論　文

ウェブコミュニティチャート: 膨大なウェブページを関連する話題を通して閲覧可能にするツール

豊田　正史[†]　　吉田　聡[†]　　喜連川　優[†]

Web Community Chart: a Tool for Navigating Numerous Web Pages by Related Topics

Masashi TOYODA[†], Satoshi YOSHIDA[†], and Masaru KITSUREGAWA[†]

あらまし　同じ話題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自動的に抽出できるリンク解析の手法が，これまでに数多く提案されている．ウェブコミュニティと呼ばれるこれらの集合を用いて，ウェブ上に存在する話題を把握することが可能である．しかし既存の研究では，個々のウェブコミュニティの抽出に重きが置かれ，ウェブコミュニティ間の関連に付いては考慮されてこなかった．本論文では，大規模なウェブアーカイブからウェブコミュニティを抽出し，関連するウェブコミュニティ同士を辺で結んだグラフを作成する手法を提案する．これをウェブコミュニティチャートと呼ぶ．我々の手法は，与えられたシードページに関連するページをリンク解析によって算出する関連ページアルゴリズムに基づいている．まず我々は，既存の関連ページアルゴリズムの精度を改善して，ユーザテストによる評価を行う．次に，改良版アルゴリズムを用いて，大規模な日本のウェブアーカイブからウェブコミュニティチャートの作成実験を行う．さらに，完成したチャートについてウェブディレクトリとの比較を行い，チャートの特徴を示す．

キーワード　ウェブ，リンク解析，ウェブコミュニティ

1. はじめに

ウェブにおけるリンク解析の研究により，同じ話題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自動的に抽出できることが明らかになってきている．これらの集合はウェブコミュニティと呼ばれ，野球チームのファンページの集合や，コンピュータメーカの公式ページの集合などが例として挙げられる．既存のリンク解析手法 [3], [5], [6], [8], [9] は，ウェブページを頂点とし，ハイパーリンクを有向辺とした大規模な有向グラフとしてウェブをとらえてグラフの構造解析を行うことで，ウェブコミュニティを抽出する．これらの手法を用いて，我々は興味のある話題に関するウェブコミュニティの存在を知ることができ，中に含まれるページからその話題に関する様々な情報を収集できる．しかし既存の手法では，ウェブコミュニティ間の関連については扱われていないため，ある話題に関して他にどのような種類のウェブコミュニティがどのくらいの数存在するかを知ることはできなかった．

我々の目的は，個々のウェブコミュニティを抽出するだけではなく，ウェブコミュニティ間の関連度を算出して，関連するウェブコミュニティの間を渡り歩けるグラフを作成することにある．このグラフを我々は，ウェブコミュニティチャートと呼ぶ．チャートは次のような状況で有用である．例えば，これからコンピュータを選んで買おうとしているユーザが，様々なメーカのページで仕様を調べたり，様々なオンラインショップで値段を比較しようと考えているとする．ウェブコミュニティチャートでは，コンピュータメーカや，オンラインショップなどのウェブコミュニティが強い関連度を持って結合されているため，ユーザは容易にこれらの間を渡り歩きながら情報を集めることができる．

本論文では，大規模なウェブのアーカイブから，ほぼ全てのウェブコミュニティを抽出し，ウェブコミュニティチャートを作成する手法を提案する．本手法では，ウェブコミュニティを抽出し，その間の関連度を算出

[†] 東京大学生産技術研究所，東京都
Institute of Industrial Science, University of Tokyo,Komaba 4–6–1, Meguro-ku, Tokyo, 153–8552 Japan

するために，関連ページアルゴリズム (RPA: Related Page Algorithm) と呼ばれるリンク解析手法を用いている．RPA は，入力としてあるページを与えられると，その周辺のグラフで密に結合されているページを関連ページとして出力する．本手法の基本的なアイデアは，分類したいページ全てに RPA を適用し，各ページから導出される関連ページのリストを特徴量としてページの分類を行うことにある．

我々は，まず既存の RPA を改良し，ユーザテストによって精度の評価を行う．次に，大規模な日本のウェブアーカイブを用いて，ウェブコミュニティチャートの作成実験を行う．さらに，作成したチャートとウェブディレクトリとの簡単な比較を行い，チャートの特徴を示す．

以降，まず 2. では関連研究を列挙し，本論文の位置付けを明らかにする．次に，3. で，ウェブコミュニティチャート作成手法の直観的な解説を行う．4. では，RPA の改良とその評価を行い，5. では，チャート作成手法の詳細な説明を行う．6. では，大規模な日本のウェブアーカイブを用いてチャート作成実験を行い，その結果を示す．7. でまとめと今後の課題を述べる．

2. 関 連 研 究

ウェブコミュニティに関する研究の多くは，Kleinberg [7] によって提案されたオーソリティおよびハブの概念に基づいている．オーソリティとは，ある話題について良質な内容を持つページのことを指し，多くの良いハブからリンクを張られているページと定義される．ハブは，ある話題に関するリンク集やブックマークページのことを指し，多くの良いオーソリティにリンクを張っているページと定義される．HITS [7] は，この循環した定義に基づいて，ウェブグラフからオーソリティおよびハブを効率良く抽出するアルゴリズムである．図 1 に，HITS によって抽出される典型的なグラフ構造の例を示した．オーソリティは IBM など大手のコンピュータメーカのページである．これらのページはコンピュータメーカリンク集などのハブによって密に結合されている．このように，HITS はオーソリティおよびハブから成る密な 2 部グラフ構造を抽出する．ウェブコミュニティは，実社会のコミュニティとは若干意味が異なる．HITS では，互いに知らない著者の集合，および競合企業の集合が抽出され得るからである(注1)．

