ウェブコミュニティ出現におけるリンク構造成長パターン分析

今藤　紀子†　　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: †{imafuji,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　本論文では，ウェブ空間の構造変化過程を解明するための一つのアプローチとして，ウェブコミュニティにおけるリンク構造の成長過程を分析する．ウェブコミュニティとは，共通する話題を取り上げているページの集合を意味し，それらを指すページ群と共に密に結びついたリンク構造を有する．1999 年から 2002 年に収集した国内のウェブアーカイブを基に作成した各年のコミュニティ集合を利用し，一定期間中におけるリンク構造の変化を調査する．とりわけ，その年に新たに出現したコミュニティに着目し，新たに作成されたウェブページがウェブコミュニティの一部として抽出される間におけるリンク構造の成長を数種のパターンに分類して捉え，各成長パターンが示すウェブ上における意味を分析する．

キーワード　ウェブコミュニティ，ウェブグラフ，リンク構造，HITS,

An Analysis on Link Structure Evolution Pattern of Web Communities

Noriko IMAFUJI† and Masaru KITSUREGAWA†

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

Abstract　In this paper, we analyze the growing process of link structure of web communities, which is an approach for understanding evolution of the web. A web community is a set of web pages created by individuals or associations with a common interest on a topic. These pages are co-cited by multiple pages, and form densely connected link structure. We examine the transition of link structure within a certain period of time using four sets of web communities created from Japanese web archives crawled in four periods between 1999 and 2002. Especially, we focus on the web communities which did not exist in the previous year and classify the evolution of link structures into some patterns. We analyze the semantics of each evolution pattern on the web.

Key words　web community, web graph, link structure, HITS

1. はじめに

ウェブページとその間に張られたハイパーリンクをそれぞれノード，エッジと見なせば，ウェブは巨大な有向グラフ（ウェブグラフと呼ばれる）である．サーチエンジンに代表される情報検索技術のめざましい発展は，混沌としたウェブ空間からページ単位での様々な情報を引き出すことを可能にした．局所的に見たとき，ウェブ空間はウェブページの追加・削除により無秩序に変化し続けているごとく振る舞っているが，一方で巨視的に捉えると，実社会の動きを如実に映し出すという側面を持つといわれている．例えば，実社会においてある話題が注目されると，その話題に関する多くのウェブページがウェブ上に現れ，質の良いページはブックマークやリンク集などからリンクされることにより，非常に密なリンク構造を構築していく．このことから，ウェブの巨視的な構造を理解し，それを時系列で捉えることで得られる情報の有用性は高く，それが生み出す様々な可能性に対する注目が集まっている．

我々の目的は，ウェブ空間の構造変化過程を解明することにある．そのための一つのアプローチとして，ウェブコミュニティにおけるリンク構造の成長過程を分析する．ウェブコミュニティ

（以降，単にコミュニティと呼ぶ）とは，話題が共通するウェブページの集合を意味する．コミュニティの存在が，ウェブ上に存在する一つの話題を意味することから，コミュニティは，ウェブにおける現象を巨視的に捉える一つの指針として利用できる．本論文では，新たにコミュニティがウェブ上に出現するまでのリンク構造の成長過程を分析する．換言すれば，ウェブ上に新たに作成され存在する個々のウェブページがハイパーリンクにより結びつき，初めてコミュニティとして抽出されるに至るまでのリンク構造の成長のメカニズムを解明する．

我々は，共参照ページの存在に着目し，新規出現コミュニティ周辺のリンク構造の成長を 4 種のパターンに分類して捉える．一方で「認知度」と「純度」という二つの側面からウェブ上におけるコミュニティの存在を位置づける．それらを軸にした平面を導入し，各成長パターンのこの平面上で示す．次に，1999 年から 2002 年に収集した国内のウェブアーカイブを基に作成した各年のコミュニティ集合を利用して実データにおける新規コミュニティの各成長パターンの割合を調査し，新規出現コミュニティの認知・純度平面上における推移を分析する．

本論文の構成は以下の通りである．第 2 節では，関連研究について述べる．第 3 節では，コミュニティの抽出手法について簡単に解説した後，新規出現コミュニティを定義し，それらのリンク構造成長パターンの分類を行う．第 4 節では，大規模ウェブデータベースを用いた実験により，実データにおける成長パターンの検証結果を示す．最後にまとめと今後の課題を述べる．

2. 関 連 研 究

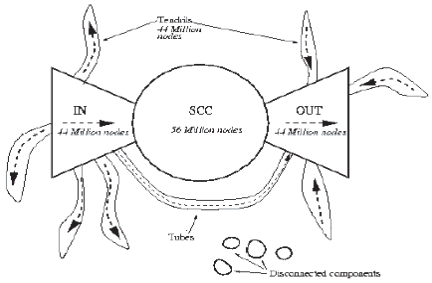

図 1　ウェブのボウタイ構造
Fig. 1　The bow-tue structure of the web

ウェブの巨視的な構造を理解する試みとしては, [6], [7] などがある．Borodin らは，大規模なウェブのリンク構造解析を行い，ウェブ上の約 92%のページが連結であり，さらにそれは同等の大きさから成る巨大な 3 つ連結成分に分類できることを示した [6]．この 3 つの連結成分は，ウェブのボウタイ構造としてよく知られている (図 1 参照)．また，Albert らは，任意の 2 つのページが平均 19 クリックで到達できる事を示した [7]．これらの研究は，ある一時点におけるウェブの構造を理解する試みである．我々の研究は，ウェブの構造の成長過程を理解する試みであり，ウェブを時系列的に捉えるという点で，これらの研究とは異なる．

