

# Max-Flow コミュニティグラフと その特徴分析

## Constructing Max-Flow Community Graph and Its Feature Analysis

今藤 紀子<sup>▼</sup> 喜連川 優<sup>◆</sup>

Noriko IMAFUJI Masaru KITSUREGAWA

ウェブコミュニティとは、似たようなトピックを持つウェブページの集合を意味する。最大流アルゴリズムを利用して抽出したウェブコミュニティ（以降、Max-Flowコミュニティと呼ぶ）は、シードとして与えたページを中心として内側に密にリンクが存在するページの集合となる。これまでの研究で、同じシードページを与えるとHITS系手法によるコミュニティではトピックが一般的になり、Max-Flowコミュニティではより詳細になる傾向があることがわかった。本論文では、日本国内のウェブグラフデータを用いて、ウェブ空間全域から抽出したMax-Flowコミュニティをノード、コミュニティ間の関連をエッジとしたコミュニティグラフ（これをMax-Flowコミュニティグラフと呼ぶ）を構築し、HITS系手法を利用して得られたコミュニティに基づく既存のコミュニティグラフと比較し、それぞれの手法で得られたコミュニティの大域的な特徴差を分析する。

A web community is a set of web pages created by individuals or associations with a common interest on a topic. Maximum-flow algorithm extracts web communities which are sets of web pages having more links within the community than the outside. We call the web community *Max-Flow community*. Our previous work has characterized the properties of Max-Flow community (i.e. local properties) by comparing communities obtained by other methods and confirmed that while topics of communities by HITS and its derivations tend to be rather general, those of Max-Flow communities tend to be specific. In this paper, we analyze the global properties of Max-Flow community. First we construct *Max-Flow community graph* whose nodes and edges are Max-Flow communities and relevancies between the communities respectively. The next, we study the differences between Max-Flow community graph and an existing community graph based on HITS related methods.

### 1. はじめに

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し、ウェブ検索やウェブの成長分析、ウェブからのトレンド発見など幅広いウェブ技術への応用が期待できる概念として近年非常に重要視されている。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、エッジと見れば、ウェブは巨大な有向グラフを成している。コミュニティ特有のグラフ構造を解析することにより、コミュニティを効率よく発見するための種々の手法が提案されてきた[1,2,7]。また、HITS[3]などの関連ページアルゴリズムを利用してコミュニティを抽出する方法も存在する[4-6]。

我々は、これまでに、Max-Flowコミュニティと完全2部グラフによるコミュニティを比較しMax-Flowコミュニティの特徴を分析した[8]。また、既存のMax-Flowコミュニティ抽出手法の問題点を明示し、新たな抽出手法を提案した[9]。一方、豊田らによる研究では、日本国内のウェブグラフデータを用いて、HITS系手法によりウェブ空間全域からコミュニティを抽出し、コミュニティをノード、コミュニティ間の関連をエッジとしたコミュニティグラフを構築し、ウェブ空間に内在する興味深い様々なコミュニティ間の関連を浮かび上がらせることができた[10]。同じシードページを与えると抽出手法によりコミュニティは異なる特徴を持つ。それら個々の特徴の差は、コミュニティを巨視的に捉えた場合、どのような特徴差となって現れてくるかについてはまだわかっていない。本論文では、Max-Flowコミュニティを利用してMax-Flowコミュニティグラフを構築し、上述の既存のコミュニティグラフと比較する。互いのコミュニティグラフの相違点を明確にし、それぞれのコミュニティ抽出手法の相違から生じる大域的な特徴差を分析する。

### 2. Max-Flow コミュニティグラフの構築

Max-Flowコミュニティグラフとは、Max-Flowコミュニティをノードとし、関連するコミュニティ間に重み付きのエッジを張ったグラフである。エッジの重みは、類似度に基づくコミュニティ間の関連度を示すものとする。Max-Flowコミュニティグラフの構築は、ウェブ全域のMax-Flowコミュニティの抽出（ノード抽出）、及び、抽出したMax-Flowコミュニティ間のエッジ生成の2ステップで実現する。また、コミュニティに重複がある場合は、それらをマージしてグラフ表示する。

