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

An Analysis on Link Structure Evolution Pattern of Web Communities

今藤 紀子♥ 喜連川 優♦

Noriko IMAFUJI Masaru KITSUREGAWA

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．

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.

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度・純度を指標とした変化を分析する．

2. 関連研究

ウェブの巨視的な構造を理解する試みとしては，[1], [2]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[1]．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．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[2]．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, [3]〜[6]などがある．[3],[4]は，ウェブページ単位での解析であり，[5]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[6]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

♥ 正会員 東京大学 生産技術研究所
imafuji@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

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[7]〜[10]．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．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[7]．詳細は，[7]を参照されたい．

3.2 新規出現コミュニティ

t1,...,tn をクロールした時期，ti におけるコミュニティ集合を C(ti) = {c1(ti), c2(ti),...,cm(ti)} とする．C(ti) 内のコミュニティのうち ti において新たに出現したコミュニティ c^+ti（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]：$|c^+(ti)| \geqq 5$

[条件2]：$c^+(ti)$ 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]：$c^+(ti)$ のメンバーページのうち，ti-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) のいずれのメンバーにもなっていない

・cj(ti-1)∈C(ti-1) のメンバーのとき，$|cj(ti-1)| \leqq 2$

条件1，2は，C(ti)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 ={a1, a2,...,al}をコミュニティcのメンバーページとし，H ={h1, h2,...,hm}をAに2つ以上のリンクをもつページ集合とする．E⊂A×Hとするとき，有向2部グラフGc(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．

$Ho = \{x \mid x \in H, |(x, \bar{A})| > |(x, A)|\}$

$Hi = \{x \mid x \in H, |(x, \bar{A})| \leqq |(x, A)|\}$

ただし，$Ho \cup Hi = H$, $Ho \cap Hi = \phi$．

図1にコミュニティのリンク構造Gc(V,E)の模式図を示す．この例においては，V = H∪A ={a1,...,a6, h1,...,h5}である．Hoは，A内のページへのリンクよりもA外へのリンクを多く持つページ，Hi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，Hi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．

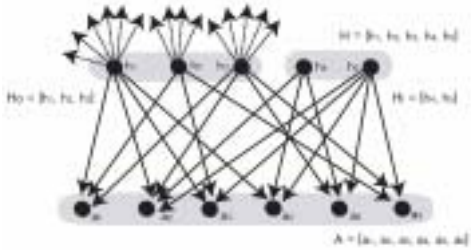

図 1 コミュニティのリンク構造 Gc(V,E)

Fig.1 Link structure of web communities, Gc(V,E)

ここで集合 Ho，Hi に属するページの性質を考える．Ho, Hi 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．Ho 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，Hi 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 = {a1, a2,...,am}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 Ho 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 Hi 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 Hi 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 Ho，純度を示す Hi を用いて，コミュニティのリンク構造の成長パターンを定義する．図 2 に定義と新規出現コミュニティにおける成長パターンごとのリンク構造の変化例を示す．

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．

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．

4. 実証分析

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．

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を基に構築したウェブグラフデータベースを用いて得られたコミュニティ集合を利用する．前述の新規出現コミュニティ定義を満たすコミュニティは，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あり，計4311のコミュニティを分析対象とする．これらの分析対象コミュニティにおいて，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．

- $|Ho(ti-1)|$: ti-1におけるコミュニティの認知度
- $|Hi(ti-1)|$: ti-1におけるコミュニティの純度
- $|Ho(ti)|$: tiにおけるコミュニティの認知度
- $|Hi(ti)|$: tiにおけるコミュニティの純度

4.1 実験結果

図2に一定の期間の前後で$|Ho|>|Hi|$と$|Ho|\leqq|Hi|$となるコミュニティの割合を示す．各期間における$|Ho|>|Hi|$の割合の平均値はti-1のとき69.87%，ti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，ti-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．

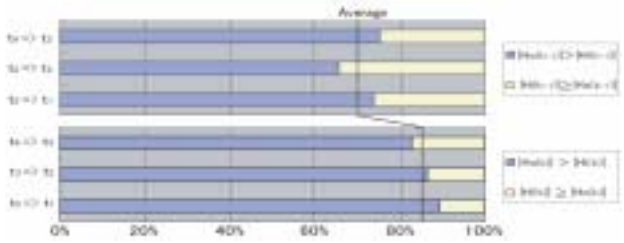

図 2 $|Ho|$及び$|Hi|$の比較

Fig.2 Comparison between $|Ho|$ and $|Hi|$

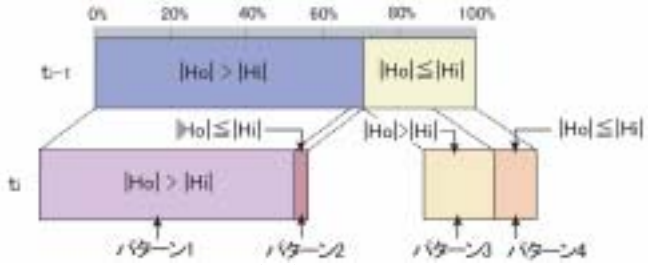

図 3 各成長パターン割合

Fig.3 Percentages of link evolution patterns

図3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．

表1,2に，各調査期間における各成長パターンの$|Ho(ti-1)|$, $|Hi(ti-1)|$(表1)，$|Ho(ti)|$，$|Hi(ti)|$(表2)の平均値，および，全体の平均値を示す．