ウェブの時系列分析に関する研究としては, [8]〜[10], [17] などがある. [8], [9] は，ウェブページ単位での解析であり, [10] は，サイト単位での解析である．これらは，ウェブの局所的な時系列変化を捉えているに過ぎない．一方，豊田らによる [17] は，コミュニティ単位での時系列解析である．4 年分の実データを利用し，HITS 系手法によって得られたコミュニティ集合における成長過程を分析している．しかしながら，HITS 系手法により抽出可能なコミュニティは，ある程度リンクが密に張りめぐらされたページ集合のみであるので，コミュニティとして抽出されるに至らない未熟なリンク構造を持つページ集合の成長過程については，解析されていない．我々は，コミュニティを軸としたウェブの時系列解析を行う．とりわけ，その構造が最も変化すると考えられる，ウェブ上に新たに追加されたウェブページの集合がコミュニティとして抽出可能になるまでのリンク構造の成長過程を分析する．

3. コミュニティの出現と成長

コミュニティとはある共通する話題について書かれたウェブページの集合を意味する．本章では，コミュニティの抽出手法，特に，以降の分析において利用するコミュニティセットの抽出手法について説明し，分析対象となる新規コミュニティの抽出条件を示す．また，4 種に分類されたリンク構造成長パターンを示し，各成長パターンが持つウェブ上における意味について述べる．

3.1 ウェブコミュニティの抽出

これまでにウェブから効率よくコミュニティを抽出する手法が多数提案されてきた [2], [3], [11]〜[13]．それらの手法は，ウェブにおけるハイパーリンク構造の特徴をそれぞれ異なる視点から捉え，それを反映させたリンク構造でコミュニティを表現する. [13] では，少なくとも一つ以上の完全 2 部グラフを含む 2 部グラフとしてコミュニティを認識する．このコミュニティは，代表的な関連ページアルゴリズムである HITS [1]〜[4] におけるハブ・オーソリティの概念(注1)と基本的な考え方は同じであるため，両者で得られるコミュニティのメンバーは概ね一致している．また, [11], [12] においては，コミュニティとは，「コミュニティの外のページへの（又は，からの）リンクよりもコミュニティ内のページ同士のリンクを多くもつ」という条件を満たすウェブページの集合と定義している．このような集合が最大流アルゴリズム (maximum-flow algorithm) [14]〜[16] により得られることから Max-Flow 手法と呼ぶ．

図 2 に各手法が定義するコミュニティにおけるリンク構造の典型例を示す．図中の点，矢印はそれぞれ，ウェブページ，その間に張られたハイパーリンクを意味する．(a) は，完全 2 部グラフから成る HITS 系手法によるコミュニティ，(b) は，Max-Flow 手法によるコミュニティである．前者は，オーソリティの集合をコミュニティメンバーとして抽出することが多い．後者では，円で包囲されたページ集合が抽出される．後者ではコミュニ

(注1)：多くのページからリンクされている良質なページをオーソリティ(ページ)といい，逆に，多くの良質なページへリンクしているページをハブ (ページ) という．

ティ内部と外部の連結度差にのみ着目しているため，内部のリンク構造が疎であっても抽出される．一方，HITS 系手法では，リンク構造が成熟していないページ集合はコミュニティとして発見されない．

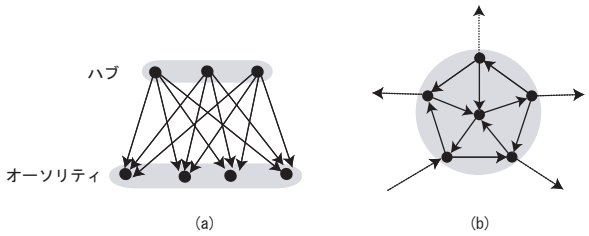

図 2　コミュニティとリンク構造; (a)HITS 系手法 (b)Max-Flow 手法
Fig. 2　Communities and link structures
by HITS family method(a) and Max-Flow method(b)

本論文では，HITS 系手法によるコミュニティ集合を利用する．この手法は，ウェブグラフ全体からメンバーが重複することなくコミュニティを効率よく抽出することを主眼として，豊田らにより提案された [5]．以下にその手順を示す．詳細は, [5] を参照されたい．