#### 2.1 Max-Flowコミュニティグラフ構成ノード

十分に多くのシードノードを用いてウェブ全域から抽出したMax-Flowコミュニティをグラフのノードとする。ここで言うウェブ全域とは、日本国内のウェブを意味する。以下にMax-Flowコミュニティの全域抽出の詳細を示す。

**利用データ:** 2002年2月にクロールした日本国内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。実際にクロールしたウェブページ数は、約4500万ページ、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の総数は約8400万、また総リンク数は、約3億7500万である。

**シードノード:** シードノードの入リンク数が少なすぎる場合、周辺グラフは極めて小さくなり、意味のあるMax-Flowコミュニティの抽出が期待できない。そこで、上記のウェブグラフデータベースにおいて、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をシードノードとして選択した。このとき総シードノード数は、約132万である。

Max-Flowコミュニティ抽出手法[9]を用いて、上記の各シードノードに対して、Max-Flowコミュニティを求める。ウェブ内から十分に多くのシードノードを利用しMax-Flowコミュニティを抽出するため、コミュニティのメンバーには重なり

<sup>▼</sup> 正会員 東京大学 生産技術研究所

[imafuji@tkl.iis.u-tokyo.ac.jp](mailto:imafuji@tkl.iis.u-tokyo.ac.jp)

<sup>◆</sup> 正会員 東京大学 生産技術研究所

[kitsure@tkl.iis.u-tokyo.ac.jp](mailto:kitsure@tkl.iis.u-tokyo.ac.jp)

が生じる。各コミュニティの上位20ノードについての重なり度を検証したところ、約8割のノードが単一のコミュニティに属しており、95%強のノードは、3つ以下のコミュニティに属していることがわかった。また、10以上のコミュニティに属するノードは、約2% (約55000ノード) 存在した。

## 2.2 Max-Flowコミュニティグラフ構成エッジ

Max-Flowコミュニティグラフでは、コミュニティ間の関連度をコミュニティの類似度で求め、コミュニティ間に重み付きのエッジを張る。コミュニティの類似度は、それぞれのコミュニティにおける上位 $n$ ノード (ただし、後述の例は全て $n=20$ ) の重複メンバー数で求める。エッジは無向であり、その重みの最大値は、 $n$ となる。図1は、属しているコミュニティ数が50以下のノードのみで得られたエッジに関して、その重み (横軸) と対応する重みのエッジ数 (縦軸) を示したものである (ただし、 $n=20$ )。グラフ右上の棒グラフは、重みが1から4およびそれ以上のエッジの割合を示している。重み1から4のエッジ数は、それぞれ、約669万、93万、34万、16万であった。また折れ線グラフは、重み5から20のエッジ数を示している。重みが大きくなるにつれ、コミュニティの類似度は増す。コミュニティ間の重みが20の場合は、両コミュニティの上位20ノードは完全に一致していることから、コミュニティのメンバーがほぼ全て重複していると予想される。

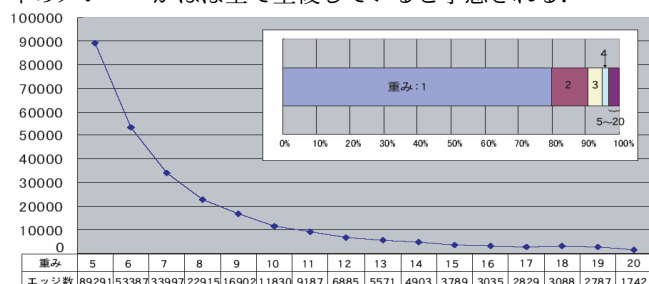


図1 エッジの重み (横軸) とエッジ数 (縦軸)

Fig.1 Edge weights and number of edges

## 2.3 Max-Flowコミュニティグラフ表示方法

現在のところは、着目したいコミュニティ近辺のグラフのみを取り出し、Max-Flowコミュニティグラフの部分グラフとしてそれを表示する。具体的には、以下の手順で行う。

- (1) URLを一つ指定し、そのURLが含まれているコミュニティセット  $C = \{C_1, C_2, \dots, C_n\}$  を取り出す。
- (2) さらに広範囲のグラフを検証する場合は、取り出した  $C_i \in C$  の各コミュニティを抽出したシードページ  $s_i$  が含まれているコミュニティセット  $C_{s_i} = \{C_{s_i1}, C_{s_i2}, \dots, C_{s_in}\}$  を取り出す。
- (3)  $C = C \cup C_{s_1} \cup C_{s_2} \cup \dots \cup C_{s_n}$  をベースノードセットとし (図2参照)、この中でメンバーの類似度が設定閾値以上のコミュニティ同士をマージする。