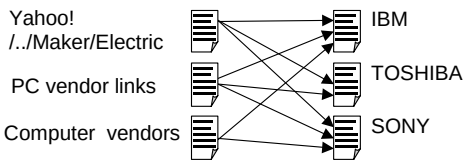


図 1　オーソリティおよびハブからなる典型的なグラフ
Fig. 1　Typical graph of hubs and authorities

HITS には，アンカーテキスト，リンクへの重み付け，および文書構造などを用いた様々な改良が施されてきている [1]〜[3]．また Dean らは，HITS を応用して，与えられたページに対し関連するページ群を結果として返す，Companion [4] という関連ページアルゴリズム (RPA) を開発した．Companion は，与えられたページ周辺のグラフからオーソリティを抽出して関連ページとして返す．我々が用いる RPA はこの Companion を改良したものである．

ウェブコミュニティに関する主な研究 [3], [6], [9] は，オーソリティおよびハブの集合をコミュニティとして扱っている．Gibson らは HITS で得られるコミュニティの性質について調査を行っている [6]．Chakrabarti らは，HITS の精度をアンカーテキストを用いて改善し，コミュニティの評価を行っている [3]．Kumar らは，大規模なウェブのアーカイブに対して，trawling と呼ばれる手法を適用し 10 万を越えるコミュニティのコアを発見した [9]．コアとはオーソリティとハブから成るサイズの小さい完全 2 部グラフであり，ほとんどのコミュニティはコアを含むという仮定に基づいている．我々のチャート手法は，Companion アルゴリズムを用いているが，グラフ全体を解析してコミュニティをリストアップする点では trawling と類似している．

また，HITS とは異なるモデルに基づいたウェブコミュニティ抽出方法も幾つか提案されている．Lempel らは，ランダムウォークモデルを用いてオーソリティを抽出する SALSA [8] アルゴリズムを提案しており，Flake らは [5] において，最大フローおよび最小カットを用いたウェブコミュニティ抽出法を提案している．

グラフに基づいたコミュニティ抽出法は，キーワードを用いた文書の類似度に基づくクラスタリングに比べて，高速に多量のウェブコミュニティを抽出できることに特徴があり，巨大なウェブを対象とするのに

(注1)：以降,「コミュニティ」は「ウェブコミュニティ」の意味で使用する．

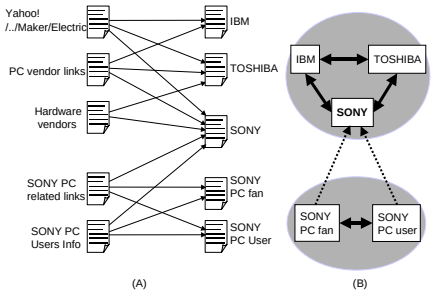


図 2　RPA による導出関係の例
Fig. 2　An example of derivation relationships

適している．しかし，現時点では個々のウェブコミュニティを抽出することにのみ焦点が当てられており，ウェブコミュニティ間の関連についての研究は少ない．我々の提案するウェブコミュニティチャートは，互いに関連するウェブコミュニティの閲覧を可能とするものである．

3. ウェブコミュニティチャートの概要

ウェブコミュニティチャートは，ウェブコミュニティを頂点とし，コミュニティ間の関連度を重み付きの辺で表したグラフである．本節では，まずチャート作成に用いる手法の直観的な説明を行う．本手法は，分類を行いたいページ (シードページ) の集合を入力として受け取り，チャートを出力する．基本的なアイデアは，与えられたシードページ全てに関連ページアルゴリズム (RPA) を適用し，各ページが他のページをどのように導出するかを調べる事にある．

図 2 に，RPA による導出関係の例を示す．この図では，IBM，TOSHIBA，SONY，および 2 つの SONY PC ファンページをシードページとして，それらの間の導出関係を図示している．グラフ (A) は各シードページとハブページのリンク関係を示し，グラフ (B) は各シードが他のページを関連ページとして導出する様子を表している．

IBM，TOSHIBA，および SONY は，互いに他のページを RPA によって導出する関係にある．これらのページは主に電器メーカリンク集等からリンクされているためである．グラフ (B) において，これらのページは対称的にお互いを導出する関係にある．

一方，2 つの SONY PC ファンは，互いに相手を導出しさらに，SONY を導出する．これらのページが SONY PC に関するリンク集から主にリンクされているためである．電器メーカのリンク集の数に比べて，SONY PC 関係のリンク集の数は少ないため，SONY PC ファンは SONY を上位の関連ページとして導出するが，SONY は SONY PC ファンを上位の関連ページとしては導出しないことになる．

グラフ (B) から，対称的な導出関係にあるページ同士には強い関連性があり，非対称的な導出関係にあるページ同士は関連性が弱いことが見て取れる．この観察に基づき，我々は対称導出関係で密に結合されたページの集合をウェブコミュニティとして扱い，2 つのウェブコミュニティのメンバー間に非対称導出関係がある場合に，これらのウェブコミュニティが関連しているとみなす．

