の数	18,060	22,299	44,425
t_k での支流の数	29,722	41,752	44,305

表 3　分裂または合併したコミュニティ群における本流および支流の数

Table 3　Number of main lines in split or merged communities

を起こしているかを表している．2000 年と 2001 年では前後の時間と比較をしなければならないので 2 本の棒グラフを並べてある．各棒グラフは，コミュニティの個数を表しており，起こった変化の種類によってブロックに分割されている．点線のブロックは，解散したコミュニティ数を表し，白いブロックは発生したコミュニティ数を表す．灰色のブロックは，合併または分裂を起こしたコミュニティ数を表す．最後に，黒いブロックは，対応コミュニティが前向きにも後向きにも 1 つしかなく成長または縮小しか起こさない単独のコミュニティ数である．

図 7 から，コミュニティの過半数は合併と分裂を起こしていることが分かる．解散するコミュニティの数は，1 年間で全体の約 25%程度である．一方，上記の単独のコミュニティ数は，1999 年から 2001 年まで約 10%程度と非常に少ない．

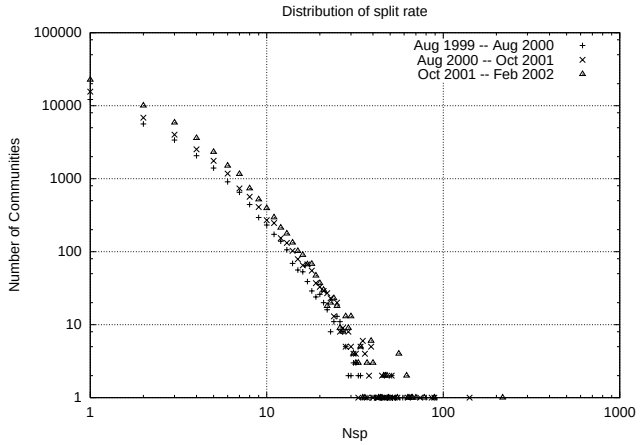

図 8　分裂率の分布

Fig. 8　Distribution of split rate

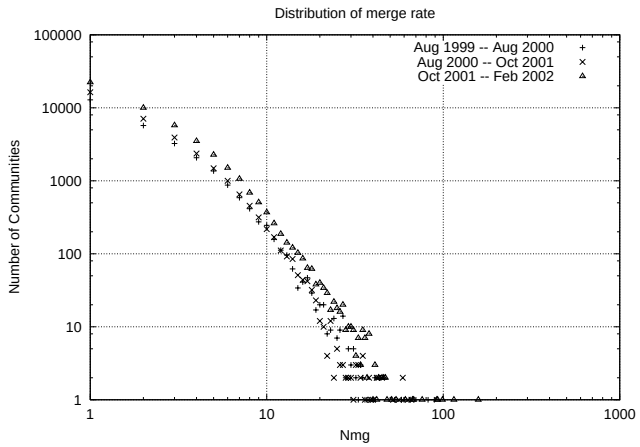

図 9　合併率の分布

Fig. 9　Distribution of merge rate

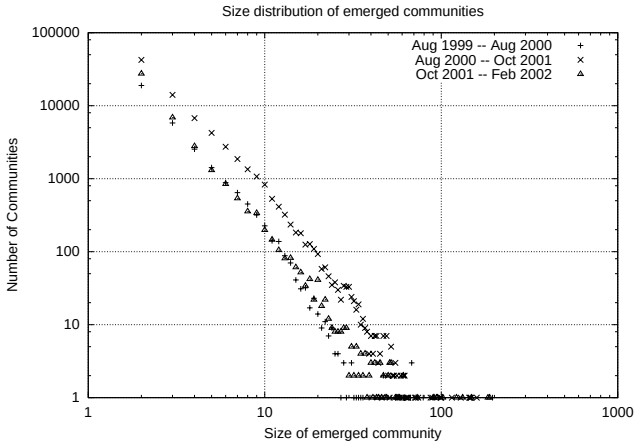

図 10　発生したコミュニティのサイズ分布

Fig. 10　Size distribution of emerged communities

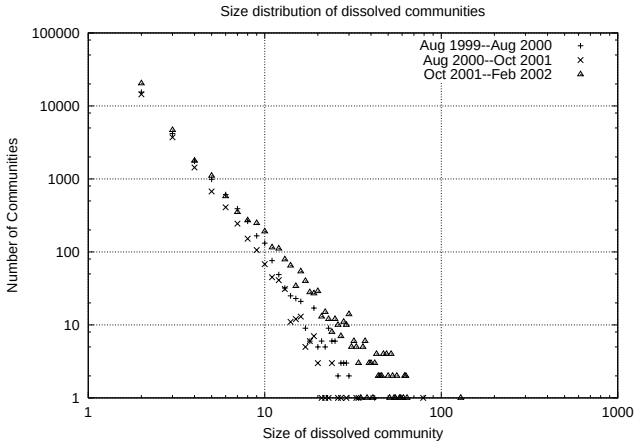

図 11　解散したコミュニティのサイズ分布

Fig. 11　Size distribution of dissolved communities

6.2.1　分裂と合併

分裂と合併による変化は複雑である．分裂したコミュニティは，次の時期には合併しているかも知れず，その逆もあり得る．しかし，分裂や合併を起こしているコミュニティ群のなかでも，本流 (第 3.2 節参照) を取り出すことで比較的安定したコミュニティを中心としてその周辺の変化をとらえることが出来る．表 3 は，分裂または合併したコミュニティ群における本流および支流の数を示している．1999 年から 2001 年の 1 年毎の計算では，本流の数は t_{k-1} に分裂したコミュニティの数の約 40%である．2001 年から 2002 年の計算では期間が 4ヵ月と短いため約 50%が本流となっている．