（1） シードページ(注2)の集合を入力する．

（2） 各シードページに対し，HITS 系関連ページアルゴリズムを施す．

（3） オーソリティ導出グラフ (ADG) を構築する．

（4） ADG から対照導出グラフ (SDG) を構築する．

（5） SDG 内で連結なノード集合をコミュニティのメンバーとして出力する．

オーソリティ導出グラフ (ADG) とは，各シードが，どのように *Companion-* [5] と呼ばれる HITS 系関連ページアルゴリズムによって他のシードを導出するかを示す有向グラフである．ノードは，シードセット内のシードページから成る．また，ノード u から v への有向辺は，u が Companion-で求められた関連ページのうちの一つとして v を導出していることを表している．ADG において u，v 間に双方向にエッジが存在するとき，u，v 間に無向辺を生成し，これらの無向辺とその両端点のノード集合により得られた無向グラフが，対照導出グラフ (SDG) である．SDG において密に連結しているサブグラフをコミュニティの核として抽出し，これらに残りのノードを加えることによりコミュニティを得る．

3.2　新規出現コミュニティ

$t_1, ..., t_n$ をクロールした時期(注3)，t_i におけるコミュニティ集合を $C(t_i) = \{c_1(t_i), c_2(t_i), ..., c_m(t_i)\}$ とする．$C(t_i)$ 内のコミュニティのうち t_i において新たに出現したコミュニティ $c^+(t_i)$（つまり，以降の分析対象となるコミュニティ）とは，以下の条件をみたすコミュニティを意味する．

条件 1　：$|c^+(t_i)| \geqq 5$

条件 2　：$c^+(t_i)$ のメンバーページのうち，50%以上が異なるサーバの URL

条件 3　：$c^+(t_i)$ のメンバーページのうち，t_{i-1} においても存在していた URL が 80%以上でかつ，それらのメンバーは以下のいずれかの条件を満たす．

- $C(t_{i-1})$ のいずれのメンバーにもなっていない
- $c_j(t_{i-1}) \in C(t_{i-1})$ のメンバーのとき，$|c_j(t_{i-1})| \leqq 2$

条件 1，2 は，$C(t_i)$ に対する分析対象コミュニティの絞り込みである．コミュニティサイズが極めて小さい，つまり，数個のメンバーのみからなるコミュニティの場合，それらはコミュニティとして成長途上にあり，明確なトピックを持たない場合も多い．よって，条件 1 により，コミュニティサイズに関する閾値を与える．一つのコミュニティ内に見られる同じサーバの URL は，同じサイト内のページであることが多い．それらのページが大部分を占めるとき，コミュニティは一つのサイトを意味することとなり，コミュニティとしての本質を持たない．よって，条件 2 により，サイトから成るコミュニティの排除を行う．なお，同じサーバであるか否かは，ドメインの一致・不一致により判断する．このため，同じサーバではあるが，わずかに異なるドメイン名を持つものは，異なるサーバの URL として認識されない．条件 3 は，”前年には無いコミュニティ”の具体条件を意味する．以上より，新規出現コミュニティとは，ある年抽出された意味のあるコミュニティのうち，前年には，リンク構造が未発達のため顕著なオーソリティ傾向を示さずコミュニティとして抽出されていなかったものを意味する．

3.3　リンク構造の成長パターン

分析に利用するコミュニティは，オーソリティページの集合から成る．前述したように，これらオーソリティページを指すハブページと共に密に連結している．コミュニティのリンク構造を現わすコミュニティグラフを以下で定義する．本論文では，コミュニティグラフとリンク構造を同義で用いていることに注意されたい．

> **定義**：$A = \{a_1, a_2, ..., a_l\}$ をコミュニティ c のメンバーページとし，$H = \{h_1, h_2, ..., h_m\}$ を A に 2 つ以上のリンクをもつページ集合とする．$E \subset A \times H$ とするとき，有向 2 部グラフ $G_c(V, E)$ をコミュニティ c のコミュニティグラフという．

ここで，コミュニティ c に関して以下の二つの属性を導入する．ただし，(x, Y) をページ x から集合 Y の要素へのリンク集合とする．

$$H_o = \{x | x \in H, |(x, \overline{A})| > |(x, A)|\}$$
$$H_i = \{x | x \in H, |(x, \overline{A})| \leqq |(x, A)|\}$$
ただし，$H_o \cup H_i = H, H_o \cap H_i = \emptyset$.

図 3 にコミュニティのリンク構造 $G_c(V, E)$ の模式図を示す．この例においては，$V = H \cup A = \{a_1, ..., a_6, h_1, ..., h_5\}$ である．H_o は，A 内のページへのリンクよりも A 外へのリンクを

(注2)：この手法においては，3 つ以上の異なるサーバからリンクされているページをシードページとしている．
(注3)：本論文では，$1 \leqq n \leqq 4$ とし各クロール時期は次章参照のこと．

多く持つページ，H_i は，A 外へのページへのリンクよりも A 内へのリンクを多く持つページを意味する．図は，H_i の両ページは A へのリンクのみ持つ場合を示した例である．

