日本におけるウェブコミュニティの発展過程

Evolution of Japanese Web Communities

豊田 正史♥ 喜連川 優♦

Masashi TOYODA Masaru KITSUREGAWA

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ウェブコミュニティの全体的な発展過程を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ウェブコミュニティをすべて抽出し，アーカイブ間でウェブ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．分析の結果，ウェブコミュニティの構造は全体的には大きく変化しているにもかかわらず，ウェブコミュニティのサイズの分布は常にべき乗則に従い，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．

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 examine evolution of web communities by comparing Japanese web archives crawled four times from 1999 to 2002. Statistics of these archives and the evolution are shown, and the global behavior of evolution is described. We found that the size distribution of communities follows the power-law and its exponent is not change over time, while the structure of communities changes dynamically.

1. はじめに

近年ウェブは急激に成長を続けてきており，日々多くのページが作成および削除されることで，そ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閲覧できるサービス[7]も開始されているが，まだ単独のページの変化しか見ることはできない等，まだ十分な機能を提供できていない．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．

♥ 正会員 東京大学生産技術研究所
toyoda@tkl.iis.u-tokyo.ac.jp
♦ 正会員 東京大学生産技術研究所
kitsure@tkl.iis.u-tokyo.ac.jp

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[2,3,5]．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，これまでに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[6]の成果を用いている．その上で，各コミュニティの時間変化を調査した．

本論文では，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，および，コミュニティの発展過程における全体的な挙動を示す．以下では，まずウェブコミュニティチャートの概要，およびコミュニティ発展過程の調査方法を述べ，発展過程の種類を分類する．次に，実験に用いたアーカイブと抽出したチャートの詳細を示し，具体的な発展過程の例を示す．最後に，発展過程の全体的な挙動の分析を行い，その結果を示す．

2. ウェブコミュニティチャート

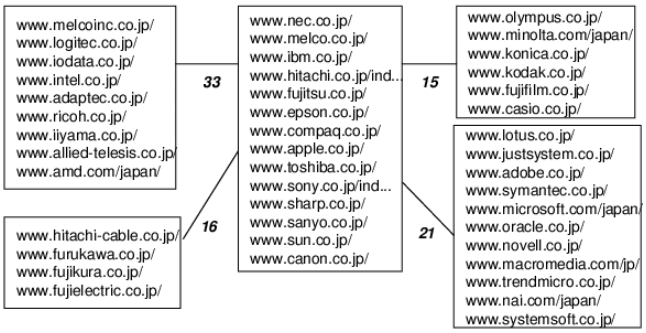

図 1 ウェブコミュニティチャートの一部
Fig.1 A part of web community chart

本節では，ウェブコミュニティチャート[6]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[6]を参照されたい．

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]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，良いオーソリティページおよび良いハブページを抽出し，(3)上位のオーソリティ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．ここで良いオーソリティとは，多くの良いハブ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．良いハブとは，リンク集およびブックマークなど，多くの良いオーソリティ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，密に結合したハブとオーソリティが抽出され，それらがよく関連したページを表すことが[6]で示されている．

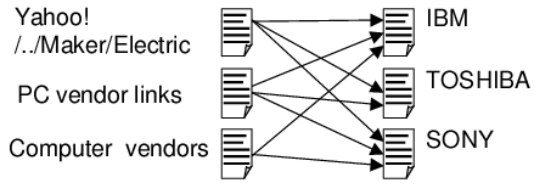

図 2 オーソリティおよびハブからなる典型的なグラフ構造

Fig.2 Typical graph structure of hubs and authorities

図 2 に典型的なオーソリティとハブのグラフ構造の 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オーソリティ 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ハブとしてある．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オーソリティ とハブ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[6]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. ウェブ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を観測する手法について述べ、発展過程の種類を分類する。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．

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分を比較調査する．

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 $C(t_{k-1})$ 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のように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，$W(t_{k-1})$ 内には存在するが，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，$W(t_k)$ にはまだ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．