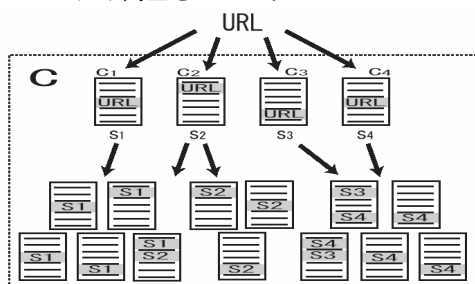


図2 Max-Flowコミュニティグラフのノード取得方法  
Fig.2 Finding nodes of Max-Flow community graph

## 3. Max-Flow コミュニティグラフの例

例1: 図3は、URL:www.miso.or.jp/ (味噌に関するサイト) を指定し、マージを行い2以上の重みを持つエッジで連結な部分を取り出したMax-Flowコミュニティグラフの例である。このURLが含まれているコミュニティは30存在し、マージステップ (閾値=0.6) により、抽出グラフ内のノード数は16となっている。Comm1は、醤油協会、醤油醸造など、Comm2は、仙台味噌、長寿味噌などの各地特産味噌、Comm3は、醤油、納豆、大豆、味噌など、Comm4は、神の舞、出羽桜などの酒造、Comm5は、マルヤ味噌などの一般的な味噌メーカー、Comm6は、飲食店、料理のレシピなどのページ群とそれぞれなっている。このように指定URLの「味噌」というトピックと現実的にも関わりが深いと思われるようなコミュニティ群が、実際にコミュニティグラフ上でも互いに関連していることが確認できる。

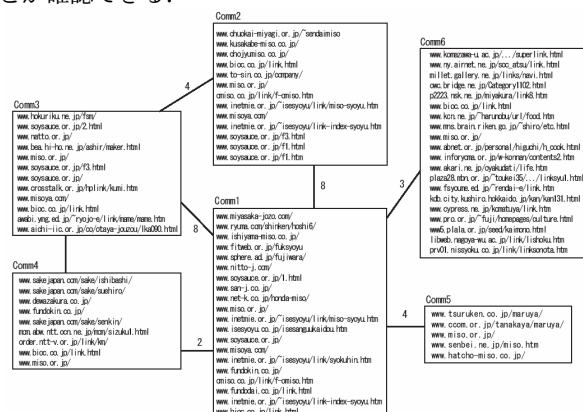


図3 Max-Flowコミュニティグラフの例1

Fig.3 Example1: Max-Flow community graph

例2: URL:www.jva.or.jp/ (バレーボール協会のサイト) を指定する。コミュニティの2段階抽出をした後のコミュニティ数は122となっており、そのうち74のコミュニティが閾値0.6でマージされ16のノードとなっている。図4は、マージされた16ノードについてのグラフを表している。表中の数字が一つのコミュニティを意味しており、右表が該当コミュニティのメンバーにおけるキーワードとなっている。また、色づけられたノードは、指定URLが含まれているコミュニティである。図中点線で示されたエッジのみが重み1であり、その他は、重み2以上のエッジのみを表示している。指定URLが含まれている6つのコミュニティは、大小様々な規模 (社会人チーム、地域のチーム、大学のチーム等) のバレーボールチーム、バレーボールに関する総合情報サイト等からなるコミュニティである。Comm8にバレーボールとスィミングが共存するスポーツ全般に関するコミュニティを介して、Comm14, Comm16と水泳関係のコミュニティが抽出されている。また、薄い繋がりではあるが、ラグビーに関するコミュニティ群が、バレーボールコミュニティの近辺に存在することがわかる。