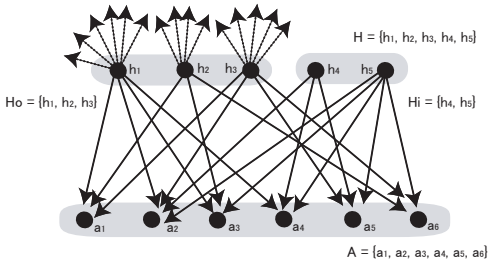

図 3　コミュニティのリンク構造 $G_c(V, E)$
Fig. 3　Link structure of web communities, $G_c(V, E)$

ここで集合 H_o，H_i に属するページの性質を考える．H_o，H_i は，両者とも複数のページへのリンクを張るリンク集やブックマークなどのページある．H_o は，コミュニティに属するページ群と共に，それ以外のページへも多数のリンクを持つことから，大規模なリンク集から成るページであると考えられる．一方，H_i は，主としてコミュニティに属するページのみにリンクを持つことから，ある特定の話題に限定されたリンク集から成るページであると考えられる．たとえば，PC メーカーに関するページの集合 $A = \{a_1, a_2, ..., a_m\}$ から成るコミュニティが存在するとする．これらのページへのリンクは，「日本企業」や「PC 関連情報」というリンク集の中のサブカテゴリとして存在する．ページ集合 A へのリンクを一つのサブカテゴリに持つ大規模リンク集が増えるにつれ，A に属するページの集合としての認知度がウェブ上において高まっていることを意味する．換言すれば，集合 H_o の構成要素数は，ウェブ上における該当コミュニティの**認知度**の指標となっていると言える．一方，H_i に属するページは「PC メーカー」に限定したリンク集から成る．このようなページが増加するにつれ，A に属するページの集合としての境界がウェブ上において明確になっていることを意味する．換言すれば，集合 H_i の構成要素数は，ウェブ上における該当コミュニティの**純度**の指標となっていると言える．

ウェブ上におけるコミュニティの認知度を意味する H_o，純度を示す H_i を用いて，コミュニティのリンク構造の成長パターンを表 1 で定義し，図 4 に新規出現コミュニティにおける各成長パターンごとのリンク構造の変化を例示する．

表 1　リンク構造の成長パターン
Table 1　Link structure evolution pattern

成長パターン	t_{i-1} のとき	t_i のとき
パターン 1	$\|H_o\| > \|H_i\|$	$\|H_o\| > \|H_i\|$
パターン 2	$\|H_o\| > \|H_i\|$	$\|H_o\| \leqq \|H_i\|$
パターン 3	$\|H_o\| \leqq \|H_i\|$	$\|H_o\| > \|H_i\|$
パターン 4	$\|H_o\| \leqq \|H_i\|$	$\|H_o\| \leqq \|H_i\|$

成長パターン 1 に属するページ群は，もともとごく小数の大規模リンク集にリンクされているのみの疎なリンク構造をしており，コミュニティとして抽出されない．一定期間後，複数の他のリンク集により認識され始め，密なリンク構造を構築しコミュニティとして抽出される．

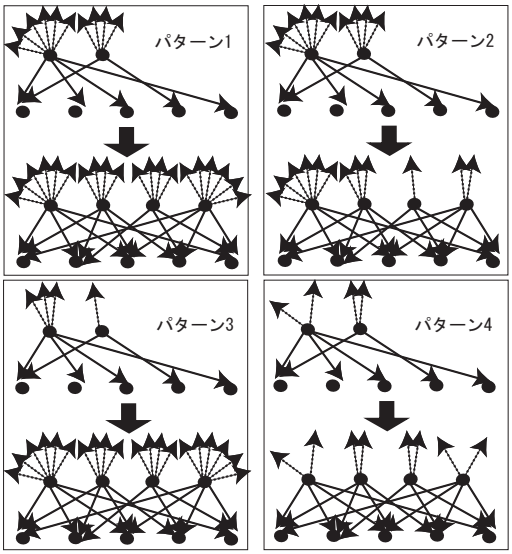

図 4　各成長パターンのリンク構造変化例
Fig. 4　Examples of link evolution patterns

成長パターン 2 に属するページ群は，成長パターン 1 と同様，もともとはごく小数の大規模リンク集からリンクされているのみでコミュニティとして抽出されるほど密なリンク構造を成していない．一定期間後，これらのページ群に関する話題に焦点を絞ったリンク集からのリンクにより密なリンク構造を構築しコミュニティとして抽出される．

成長パターン 3 に属するページ群は，もともとごく小数の小規模リンク集からリンクされているのみの疎なリンク構造をしており，コミュニティとして抽出されない．一定期間後，これらのページ群が取り上げる話題が他の複数の大規模リンク集により認識され始めることにより，密なリンク構造を構築しコミュニティとして抽出される．

成長パターン 4 に属するページ群は，成長パターン 3 と同様，もともとはごく小数の小規模リンク集からリンクされているのみでコミュニティとして抽出されるほど密なリンク構造を成していない．一定期間後，これらのページ群に関する話題に興味を持つページ制作者が増加し，そのページからのリンクにより密なリンク構造を構築しコミュニティとして抽出される．