表 1 各パターンにおける(a)$|Ho(ti-1)|$,(b)$|Hi(ti-1)|$平均値

Table1 Ave. of (a)$|Ho(ti-1)|$,(b)$|Hi(ti-1)|$

	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)
t2=>t1	5.5	0.3	2.4	0.8	0.3	0.5	0.4	3.6
t3=>t2	3.9	0.1	2.3	0.3	0.2	0.4	0.5	5.8
t4=>t3	19.8	0.2	2.1	0.4	0.3	0.6	0.7	3.1
Ave.	10.7	0.2	2.2	0.4	0.3	0.5	0.6	4.3

表 2 各パターンにおける(a)$|Ho(ti)|$,(b)$|Hi(ti)|$平均値

Table2 Ave. of (a)$|Ho(ti)|$,(b)$|Hi(ti)|$

	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)
t2=>t1	11.1	0.6	2.3	4.5	4.6	0.6	1.2	5.3
t3=>t2	11.2	0.7	2.6	4.2	5.8	0.9	2.4	19.6
t4=>t3	23.9	0.5	1.9	2.8	5.3	1.3	1.6	7.0
Ave.	16.4	0.6	2.3	3.6	5.5	0.9	1.9	12.4

4.2 考察

ここでは，特に顕著な成長を見せている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3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，Ho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 3 成長パターン 1 の実例
Table.3 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4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Hi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，Hi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Hi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 4 成長パターン 4 の実例
Table.4 An example of evolution pattern 4

1	http://...resources/cronolog/
2	http://.../.../Servers/Log_Analysis_Tools/
3	http://.../.../PostgreSQL/6.5/apache_php.html
4	http://.../Protocols/rfc2616/rfc2616.txt
5	http://.../Internet/SiteManagement/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. おわりに

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が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解析

[文献]

[1] A.Broder, R.Kumar et al.: ”Graph Structure in the web”. Proc. of 9th WWW (2000).

[2] R.Albert, H.Jeong et al.:”Diameter of the world wide web”. Nature, 401:130 (1999).

[3] B.E.Brewington and G.Cybenko.:”How dynamic is the web?”. Proc. of 9th WWW (2000).

[4] J.Cho and H.G.Molina.:” The evolution of the web and implications for an incre-mental crawler”. Proc. of 26th VLDB (2000).

[5] K.Bharat, B.W.Chang et al: ” Who Links to Whom: Mining Linkage between Web Sites”. Proc of IEEE 1st ICDM (2001).

[6] M.Toyoda and M. Kitsuregawa.: ” Extracting evolution of web communities from a series of web archives”. Proc of 14th Hypertext03 (2003).

[7] M.Toyoda and M.Kitsuregawa.:”Creating a web community chart for navigating related communities”. Proc. of 12th ACM Hypertext, pp.103-112 (2001).

[8] N.Imafuji and M.Kitsuregawa.:”Finding Web Communities by Maximum Flow Algorithm using Well-Assigned Edge Capacities”. IEICE Trans. on Inf. & Syst. Vol.E87-D No2 pp. 407-415. Feb (2004).

[9] R.Kumar, P.Raghavan et al.:”Trawling the web for emerging cyber-communities”, Proc. of the 8th WWW Conference, pp.1481-1493 (1999).

[10] G.Flake, S.Lawrence, and C.L. Giles.:”Efficient identification of web communities”, Proc. of 6th ACM SIGKDD KDD2000, pp 150-160 (2000).

[11] D.Gibson, J.Kleinberg et al.: ”Inferring web communities from link topology”, UK Conference on Hypertext, pages 225-234 (1998).

今藤 紀子 Noriko IMAFUJI

東京大学生産技術研究所産学官連携研究員. 2001 奈良女子大学大学院人間文化研究科博士後期課程修了，理学博士．グラフ理論に基づくウェブマイニング，ウェブ空間解析の研究に従事．情報処理学会正会員．日本データベース学会正会員.

喜連川 優 Masaru KITSUREGAWA

1983 年同大大学院工学系研究科情報工学博士課程修了, 工学博士．現在，東京大学生産技術研究所教授．平成 15 年 4 月より同所戦略情報融合国際研究センター長. データベース工学，並列処理，Web マイニングに関する研究に従事．本会理事，情報処理学会フェロー，SNIA-Japan 顧問，ACM SIGMOD Japan Chapter Chair(H11-H14)，電子情報通信学会データ工学研究専門委員会委員長(H9,10)．VLDB Trustee，IEEE TKDE Assoc. Editor, IEEE ICDE, PAKDD, WAIM Steering Comm. Member