4. 関連ページアルゴリズムの精度改善

ウェブコミュニティチャート作成手法は，3. で述べたように，関連ページアルゴリズムを基本的な構成要素として用いている．本節では，既存の関連ページアルゴリズムである Companion と我々が改良を施した Companion–を解説し，ユーザテストによって精度の比較を行う．

4.1 関連ページアルゴリズム

関連ページアルゴリズム (RPA) は，シードページ (複数でも可) を入力とし，上位 N 個のオーソリティおよびハブページを出力する．RPA はまず，ウェブグラフからシードページ周辺の部分グラフを抽出する．これを近傍グラフと呼ぶ．次に，この近傍グラフにおいてミラーページの削除および各辺へのウェイト付加を行う．最後に，近傍グラフの各ページについてオーソリティスコアおよびハブスコアの計算を行い，高いオーソリティスコアを持つページを関連ページとして返す．以下の節では，各手順の詳細について解説する．

4.1.1 近傍グラフの作成

シードページ周辺の近傍グラフを作成する．両手法とも，ウェブサーバ間 (ホスト間) に張られているリンクのみを考慮し，サーバ内で張られているリンクは無視する．作成方法は，以下のように異なる．

Companion では，シードページからリンクを逆にたどり，そこから順方向にリンクをたどる間に通るページの集合 (BF: Back Forward set)，および，シードページから出ているリンクをたどり，そこから逆にリンクをたどる間に通るページの集合 (FB: Forward Back set) をグラフに含める．ただし，BF で順方向にリンクをたどる際には，ページ内でのリンクの出現順

序を考慮して，シードページを指しているリンクに近い位置にあるリンクのみをたどる．実際には，シードページを指しているリンクから上下 R 個，合計 $2R+1$ 個のリンクをたどる．これは，近い場所にあるリンクは関連があるページを指しているというヒューリスティックに基づいている．また，BF および FB に含まれるページ間のリンクを近傍グラフに含める．ただし，BF に関しては，作成の際にたどったリンクのみを採用する．

Companion–では，BF のみを使用する．Companion においては，FB 中に Yahoo! のような被リンク数の多いページが頻繁に含まれるため，そのようなページにオーソリティスコアが集中して異なる話題のページが出力される．HITS に基づくアルゴリズムにおいて頻繁に起るこの現象は，トピックドリフトと呼ばれている．Companion–では，トピックドリフトを防ぐために FB を近傍グラフから削除した．

どちらの手法においても，あるページからの被リンク数が $MaxIn$ を越える場合には，ランダムに $MaxIn$ 個の被リンクを選び，それだけをたどる．本実験では，R として 10，$MaxIn$ として 2000 を使用している．R は，大きすぎると上記のトピックドリフトを起こし易く，小さすぎると 2 部グラフ構造が現れ難くなり結果が不安定になる．10 付近では上位の結果はおおむね安定しているため，この値を用いた．$MaxIn$ は，十分に大きければ結果に与える影響は少なく，2000 以上であれば十分である．

さらに，作成した近傍グラフからミラーページおよびミラーに近いページを削除する．ミラーの判定は，trawling [9] と同じ方法で行っている．詳しくは [9] を参照されたい．

4.1.2 各リンクへのウェイト付加

近傍グラフに含まれる各リンクに，0〜1 の範囲でオーソリティウェイト ($auth\_wt(m,n)$) およびハブウェイト ($hub\_wt(m,n)$) を与える．オーソリティウェイトは，そのリンクが指すページに対してどの程度のオーソリティスコアを与えるか，に影響する．ハブウェイトは，リンク元のページに対してどの程度のハブスコアを与えるか，に影響する．

Companion および Companion–では，基本的にはオーソリティおよびハブウェイトは 1 である．ただし，あるページが同じサーバにある n 個のページから指されている場合，該当するリンクのオーソリティウェイトを $1/n$ する．また，あるページが同じサーバにある n 個のページを指している場合，該当するリンクのハブウェイトを $1/n$ する．これは，1 つのサーバが大きな影響を持ちすぎないようにするためである．

4.1.3 オーソリティおよびハブスコアの計算

各ページ n のオーソリティスコア ($auth(n)$) とハブスコア ($hub(n)$) は，以下の手順で算出する:

（1） 全てのページの $auth(n)$, $hub(n)$ を 1 とする．

（2） スコアが収束するまで以下を繰り返す．

$auth(n) = \sum_{m(m \Rightarrow n)} auth\_wt(m,n) hub(m)$

$hub(n) = \sum_{m(n \Rightarrow m)} hub\_wt(n,m) auth(m)$

($x \Rightarrow y$ は x から y へリンクがあることを示す．)

全 $auth(n)$ の二乗和が 1 になるよう正規化する．

全 $hub(n)$ の二乗和が 1 になるよう正規化する．

（3） $auth(n)$ の高い順に N ページを出力する．

4.2 評価実験

Companion と，Companion–の精度を比較するため，ユーザテストを行った．このテストには，被験者として助教授，助手，ポスドク，および学生を含む 10 人が参加した．すべての被験者はウェブを日常的に利用している．