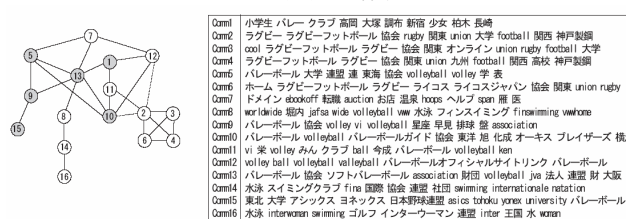


図4 Max-Flowコミュニティグラフの例2  
Fig.4 Example2: Max-Flow community graph

#### 4. コミュニティグラフの特徴比較

本節では、既存のコミュニティグラフとMax-Flowコミュニティグラフを比較し、Max-Flowコミュニティの有する特徴を分析する。

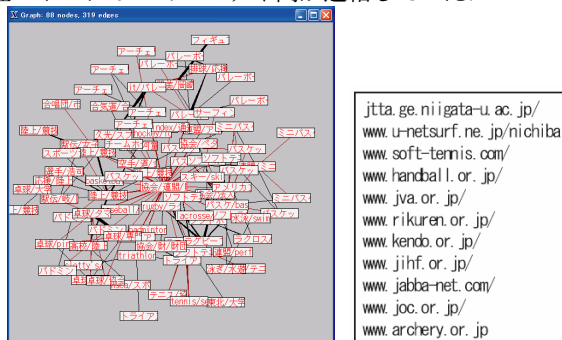
#### 4.1 既存コミュニティグラフ概要

豊田らによるコミュニティグラフ([10])においては、ウェブコミュニティチャートと呼んでいる)では、HITSを発展させた関連ページアルゴリズムが利用されている。このコミュニティグラフにおいても、ノードは、コミュニティであり、重み付きエッジは、コミュニティ間の関連度を意味している。ただし、Max-Flowコミュニティグラフとは異なり、エッジは有向である。コミュニティグラフの構築手順の詳細に関しては[10]を参照されたい。尚、比較の際に利用したコミュニティグラフは、前節で述べたMax-Flowコミュニティグラフ構築の際に利用したデータと同一のものにより作成されている。

比較には、ウェブコミュニティブラウザツール[11]により表示したグラフを用いる。ノードは、コミュニティを示しており、アンカーテキストを基に生成されたラベルが張られている。このツールでは、URLを入力するとそのURLが含まれているコミュニティが画面上に現れる。厳密には、指定URLと一部分が一致するURLが含まれているコミュニティも現れるが、実際に指定URLが含まれているコミュニティのみを残す。まず、そのコミュニティの周辺コミュニティを追加表示する機能（in-link・out-link展開）を用いて周辺コミュニティを全て表示する。表示されるノード（コミュニティ）が多すぎる場合は、ノードの大きさ、エッジの重みによる閾値操作により適度にノードおよびエッジを除去する。

## 4.2 コミュニティグラフ比較例

**比較例 1**：前節例2と同様、URLにwww.java.or.jp（バレーボール協会サイト）を指定する。既存コミュニティグラフにおいては、各種スポーツ協会のコミュニティに属している。図は、このコミュニティをin-link展開し、エッジの重みの閾値を5、コミュニティサイズを10以上に設定して表示させたグラフと、指定URLが含まれているコミュニティメンバーのURL（右表）を示している。一方、前節(図4)で示したように、Max-Flowコミュニティグラフでは、各種スポーツ協会のページが集まったコミュニティは存在せず、指定URLは複数のバレーボールに関するコミュニティに含まれていた。既存コミュニティグラフでは、指定URLが含まれているコミュニティを介して他の種々のスポーツに関するコミュニティが連結されているが、Max-Flowコミュニティグラフでは、色々なスポーツ全般のリンク集や、それらのページからリンクされているページからなるコミュニティ（例えば、Comm8）を介して異種スポーツのコミュニティ間が連結していた。



**図 5 既存コミュニティグラフ: [www.java.or.jp](http://www.java.or.jp)**  
**Fig.5 Web community chart: [www.java.or.jp](http://www.java.or.jp)**

**比較例2：**URLにwell-mannered.org/（個人が設立している犬を飼う際の情報サイト）を指定する．図6は，マージ操作をせずにエッジの重さが2以上部分を抽出したMax-Flowコミュニティグラフである．ただし，各コミュニティはそれぞれ5つずつURLを表示し，右表は，これまでと同様，該当コミュニティのアンカーテキストを基にして抽出したキーワードを示している．犬の飼い主の私的サイト集（Comm2，Comm5，Comm7，Comm8），犬の総合情報サイト・リンク集（Comm3，Comm4，Comm6）その他犬関連のページ集（Comm1，Comm6）等がコミュニティ群として抽出された．一方，既存コミュニティグラフでは，指定URLが種々の犬情報サイトから成るコミュニティに属している．図7は，このコミュニティをin-link展開し，エッジの重みを10，コミュニティサイズを10以上に設定して表示させたグラフと，指定URLが含まれているコミュニティメンバーのURL（右表）を示している．Max-Flowコミュニティグラフでは，複数のコミュニティにまたがって存在していたページが一つのコミュニティに集約されており，色々なペット，動物，犬の雑貨等関するコミュニティが周辺に存在している．