前述のように，コミュニティは「認知度」と「純度」という二つの指標で表現できる．これらを軸にした平面 (図 5 参照) を想定することにより，コミュニティのリンク構造をこの平面上で位置づけ，さらに，成長過程の視覚化も可能となる．以降，この平面をコミュニティの認知・純度平面，或いは，単に，認知・純度平面と呼ぶ．図 5 は 4 つの成長パターンを，この認知・純度平面上で表現した例である．パターン 1 及び 4 は，認知＝純度の軸の片側での変化に対し，パターン 2 及び 3 は，この軸を交差する変化を意味する．

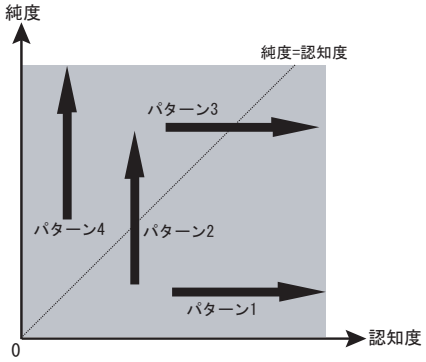

図 5　認知・純度平面上での成長パターン例
Fig. 5　Intuive graph of evolution pattern

4. 実 証 分 析

前章では，本論文で分析するコミュニティ集合と，リンク構造の成長パターンについて述べた．本章では，実データを用いて行った実証分析結果を示し，実例を基にリンク構造の成長パターンと共に分析結果を考察する．

4.1 実 験 概 要

利用データ：1999 年から 2002 年の各年にクロールされた日本国内 (主に jp ドメイン) のウェブページから成る 4 つのウェブアーカイブによるウェブグラフデータベースを利用した．これは，アーカイブの全てのページから URL とリンク情報（どの URL がどの URL へのリンクを持つか）のみを取り出し，ウェブをグラフ表現によりデータベース化したものである．クロール時期と実際にクロールしたウェブページ数，これらのページが保持するリンク情報を元に得られた URL の総数および総リンク数を表 2 に示す．

表 2　ウェブアーカイブの詳細
Table 2　Details of web archives

クロール時期	総クロールページ数	総 URL 数	総リンク数
1999 年 7 月	17M	34M	120M
2000 年 6 月	17M	32M	112M
2001 年 10 月	40M	76M	331M
2002 年 2 月	45M	84M	375M

利用コミュニティ集合：各年のウェブグラフデータベースより，3.1 節で述べた手法を用いて構築されたコミュニティ集合を利用する．入力シードページ集合は，3 つ以上の異なるサーバからリンクを張られているページから成る．各コミュニティ集合の詳細を表 3 に示す．表中 (a)，(b)，それぞれ，前述の分析対象コミュニティ検索条件における条件 1，条件 1 かつ条件 2，および全ての条件 (条件 1 かつ条件 2 かつ条件 3) を満たすコミュニティの総数を意味する．検索条件における t_i は，$1 \leqq i \leqq 4$ で t_1, t_2, t_3, t_4 はそれぞれ，1999 年 7 月，2000 年 6 月，2001 年 10 月，2002 年 2 月を意味する．ちなみに，1999 年におけるコミュニティ総数は，85713 であった．

表 3　コミュニティ集合の詳細
Table 3　Details of web community sets

クロール時期	総コミュニティ数	(a)	(b)
2000 年 6 月	94084	33605	32364
2001 年 10 月	166689	58668	50141
2002 年 2 月	185195	65022	53695

図 6 のグラフは，条件 1 および 2 を満たすコミュニティのうち，前年のコミュニティ集合内に 2 つ以上の重複メンバーが存在するコミュニティが無いもの (換言すれば，前年には該当するコミュニティが存在していなかったコミュニティ) を取り出し，それらのコミュニティのうち前年に URL が存在していたメンバーの割合を示したものである．2001 年から 2000 年の調査において 10%以下の割合が高くなっているのは，2000 年においてウェブアーカイブより一部データが消失したためである．以上より，検索条件を満たす，2000 年から 1999 年，2001 年から 2000 年，2002 年から 2001 年でそれぞれ，653，1983，1675，計 4311 のコミュニティを分析対象とする．

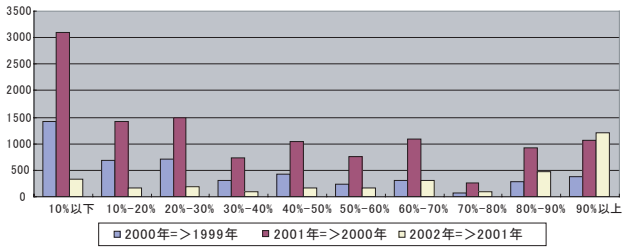

図 6　メンバーページの前年存在割合
Fig. 6　Rate of the member pages existed in the previous year

図 7 のグラフは，各調査期間の t_{i-1} での分析対象コミュニティにおける距離 2(注4)以内に含まれるメンバー数の割合を示している．これによると，前年ではメンバーの多くが非連結で存在していたというコミュニティは殆ど無く，半数以上は前年においても大部分のメンバーが既にある程度連結して存在していたことがわかる．