データセットとしては，1999 年の夏に収集した 1700 万ページからなる日本のウェブアーカイブ (主に jp ドメインに存在するページの集合) を使用した．これらのページは国内の著名なページを出発点に，幅優先探索の順番で収集した．高速に関連ページアルゴリズムを実行するために、本アーカイブを基にウェブグラフのデータベースを作成して使用した．グラフデータベースは，約 1 億 2 千万のリンクを含み，約 3000 万のページを含んでいる (1700 万のアーカイブに存在するページと，それらから指されているがアーカイブに含まれていない 1300 万ページ)．各ページの被リンク数の分布は，べき乗分布となっており，べき指数は約 2.2 となった．この結果は, [9] で発表されている全世界のアーカイブにおける結果 (べき指数 2.1 のべき乗分布) とほぼ等しい．べき指数が 0.1 高く出ているのは，日本国内ではリンクの密度が比較的薄いことを示している．

本実験では，まず被験者から,「過去に、ある話題に沿って集めたことのあるウェブページ」または「関連する情報を集めたいと思うウェブページ」を募集した．各被験者から，1〜4 個の別々な話題に関するシードページを回収し，合計 24 ページを得た．次に，各シードページに，Companion，および Companion– を適用し，上位 10 ページのリストを 2 種類作成した．各被

URL of seed pages	Short description	#inlinks	Companion	Companion–
weather.is.kochi-u.ac.jp/	Kochi Univ., Weather Home	1205	5/10	9/9
www.watch.impress.co.jp/pc/index...	PC Watch	1056	6/10	9/10
www.peugeot.co.jp/	Official Peugeot Japan	423	10/10	10/10
www.mahjong.or.jp/	Mahjong Walker	168	0/10	9/9
www.maccentral.or.jp/pokemon/	Pokemon site	164	9/10	10/10
www.ops.dti.ne.jp/~glass/	Stock market information	104	7/8	10/10
www.red-hell.com/	Urawa Reds (a soccer team) fan page	113	10/10	10/10
www.i-kochi.or.jp/prv/kochi/	Kochi Prefecture Information	109	2/10	8/8
www.japan.msf.org/	Medicines Sans Frontiers Japan	85	8/8	9/9
www2j.biglobe.ne.jp/~tatuta/	Free market information	71	0/10	9/10
www.panda.org/	WWF International	61	10/10	9/9
www.tintin.com/	Airline mileage service information	51	9/9	8/8
lang.nagoya-u.ac.jp/~matsuoka/Japan...	A Guide to Japan	43	0/10	7/8
www.spice.or.jp/~mt0711/index.html	Overseas travel Information	33	8/8	10/10
www.mars.dti.ne.jp/~o-shin/	Relational Database Information	26	8/10	10/10
www.triathlon.or.jp/	Triathlon World	26	10/10	10/10
www.alc.co.jp/nihongo/nihongo1.html	Japanese Language Center	23	10/10	8/10
plaza.harmonix.ne.jp/~kamao/	Virtual domain service information	18	2/10	7/9
www.isp.ne.jp/~nakajima/index.html	Movie information	15	2/10	8/8
islamcenter.or.jp/	Islamic Center Japan	12	7/10	7/10
www2e.biglobe.ne.jp/~TKG/	Puzzle information	8	0/10	8/8
www3.famille.ne.jp/~s370902/camera/...	Camera information	4	1/9	4/10
archives.math.utk.edu/popmath.html	POP Mathematics	1	7/9	7/10
home.att.ne.jp/green/asj	The Africa Society of Japan	1	8/10	7/10
Average precision			0.61	0.91

表 1　アクセスできたページ数に対する関連ページの数
Table 1　# of related pages to # of accessible pages

験者には，提出したシードページに対応する結果リストを返却し，結果の評価を依頼した．具体的には，結果の各ページの内容をウェブブラウザで閲覧し，そのページにアクセスできた場合には，内容がシードページに関連する話題を持っているかを主観的に判断するよう求めた．

表 1 に，被験者による評価結果を示す．この表には，シードページの URL および簡単な説明が，被リンク数が多い順に並べられており，各シードページに対して 2 種のアルゴリズムの精度が示されている．ここで精度とは，シードに関連すると判断されたページの数を，アクセスできたページの数で割った値である．

表 1 から分かるように，多くのシードは 10 以上の被リンク数を持つ著名なページである(注2)．このため，この実験の結果は，ウェブコミュニティチャートの質に対して大きな影響を持つことになる．

ほとんどのシードに関して，Companion–が，Companion に比べてより良い結果を出力しており，平均では約 0.9 の精度を示している．この精度向上は，近傍グラフから，トピックドリフトを起こしやすい部分を削ったことから得られたものである．Companion はしばしばトピックドリフトを起こし，精度を落としている．

Companion は，Yahoo!などの非常に有名な 20URL をストップ URL リストに登録しておき，近傍グラフ作成の際にこれらを取り除くことになっているが，Companion–はストップ URL を使用しない．このため，今回の実験では双方ともストップ URL リストを使用しない条件での比較を行った．このため，Companion の精度は本来の精度よりも落ちていることに注意されたい．Companion でトピックドリフトを起こしている 9 件のケースのうち 6 件まではストップ URL を使うことで防ぐことが可能である．しかし，ドリフトを起こす URL が必ずストップ URL に含まれているとは限らないため，ストップ URL には限界がある．実際，残りの 3 件についてはストップ URL に含まれるほど有名ではないページがドリフトの原因になっている．このため，ストップ URL を採用したとしても，Companion– は Companion より良い精度を示す．また，Companion–はストップ URL を使わずに良い精度を得られる点で Companion より優れていると言える．