次に，分裂率と合併率の分布を示す．図 8，9 は，分裂または合併したコミュニティの数を，分裂または合併した URL 数 (第 3.2 節における N_{sp}，N_{mg}) の関数として両対数グラフで表したものである．分布は両方のグラフにおいてほぼべき乗則に従い，グラフの形状と値は分裂でも合併でもほぼ同じである．これは，分裂および合併が変化前後においてほぼ対象的な変化であることを表している．

6.2.2　発生と解散

発生および解散したコミュニティのサイズの分布は，またべき乗則に従う．図 10，11 には，これらの分布を両対数グラフで示した．多くの場合，べき指数は 3.2 より大きい．チャート全体のサイズ分布におけるべき指数が約 3.0 であることを考慮すると，コミュニティは小さい程発生し易く，かつ解散し易いということが分かる．

6.2.3　成　長　率

最後に，コミュニティの成長率の分布を示す．図 12 には，コミュニティの数を成長率の関数として表した．縦軸には対数軸を使用している．このグラフでは，本流の成長率のみを示してある．成長率の絶対値はほとんどのコミュニティにおいて小さく，グラフの形状はきれいな対称性を示している．この対称性は，コミュニティ全体のサイズ分布の形状が時間が経っても保たれることの理由の 1 つであると考えられる．

7.　まとめと今後の課題

本論文では，1999 年から 2002 年の間に 4 回収集した日本のウェブアーカイブを用いてウェブコミュニティの発展過程を詳細に分析した．我々の手法は，各アーカイブから全てのウェ

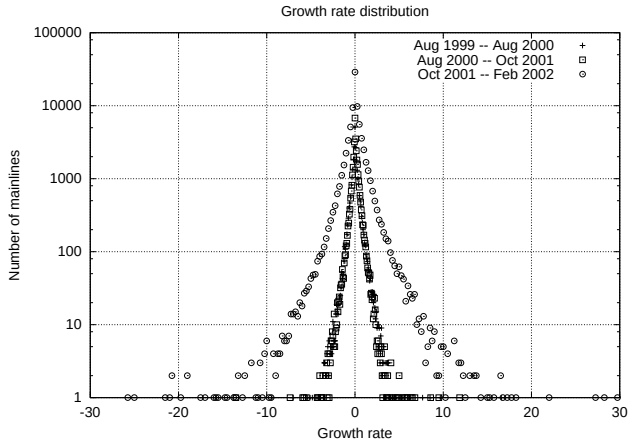

図 12　成長率の分布

Fig. 12　Distribution of growth rate

ブコミュニティを抽出し，時間軸に沿って比較を行うことでコミュニティの発展過程を把握する．この際，コミュニティの成長率，新規率，安定率など，興味ある発展過程の抽出に有用なメトリックスを導入して，それらの値の分布や，時間変化を示した．また，コミュニティの発展過程の例を発展過程ビューアを用いて示した．

分析の結果，コミュニティの構造は，大きく変化しているにもかかわらず，コミュニティのサイズの分布はべき乗則に従い，時間を経てもべき指数は大きく変化しないことが判明した．これは以下のように，すべての発展過程がべき乗則に従って起こり，分裂と合併の間，および成長と縮小の間に，対称性があることによると考えられる．

- 発生および解散したコミュニティのサイズ分布はべき乗則に従う．
- 分裂および合併を起こすコミュニティについて，分裂率および合併率はほぼ同じべき乗分布に従う．これは，t_{k-1} において，分裂したコミュニティ群が，t_k において分裂率とほぼ同じ分布の合併率により合併していることを表す．
- 本流コミュニティにおいて，成長率の分布は 0 をはさんで左右対称のグラフとなる．これはコミュニティの成長と縮小が同じ分布で起きていることを示す．

現在のウェブスナップショットは 1 年毎の収集であるため，周期が長すぎて詳細なコミュニティの変化を追うことができない．今後は収集の周期を短くして，より詳細で連続的な発展過程を調査する予定である．

文　献

[1] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The Connectivity Server: fast access to linkage information on the Web. In *Proceedings of the 7th International World Wide Web Conference*, 1998.

[2] K. Bharat and G. A. Mihaila. When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics. In *Proceedings of the 10th World-Wide Web Conference*, 2001.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web. In *Proceedings of the 9th World-Wide Web Conference*, 2000.

[4] S. Chakrabarti. Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation. In *Proceedings of the 10th World-Wide Web Conference*, 2001.

[5] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*, 1998.

[6] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. In *Proceedings of the 8th World-Wide Web Conference*, 1999.

[7] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient Identification of Web Communities. In *Proceedings of KDD 2000*, 2000.

[8] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web Communities from Link Topology. In *Proceedings of HyperText98*, 1998.

[9] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[10] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In *Proceedings of the 25th VLDB Conference*, 1999.

[11] R. Lempel and S. Moran. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. In *Proceedings of the 9th World-Wide Web Conference*, 2000.

[12] S. R. Ravi Kumar, Prabhakar Raghavan and A. Tomkins. Trawling the Web for emerging cyber-communities. In *Proceedings of the 8th World-Wide Web Conference*, 1999.

[13] M. Toyoda and M. Kitsuregawa. Creating a Web Community Chart for Navigating Related Communities. In *Conference Proceedings of Hypertext 2001*, pp. 103–112, 2001.

[14] Wayback Machine, The Internet Archive. http://www. archive. org/.

[15] 豊田, 喜連川. ウェブコミュニティの発展過程抽出手法. 電子情報通信学会データ工学研究会, 5 月 2002 年.