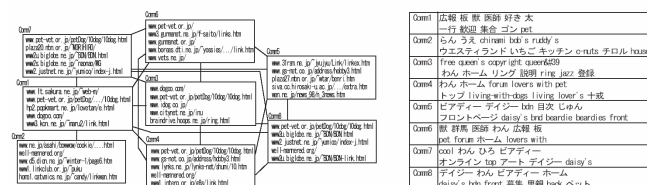


図 6 Max-Flow コミュニティグラフ:well-mannered.org  
Fig.6 Max-Flow community graph:well-mannered.org

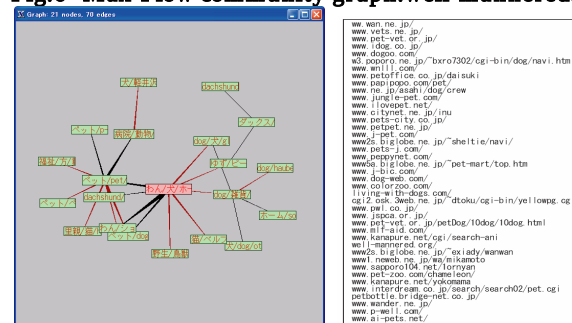


図7 既存コミュニティグラフ:well-mannered.org  
Fig.7 Web community chart:well-mannered.org

### 4.3 考察

比較例1より、トピックAに関して極めて重要度が高いページB（例えば、トピック「バレーボール」に対してバレーボール協会、トピック「倉敷観光」に対して倉敷市サイト内の観光ページ等）が含まれるコミュニティは、Max-Flowコミュニティグラフでは、トピックAと階層的に関連しているページ集合（例えば、「バレーボール」に対して、地域のバレーボールチームのページなど、「倉敷観光」に対して、私的な倉敷紹介サイト、倉敷市所在の公園に関するサイトなど）となっており、それらがグラフの一部分を形成していた。一方、既存コミュニティグラフでは、Bと並列的に関連しているページ（例えば、バレーボール協会のページに対して、その他各種スポーツの協会・団体ページ、倉敷市サイト内の観光ページに対して、市町村が運営する観光ページ）が、一つのコミュニティを形成していた。このことから、Max-Flowコミュニティグラフからは、一つのトピックに関して、階層的なコミュニティの分布を検証するのに適していると言える。

比較例2のように、既存コミュニティグラフでは単一のコミ



コミュニティに属しているページ群が、Max-Flowコミュニティグラフでは、複数のコミュニティに分散しグラフの一部を形成している場合が多く見られた。また、前述のように比較例1の既存コミュニティグラフにおいて指定URLが含まれていたコミュニティは、各種スポーツ協会・団体のページであったが、Max-Flowコミュニティグラフでは、そのようなコミュニティは存在しない。しかしながら、各種スポーツリンク集と共に種々のスポーツのページが集まったコミュニティは存在する。これらのことから、既存コミュニティグラフに比べてMax-Flowコミュニティグラフは、細かい粒度のページ集合となっていると言える。両コミュニティグラフの関係は模式的に図8のように表すことができる。