(注2)：我々のアーカイブでは，10 以上の被リンク数を持つページは全体の 5%しか存在しない

5. ウェブコミュニティチャート作成手法

本節では，大規模なウェブアーカイブからコミュニティチャートを作成する手法を説明する．

5.1 シードページ集合の選択

まず，チャートによって分類されるべきページを，シードページ集合として選択する．シードページ集合は，ウェブ上で評価の高いページをできるだけ網羅したものとしたい．評価の高さは，ページの被リンク数を用いてある程度判断できるため，これを利用した．本手法では，全てのページについて被リンク数を調べ，異なるサーバ (ホスト) から IN 本以上のリンクが張られているページを，シードとして選択する．こうして得られたシードページ集合を S とする．

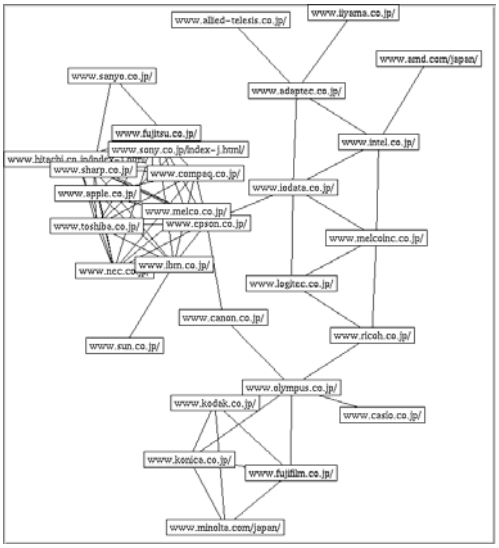


図 3　SDG 中の連結成分
Fig. 3　A connected component in SDG

5.2　オーソリティ導出グラフの作成

次に，各シードページ間の関連ページアルゴリズム (RPA) による導出関係を表す有向グラフを作成する．これをオーソリティ導出グラフ (ADG: Authority Derivation Graph) と呼ぶ．ADG の頂点集合は，シードページ集合 S に等しい．ADG において頂点 s から t への有向辺があるとき，s が t を Companion– によって上位 N 個以内のオーソリティとして導出することを示す．以下に ADG の作成手順を示す．

- 各シードページ $s \in S$ に Companion–を適用し，上位 N 個のオーソリティページ A_s を算出する．
- A_s に含まれる各オーソリティページ t について，$t \in S$ ならば，s から t への有向辺を作成する．

5.3　対称導出グラフの抽出

ADG から，対称導出グラフ (SDG: Symmetric Derivation Graph) を作成する．SDG では，2 つの頂点が互いに相手を導出する，対称的な導出関係のみに着目する．SDG は，頂点集合としてシードページ集合 S を持つ無向グラフである．SDG では，ADG において s から t，t から s への辺が両方存在する場合にのみ，頂点 s から t への辺が存在する．

SDG における個々の連結成分は，ウェブコミュニティとして扱うには粒度が大きく，複数の話題を含むことが多いので，さらなる分割を必要とする．図 3 に分割が必要な連結成分の 1 例を示す．この連結成分は，全体としてはコンピュータに関係する会社の集まりであるが，より細かく見ると 3 種類の企業の集合を含むことが分かる．左上には，NEC などのコンピュータメーカの集合があり，右上には Adaptec などデバイス系メーカの集合が，下には OLYMPUS などのデジタルカメラメーカの集合が存在する．この例では，3 つの集合の間にそれぞれ 1 本の辺しかないのでこれらを切断することで，分割が可能である．次の節では，SDG の分割によるコミュニティ抽出方法を示す．

5.4　ウェブコミュニティの抽出

SDG の分割は，SDG 中の 3 角形 (3 ページからなるクリーク) を基本単位として行う．3 角形に含まれるページは，多くのケースにおいて同じ話題を共有しており，分割の単位とするのに適している．例えば，図 3 中には 3 角形が幾つも存在するが，どの 3 角形も同じ業種の会社を含んでいる．SDG 中の最大クリークを用いて分割を行うことも考えられるが，我々の経験では，最大クリークでは条件が厳しい状況が多い．例えば，図 3 左上のコンピュータメーカのコミュニティは最大クリークで抽出可能であるが，右上および下のデバイスメーカやデジタルカメラメーカのコミュニティは，3 角形が辺を共有しながら緩く結合された形をしており，最大クリークでは捕え難い．このため，条件を緩める必要がある．

SDG を詳細に観察した結果，我々は，辺を共有する 3 角形の集合をコミュニティのコアとして抽出し，コアに含まれないページを隣接するコアに付加してコミュニティを抽出する手法を採用した．コアは，定義より少なくとも 1–連結の部分グラフとなる．また，簡単のため，コミュニティは，互いに重複するページを持たないようにしている．以下に，コミュニティ抽出手法の手順を示す．