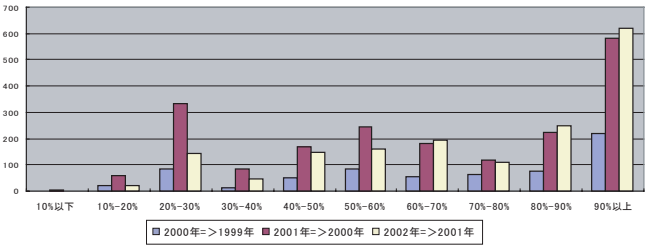

図 7　距離 2 以内に含まれるメンバーの割合
Fig. 7　Rate of members within depth 2.

上述のコミュニティ集合において，以下の値を求める．これらの値により，新規出現コミュニティ集合における成長パターンの分布を調査するとともに，各成長パターンが示す認知・純度平面上での推移を検証する．

(注4)：最短，2 リンクで辿れるという意味

- $|H_o(t_{i-1})|$；t_{i-1} におけるコミュニティの認知度
- $|H_i(t_{i-1})|$；t_{i-1} におけるコミュニティの純度
- $|H_o(t_i)|$；t_i におけるコミュニティの認知度
- $|H_i(t_i)|$; t_i におけるコミュニティの純度

4.2 実験結果

図 8 に一定の期間の前後で $|H_o| > |H_i|$ と $|H_o| \leqq |H_i|$ となるコミュニティの割合を示す．各期間における $|H_o| > |H_i|$ の割合の平均値は t_{i-1} のとき 69.87%, t_i のとき 85.32%であった．新しいウェブページが出現した時点 (つまり，t_{i-1} の時点で)，約 7 割のコミュニティが，既に何らかの大規模リンク集からリンクされているということがわかる．一定期間後，コミュニティとして抽出されるに十分なほど密になったリンク構造の約 85%が，大規模なリンク集によるものであることがわかる．

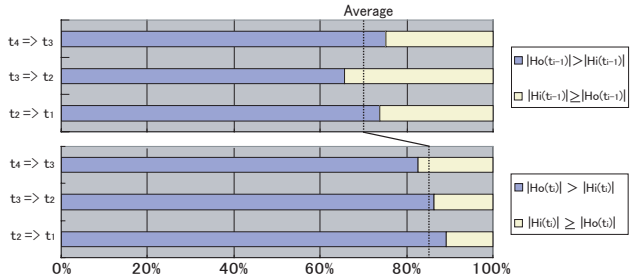

図 8　$|H_o|$ 及び $|H_i|$ の比較

Fig. 8　Comparison between $|H_o|$ and $|H_i|$

図 9 は，3 期間全ての分析対象コミュニティにおける各成長パターンの割合を示したものである．成長パターン 1 は，約 67%で最も多く，成長パターン 3 の約 18%，成長パターン 4 の約 11%と続き，成長パターン 2 は，約 4%と最も少なかった．初めに大規模リンク集によりリンクされ始めたページ集合は，新たな大規模リンク集によるリンクを増加させることが多く，逆に，これらのページ集合のみを指すような小規模リンク集によって純度をあげることは殆ど無い．一方，もともとそれらのページのみを指すような小規模リンク集からリンクされたページ集合のうち約 62%は，一定の期間内に認知度を高め，新たな大規模リンク集によるリンクを増加させている．逆に，約 38%は，複数の小規模リンク集によって純度を高める．

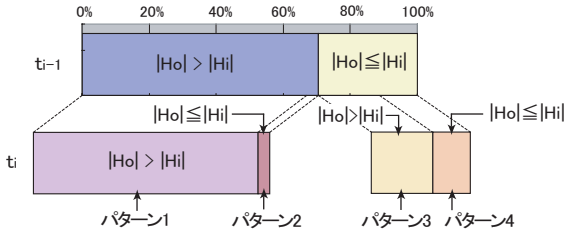

図 9　各成長パターン割合

Fig. 9　Pacentages of link evolution patterns.

表 4 に，各成長パターンにおける $|H_o(t_{i-1})|$，$|H_i(t_{i-1})|$ の各調査期間ごとの平均値，および，全体の平均値を示す．

表 4　各パターンにおける (a)$|H_o(t_{i-1})|$, (b)$|H_i(t_{i-1})|$ の平均値

Table 4　Ave. of (a)$|H_o(t_{i-1})|$, (b)$|H_i(t_{i-1})|$ with respect to each link evolution pattern

	パターン 1		パターン 2		パターン 3		パターン 4	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
$t_2 => t_1$	5.54	0.32	2.35	0.75	0.33	0.50	0.43	3.55
$t_3 => t_2$	3.89	0.11	2.30	0.34	0.22	0.39	0.51	5.76
$t_4 => t_3$	19.83	0.17	2.14	0.39	0.34	0.61	0.65	3.05
Average	10.72	0.17	2.23	0.42	0.27	0.46	0.57	4.29