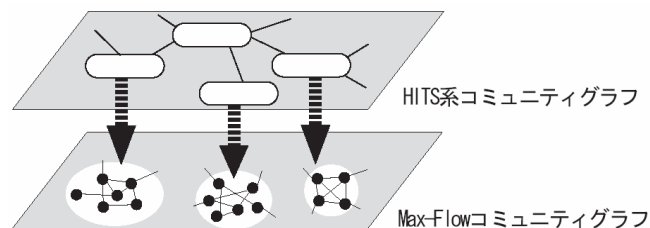


図8 コミュニティグラフの相違

Fig.8 Difference between community graphs

また、Max-Flowコミュニティグラフでは、全く関連のないと思われるトピックのコミュニティ同士が近隣コミュニティとして、抽出されることがある。これらは、それらのトピックを結びつけるリンクが偶然的に何らかの理由であるページに張られることにより生じる。実際に確認できた例としては、あるスポーツ選手のオフィシャルホームページとそのページの制作会社を結ぶリンクにより、そのスポーツに関するコミュニティとネット系デザイン会社のコミュニティが隣接していた。コーヒーのコミュニティの近辺にJAVA言語に関するコミュニティが存在しているといった興味深い例も確認できた。よって、Max-Flowコミュニティグラフでは、表だってその存在がわからないようなトピック間の隠れた関係を導き出すことができる。

## 5. おわりに

本論文では、最大流アルゴリズムを用いて抽出したウェブコミュニティをMax-Flowコミュニティと呼び、Max-Flowコミュニティをノードとし、コミュニティ間の類似度に基づく重み付きエッジを張った、Max-Flowコミュニティグラフを構築する方法を述べた。また、豊田らによるHITS系手法を利用して得たコミュニティをノードにした既存のウェブコミュニティグラフと比較し、Max-Flowコミュニティグラフが有する特徴を示した。これまでの研究では、同じシードを与えると、Max-Flowコミュニティは、シードトピックよりもより詳細なトピックになり、逆に、HITS系手法によるコミュニティでは、より一般的になるといった、一つ一つのコミュニティレベルでの特徴の相違が明らかになっていた。本研究により、それらのコミュニティを巨視的に見た場合、Max-Flowコミュニティグラフは、HITS系手法によるコミュニティグラフに比べて、細かい粒度のグラフとなっていることがわかった。

現在のところ、コミュニティ間のエッジには、コミュニティの重複に基づく類似度を利用しており、重複のあるコミュニティに関するグラフのみが得られるようになっている。しかしながら、この方針では、ウェブグラフ上では隣接関係にあるようなコミュニティであっても、重複がなければ表示することはできない。よって、基のウェブグラフにおけるリンク

関係を反映させるなどして、ウェブグラフにおいて周辺にあるコミュニティをも抽出可能にする方法を考えなければならない。それと同時に、より簡単に、Max-Flowコミュニティグラフを表示し特徴を検証するために、URLの指定、キーワード及びグラフの表示などの一連の操作を可能にするツールを開発することが今後の大きな課題である。

## [文献]

- [1] G.Flake, S.Lawrence, and C.L. Giles.: "Efficient identification of web communities", Proc. of 6th ACM SIGKDD KDD2000, pp 150-160 (2000).
- [2] 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).
- [3] J.M.Kleinberg: "Authoritative Sources in a Hyperlinked Environment", Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, pp.668-677 (1998).
- [4] D.Gibson, J.M. Kleinberg, and P.Raghavan.: "Inferring web communities from link topology", UK Conference on Hypertext, pp.225-234 (1998).
- [5] S.Chakrabarti, B.Dom, D.Gibson et al.: "Automatic resource list compilation by analyzing hyperlink structure and associated text", Proc. of the 7th WWW Conference, pp.65-74 (1998).
- [6] R. Kumar, P.Raghavan et al.: "Extracting large-scale knowledge bases from the web", VLDB Journal, pp. 639-650 (1999).
- [7] R.Kumar, P.Raghavan et al.: "Trawling the web for emerging cyber-communities", Proc. of the 8th WWW Conference, pp.1481-1493 (1999).
- [8] N.Imafuji and M.Kitsuregawa.: "Effects of maximum flow algorithm on identifying web community", Proc. of the 4th WIDM Workshop, pp 43-48 (2002).
- [9] 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).
- [10] M.Toyoda and M.Kitsuregawa.: "Creating a web community chart for navigating related communities". Proc. of 12th ACM Hypertext, pp.103-112 (2001).
- [11] 福地健太郎, 豊田正史, 喜連川優.: "Web Community Browser: Webコミュニティ構造の可視化と探索機構の実現", FIT 2002 論文集 (2002).

## 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

1978 年東京大学工学部電子工学科卒。1983 年同大学院工学系研究科情報工学博士課程修了，工学博士。同年同大生産技術研究所講師，現在，同教授。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，情報処理学会理事・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