（1） SDG 中の 3 角形を列挙し，辺を共有する 3 角形の集合からなる部分グラフをコアとして列挙する．

（2） 複数のコアに属するページ p が存在する時，所属するコアを 1 つ選択し，他のコアから p を取り除く．コアの選択には ADG を用いる．つまり，p から ADG の有向辺が一番多く延びているコアを選択する．

（3） コアに含まれていない各ページを隣接するコアに付加する．ただし隣接とは，SDG 内でコア中のページに対して辺を持っていることである．隣接するコアが複数存在する場合，(2) と同様に ADG を用いて付加するコアを選択する．各コアに隣接ノードを付加した部分グラフをコミュニティとして出力する．

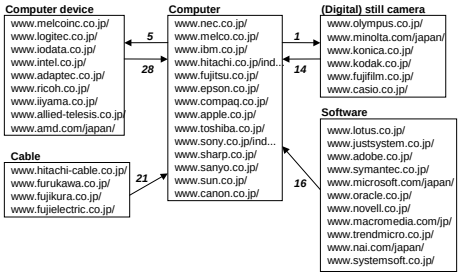


図 4　ウェブコミュニティチャートの一部
Fig. 4　A part of the web community chart

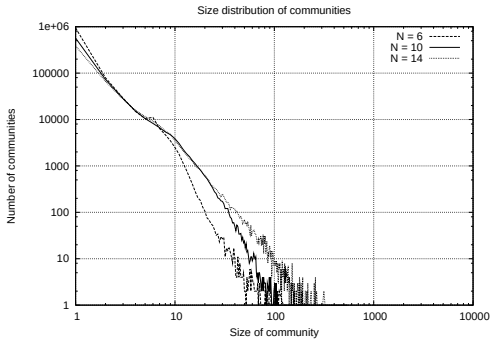


図 5　ウェブコミュニティのサイズ分布
Fig. 5　Size distribution of web communities

（4） この時点でコミュニティに含まれていないページからなる連結成分が存在すれば，それらもコミュニティとして出力する．

5.5　ウェブコミュニティチャートの作成

最後に，コミュニティ間の関連度を算出して有向辺を作りチャートを完成する．ウェブコミュニティチャートは，コミュニティを頂点とし，コミュニティ間に関連度を付加した有向辺を持つ有向グラフである．辺とその関連度は，ADG を用いて算出する．ADG においてコミュニティ c のメンバーから他のコミュニティ d のメンバーへの有向辺が w 本存在する時，c から d への有向辺を作成し，その関連度を w とする．

図 4 に，作成したウェブコミュニティチャートの一部を示す．コンピュータメーカのコミュニティを中央に置き，その周辺に存在する関連業種のコミュニティを図示した．この図では関連度の合計が 15 以上のコミュニティのみを示しているが，実際にはこの他にも多数のコミュニティが周辺に存在している．

辺の方向は，有名でないコミュニティから有名なコミュニティへ，または，細かい話題に特化したコミュニティからより一般的な話題のコミュニティへと向かう傾向がある．例えば，図 4 ではデバイスメーカからコンピュータメーカへの辺は多いが，その逆は少ない．

6.　ウェブコミュニティチャート作成実験

本節では，大規模なデータセットを用いたウェブコミュニティチャートの作成実験とその結果を述べる．

6.1　データセット

データセットとしては，2002 年 2 月に収集した 4500 万ページからなる日本のウェブアーカイブ (jp ドメインに存在するページの集合) を使用した．このアーカイブは，これ以前に収集したアーカイブから抽出した約 38 万のウェブサーバのトップページを出発点として，幅優先探索の順番で収集したものである．本アーカイブを基にしたウェブグラフのデータベースは，約 8400 万ページ (アーカイブ内部 4500 万および外部 3900 万ページ)，および約 3.7 億のリンクを含む．被リンク数の分布においても，1999 年のアーカイブと同様にべき指数約 2.2 のべき乗分布を示し，全世界のアーカイブにおける結果とほぼ一致している．

6.2　チャート作成とコミュニティのサイズ分布

シードページ選択時のパラメタ IN(5.1 参照) としては，3 を用いた．すなわち，異なるサーバからの被リンク数が 3 以上のページをシードとした．被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．最終的にシードとして選択されたページの数は，約 160 万ページであった．

まず，関連ページアルゴリズムの結果を上位から何個使用するかを決めるパラメタ N(5.2 参照) を，6 から 14 まで変化させ，コミュニティのサイズの分布を調べた．図 5 にコミュニティのサイズ (含まれるページ数) と，そのサイズのコミュニティの個数を両対数グラフで示す．サイズ 1 は，孤立しているページを意味する．$N = 10$ の分布を中心に，$N = 6$，および $N = 14$ の分布をプロットしてある．サイズの分布は，ほぼべき乗則に従う．N が大きくなるほど SDG の密度が上がるためサイズは全体的に大きくなり，N が小さくなるとサイズは全体的に小さくなる．また，N が大きくなるほど分類できるページ数が多くなり孤立するページが少なくなるが，$N = 14$ 以上では分類の粒度が大きくなりすぎる上，RPA の結果にノイズが多