表 5 に，各成長パターンにおける $|H_o(t_i)|$ 及び $|H_i(t_i)|$ の各調査期間ごとの平均値と，全体の平均値を示す．

表 5　各パターンにおける (a)$|H_o(t_i)|$, (b)$|H_i(t_i)|$ の平均値

Table 5　Ave. of (a)$|H_o(t_i)|$, (b)$|H_i(t_i)|$ with respect to each link evolution pattern

	パターン 1		パターン 2		パターン 3		パターン 4	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
$t_2 => t_1$	11.05	0.64	2.25	4.50	4.57	0.55	1.21	5.26
$t_3 => t_2$	11.17	0.67	2.63	4.24	5.75	0.90	2.41	19.64
$t_4 => t_3$	23.87	0.46	1.93	2.83	5.25	1.25	1.58	7.03
Average	16.38	0.58	2.25	3.62	5.45	0.94	1.90	12.35

表 4, 5 の全調査期間における各成長パターンの平均値を用いると，図 10 のような認知・純度平面上での各成長パターンの推移が見てとれる．

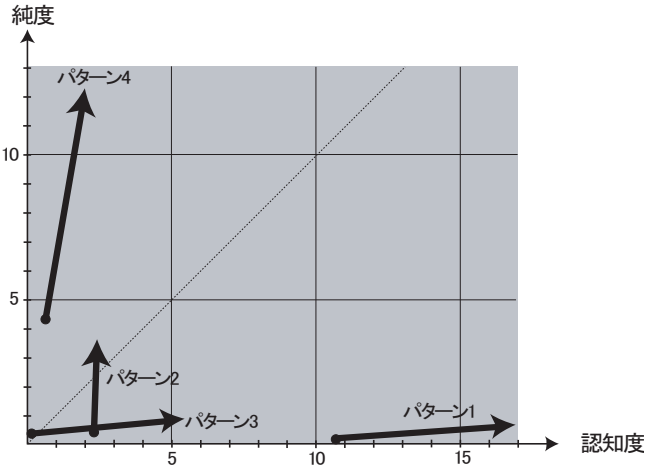

図 10　認知・純度平面上における各成長パターンの推移

Fig. 10　Intuitive graph of link evolution patterns.

4.3 考　　察

図 10 で示された認知・純度平面上に描かれた成長パターンの推移から，どの成長パターンも認知度を増す方向にあることがわかる．ここで，平面上においても特に顕著な推移を見せている 2 つのパターン，成長パターン 1 及び成長パターン 4 に着目する．

成長パターン 1 を示すコミュニティは全体の 67%と最も多い．この成長パターンを示す中でも急激にリンク構造を成長させた例の一つとして，表 6 のようなコミュニティが挙げられる．これは，大分サッカーチーム (1)，大分県に関する情報のポータ

ル的なページ (2)，福岡・大分間の観光マップ (3)，大分県臼杵市のページ (5) からなり，大分県に関するコミュニティであると言える (ただし，4 番目のページは，ページ削除のため確認できず)．この例では，2001 年には，H_o は一つ存在し，これがこのうち 3 つのページへリンクを張っているのみであった．2002 年には，76 ものページがこれらのページへ合計 378 のリンクが張られ，1 年の間に非常に密なリンク構造を構築していた．

表 6　成長パターン 1 の実例
Table 6　An example of evolution pattern 1

1 http://.../common/trinity/TRINITY.HTML
2 http://www.coara.or.jp/oitanetnavi/
3 http://tenjin.coara.or.jp/TOPIC/kankomap/
4 http://.../VSHOP/NewVSHOP/shop/37shop/
5 http://www.city.usuki.oita.jp/from/index98.html

上記の例におけるコミュニティは，2002 年の時点では大分県という枠組みで浮かび上がってきている．しかしながら，例えば，大分サッカーチームは，地方サッカーチームのページから成るコミュニティ，観光マップは，観光情報を集めたページから成るコミュニティ，といった具合に，他にも色んな視点からの分類が可能である．このように，認知・純度平面上を低い純度で水平に推移するコミュニティは，未だコミュニティとしての成長過程にあると考えられ，全体の約 67%がこの成長パターンを示すことから，ウェブ上においてコミュニティとしてページの集合は定常を保つのではなく，変化し続ける部分が多いということを表している．

成長パターン 4 を示すコミュニティは全体の 1 割強とそれほど多くない．この成長パターンを示す中でも急激にリンク構造を成長させた例の一つとして，表 7 のようなコミュニティが挙げられる．これは，colonolog と呼ばれるウェブログプログラムのページ (1)，ログ解析ツールディレクトリのページ (2,5)，PostgreSQL のインストール方法解説ページ (3)，HTTP の詳細が記述されているページ (4) などからなり，直接，間接的にログ解析に関連するページの集合であるといえる．この例では，2001 年には，2 つの H_i が存在し，このページからメンバーページへ合計 8 つのリンクが張られていた．2002 年には，H_i は 187 にもなり，これらのページからメンバーページへ合計 748 のリンクが張られていた．平均すれば，どの H_i からもちょうど 4 つのメンバーページへのリンクを持っていることになり，一年の間にリンク構造としては非常に密に連結した 2 部グラフに成長している．