4. ウェブアーカイブとウェブコミュニティチャートの詳細

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．

Year	Period	#Pages	#URLs	#Links	#Seeds	#Communities
1999	Jul. to Aug.	17M	34M	120M	657K	79K
2000	Jun. to Aug.	17M	32M	112M	737K	88K
2001	Early Oct.	40M	76M	331M	1404K	156K
2002	Early Feb.	45M	84M	375M	1511K	170K

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

まず，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．

次に，これら 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，第 2 節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[6]を参照されたい)．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は表 1 に示してある．

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，および成長率等の発展の度合いによる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[8]を参照されたい．

図 3 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 3 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．これは新しく現れた URL の割合で，周辺のコミュニティをソートすると得られる．図 3 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．テロ事件の直後から平和活動に対して人々の関心が集ってハブページが多数作成され，急速にコミュニティとして発生したことが分かる．

図 3 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. 3 An emerged community around an Islam information community

6. 全体的な発展過程の分析

本節では，ウェブコミュニティの全体的な発展過程を分析する．

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 4 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．

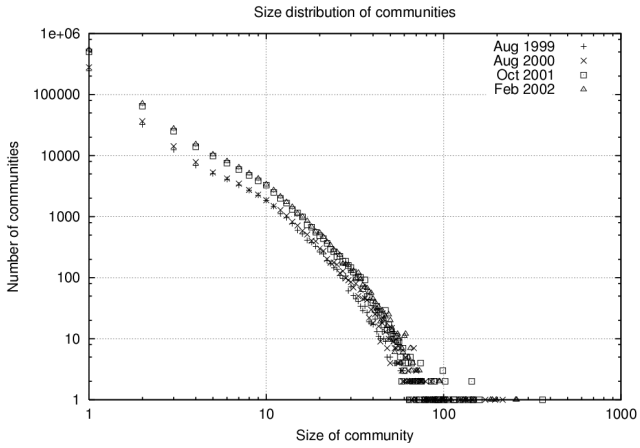

図 4 コミュニティのサイズ分布
Fig. 4 Size distribution of communities

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 5 は，t_{k-1}からt_kの間に何個のコミュニティがどのような種類の変化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 5 から，コミュニティの過半数は合併または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．

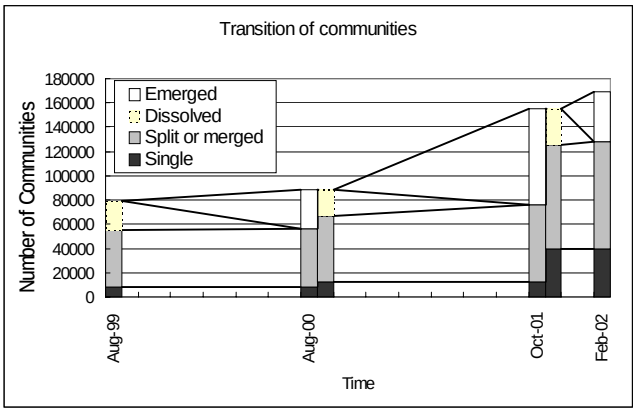

図 5 コミュニティの推移
Fig. 5 Transition of communities

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分析した．我々の手法は，各アーカイブから全てのウェブコミュニティを抽出し，時間軸に沿ってコミュニティの比較を行うことでコミュニティの発展過程を観測する．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自体は，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．

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] 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.

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

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

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

[5] 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.

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

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

[8] 豊田正史，喜連川優．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, 2002 年 5 月.

豊田 正史 Masashi TOYODA
東京大学生産技術研究所産学連携研究員．1999 東京工業大学情報理工学研究科博士後期過程修了，博士（理学）．ウェブマイニング，ユーザインタフェース，ビジュアルプログラミングの研究に従事．日本データベース学会正会員．情報処理学会, 日本ソフトウェア科学会, ACM, IEEE CS 各会員．

喜連川 優 Masaru KITSUREGAWA
東京大学生産技術研究所教授．1983 東京大学大学院工学系研究科電子情報工学専攻博士課程修了．工学博士．データベース工学, 並列処理, ウェブマイニングに関する研究に従事．情報処理学会理事．SNIA-Japan 顧問．日本データベース学会理事．ACM SIGMOD Japan Chapter Chair．VLDB Trustee，IEEE ICDE，PAKDD，WAIM ステアリングコミティメンバ．