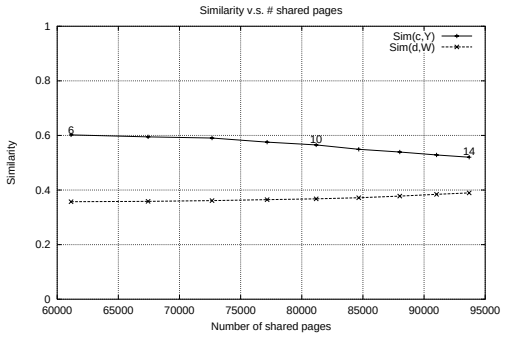


図 6　類似度と共有ページ数
Fig. 6　Similarity v.s. number of shared pages

くなるため，現実的にはチャートとしての使用は難しくなる．

6.3　ウェブディレクトリとの比較

次に，コミュニティの分類傾向と，パラメタ N の影響を調べるため，日本最大のウェブディレクトリである，Yahoo! Japan(以下，Yahoo!) との簡単な比較を行った．比較に用いたのは，2002 年 9 月の時点での Yahoo!であり，約 3.1 万のカテゴリおよび 17.7 万のユニークなページが登録されている．以降ではチャートと Yahoo!の共通部分に注目して比較を行う．つまり，チャートと Yahoo!に共有されているページのみを対象とし，他のページは存在しないものとする．

比較のため，個々のコミュニティ(c) の Yahoo!(Y) に対する類似度 $Sim(c,Y)$，個々のカテゴリ (d) のチャート (W) に対する類似度 $Sim(d,W)$ を以下のように定義する．
$Sim(c,Y) = |c \cap d'|/|c|$ (ただし、$d' \in Y$ は c と最も多くのページを共有するカテゴリ)
$Sim(d,W) = |d \cap c'|/|d|$ (ただし、$c' \in W$ は d と最も多くのページを共有するコミュニティ)

図 6 は，パラメタ N を変化させた時の全共有ページ数 (x 軸) および類似度の平均 (y 軸) をプロットしたものである．N の値は，点の上に示してある．類似度はサイズが小さいコミュニティやカテゴリでは安定しないので，共通部分においてサイズが 5 以上のコミュニティおよびカテゴリについてのみ計算し平均を取った．図 6 から分かるように，パラメタの変化に対して類似度の変化は緩やかであり，全体的に $Sim(c,Y)$ が，$Sim(d,W)$ よりも高い値となっている．$Sim(c,Y)$ は 0.5 以上を保っており，コミュニティ内の半分以上のページが 1 つのカテゴリに含まれていることを示す．これは，コミュニティの方がカテゴリよりも分類の粒度が小さく，N の変化に対して分類はおおむね安定していることを示している．

図 6 では N の増加に伴い，$Sim(c,Y)$ は減少し，$Sim(d,W)$ は増加している．$Sim(c,Y)$ の減少は，コミュニティの粒度が増加し，RPA の結果にノイズが多くなるために起こる．また，N が増加するとチャート内で距離の近いページから合併していくため，$Sim(d,W)$ の増加はチャートにおける関連度が，Yahoo!の分類とある程度一致していることを示している．

この結果は，分類の善し悪しを示すものではなく，ここから適切な N の値を決めるのは難しい．我々は経験的に，$Sim(c,Y)$ の減少が比較的少なく，孤立ページ数もシード全体の 3 分の 1 程度となる $N = 10$ 付近を使用している．

6.4　関連度とウェブディレクトリにおける距離の比較

最後に，ウェブコミュニティチャートにおける辺と関連度について，Yahoo!を用いて調査を行った．ここで調査したのは，チャートにおいて，ある関連度の辺が 2 つのコミュニティを結んでいる時，これら 2 つのコミュニティに対応する Yahoo!の 2 つのカテゴリが Yahoo!においてどの程度の距離にあるか，である．

本調査では，チャート内の辺のうち，両端のコミュニティに対応するカテゴリが唯一に定まるもののみを調査対象とした．この辺の集合は以下のように選択した．

- 第 6.3 節と同様に，チャートと Yahoo!の共通部分においてサイズが 5 以上のコミュニティを取り出す．これらのコミュニティの集合を S とする．
- 各コミュニティ $c_i \in S$ と，最も多くのページを共有するカテゴリ d_i を，c_i に対応するカテゴリとする．ただし，共有ページ数が最多となるカテゴリが複数ある場合には，c_i を S から除く．この結果，対応カテゴリが唯一に定まるコミュニティのみが S に残る．
- S に含まれるコミュニティ同士を結ぶ辺の集合 $\{(c_i, c_j, w_{i,j}) | c_i, c_j \in S\}$ を調査対象とする．ただし，本調査では簡単のため辺は無方向として扱い，辺の関連度 $w_{i,j}$ は，両方向 (c_i から c_j，および c_j から c_i) の関連度の合計とする．

本調査では，$N = 10$ で算出したコミュニティチャート (第 6.3 節参照) を用いた．共通部分におけるサイズが 5 以上のコミュニティが 4,080 個存在したのに対し，対応カテゴリが唯一に定まるコミュニティの数は