表 7　成長パターン 4 の実例
Table 7　An example of evolution pattern 4

1 http://...resources/cronolog/
2 http://.../World_Wide_Web/Servers/Log_Analysis_Tools/
3 http://.../t-ishii/PostgreSQL/6.5/apache_php.html
4 http://.../Protocols/rfc2616/rfc2616.txt
5 http://.../Internet/Site_Management/Log_analysis/

この成長パターンに属するページ集合としては，ニュース記事や新商品に関する紹介ページなどが多く見られ，成長パターン 1 とは異なり，別の視点からの分類が存在しない．一方，上記の例のように，急激なコミュニティのリンク構造の成長は，2001 年から 2002 年にかけてウェブログ解析に対する興味が高くなっていることを如実に示している．このように，認知・純度平面上を垂直に推移するコミュニティは，ウェブから急激に盛り上がっている話題やトレンドを抽出するのに適した成長パターンであるといえる．

5. ま と め

本論文では，コミュニティにおけるリンク構造の成長過程を分析した．とりわけ，ウェブ上に新規に出現したコミュニティに着目し，新たに作成されたウェブページがウェブコミュニティの一部として抽出される間におけるリンク構造の成長を数種のパターンに分類して捉え，実データを用いた実験により，各成長パターンに分類されるコミュニティの割合を検証した．また，認知度，および純度という二つの指標によりコミュニティを捉え，それらの指標を軸とした平面上でウェブ上における成長過程を表現した．各成長パターンのウェブ上における意味を捉えることにより，成長パターンを特定することにより，実社会での流行や，人々の興味の盛り上がりをウェブを通して抽出できる可能性があることがわかった．

今後の課題としては，以下のことが含まれる．

- 新規出現コミュニティ以外のコミュニティの成長パターンの分類と検証
- 2002 年以降のデータにおける成長パターンの検証
- ウェブから抽出できるトレンドや流行のリンク構造による解析

一つ目は，今回は，新規に出現するコミュニティのみに焦点を絞りの検証を行った．また，コミュニティのメンバーであるオーソリティ側のページは固定し，ハブ側のページの変化を解析した．今後は，ハブ側を固定したオーソリティ側のページの変化についても考えていきたい．二つ目は，我々の研究室では，2002 年以降もウェブのクロールを続けている．2002 年以降のデータからは，クロール方針が異なり，jp ドメインの他に".com"も含まれるようになっている．".com"が含まれることで今回の検証とどのような違いを生じるか非常に興味深い．三つ目は，今回の検証で，急激に盛り上がっている話題やトレンドを抽出するにふさわしい成長パターンを特定することができた．リンク構造の成長特性を利用したトレンドの抽出手法を考案していきたい．

文　　献

[1] J.M.Kleinberg: *Authoritative Sources in a Hyperlinked Environment*, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677, 1998.

[2] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

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

[4] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In *The VLDB Journal*, pages 639–650, 1999.

[5] M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. *12th ACM Hypertext*, pages 103–112, 2001.

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, A. Rajagopalan, R. Stata, A. Tomkins,and J. Wiener, “Graph strucutre in the web,” In Proc. of 9th WWW Conference, 2000

[7] R.Albert, H. Jeong, and A. L. Barabasi, “Diameter of the world wide web, ” Nature, 401:130, 1999

[8] B.E.Brewington and G.Cybenko, “How dynamic is the web?,” In Proc. of 9th WWW Conference, 2000

[9] J. Cho and H. Garcia-Molina. “The evolution of the web and implications for an incre-mental crawler,” In Proc. of 26th VLDB, 2000

[10] K.Bharat, B. W. Chang, M, Henzinger and M.Ruhl. “Who Links to Whom: Mining Linkage between Web Sites,” In Proc. of IEEE ICDM, 2001.

[11] G. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, 2000.

[12] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[13] R.Kumar, P.Raghavan, S.Rajagopalan, and A.Tomkins, “Trawling the web for emerging cyber-communities,” In Proc. 8th WWW Conference, 1999.

[14] R.K.Ahuja, T.L.Magnanti, and J.B.Orlin, “Network Flows : Theory, Algorithms, and Applications,” Prentice Hall, Englewood Cliffs, NJ, 1993.

[15] A.V.Goldberg and R.E.Tarjan, “A new approach to the maximal flow problem,” In Proc. 18th Ann. ACM Symposium on Theory of Computing, 1986.

[16] L.R.Ford Jr. and D.R.Fulkerson, “Maximal flow through a network,” Canadian J.Math.,8:399–404, 1956.

[17] M.Toyoda and M.Kitsuregawa.:Extracting evolution of web communities from a series of web archives. In Proc. of 14th Conference on Hypertext and Hypermedia(Hypertext 03), pp.28-37, 2003.