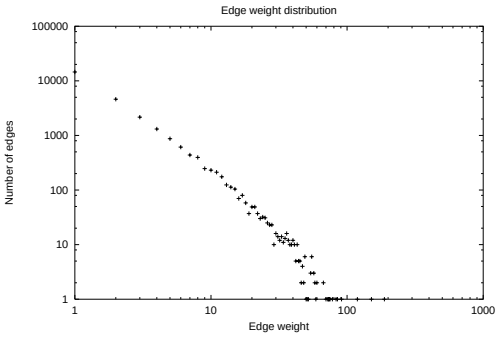


図 7　辺における関連度の分布
Fig. 7　Edge weight distribution

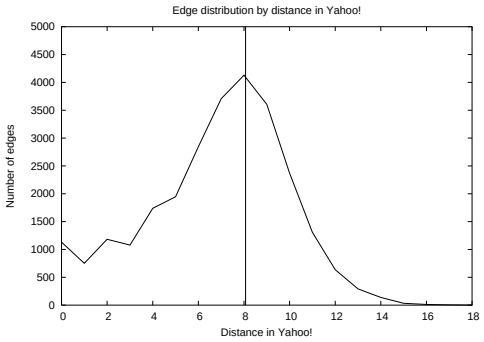


図 8　Yahoo!のカテゴリ間距離による辺の分布
Fig. 8　Edge distribution by distance in Yahoo!

3,252 個であった．これらのコミュニティを結ぶ辺の数は 26,910 本であった．辺の関連度の分布を図 7 に示す．横軸に辺の関連度を，縦軸にその関連度を持つ辺の数を，両対数で表してある．関連度の分布はべき乗分布を示しており，辺の集合が大多数の関連度の低い辺，および少数の関連度の高い辺から構成されていることが分かる．

次に，辺の両端にあるコミュニティ (c_i, c_j) に対応するカテゴリ (d_i, d_j) 間の距離について調査を行った．この距離は，Yahoo!の階層構造の中で意味的に近い程短くなるように定義した．この距離を用いて，関連度の高い辺に対応するカテゴリ間の距離が短い傾向にあるかどうかを調査した．

カテゴリ d_i と d_j 間の距離は，d_i からカテゴリ階層を辿って d_j まで到着するまでのステップ数と定義する．この距離に依ると，子カテゴリから親カテゴリまでの距離は 1 となり，同じ親カテゴリを持つ兄弟カテゴリ間の距離は 2 となる．ただし，d_i と d_j の間に直接のシンボリックリンク (階層の異なるカテゴリ間を結ぶハイパーリンク) が存在する場合には，距離は 1 とした．

図 8 に，各辺に対応するカテゴリ間距離の分布を示した．横軸にカテゴリ間の距離，縦軸にそのカテゴリ間距離に対応する辺の数を示した．カテゴリ間距離 8 付近にある縦の線は，任意のカテゴリ間の距離の平均 (8.07) を示している．図 8 からは，グラフの最大のピークが 8 付近にあり，カテゴリ間距離の短い方向へ偏った分布となっていることが分かる．これは多くの辺が，対応するカテゴリをランダムに結んでいるのとほぼ変わらないことを意味する．しかし，グラフの短距離方向への偏りから，対応するカテゴリ間の距離が短い辺もある程度存在することが分かる．

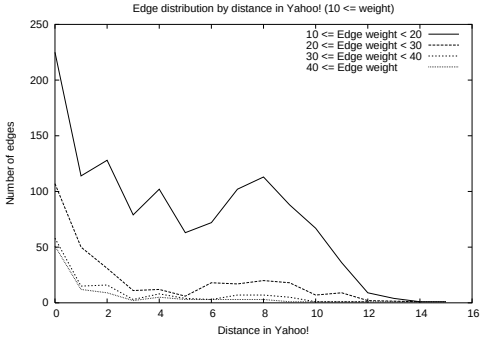


図 9　Yahoo!のカテゴリ間距離による関連度 10 以上の辺の分布
Fig. 9　Edge distribution by distance in Yahoo! (with 10 or more edge weight)

ここで辺の関連度の分布がべき乗分布に従うことに注意する必要がある．関連度の低い辺がほぼ全体を占めているため，図 8 からは高い関連度を持つ辺の傾向が読み取れない．そこで，関連度が 10 以上の辺のみを取り出し，それらの辺に対応するカテゴリ間距離の分布を図 9 に示した．10 以上の関連度を 10 毎の区間に区切り，それぞれに対して分布を示している．ただし，40 以上の辺は 1 区間にまとめてある．このグラフからは，辺の関連度が高いほど，対応するカテゴリ間距離が短い可能性が高いことが分かる．関連度 10 以上 20 未満の辺については，まだ 8 付近にピークが残っているものの，最大のピークは距離 0(対応するカテゴリが等しい) にあり，距離が短い可能性が比較的高くなっている．それ以上の関連度では，距離が短い可能性がより高くなっていることが分かる．

対応するカテゴリ間距離が長い場合に，その辺に意

味がないとは必ずしも言い切れない．そのような辺をサンプリングして調査したところ，本来なら直接のシンボリックリンクが存在するべきカテゴリ同士を結んでいる辺が数多く存在した．例えば，以下の 2 つのカテゴリは意味的には近いが，直接にはシンボリックリンクで結ばれていない．

- /コンピュータとインターネット /インターネット /WWW /CGI /スクリプト
- /コンピュータとインターネット /ソフトウェア /プログラミングツール /プログラミング言語 /Perl

7. まとめと今後の課題

大規模なウェブアーカイブから，ほぼ全てのコミュニティを抽出して，関連するコミュニティを結ぶチャートを作成する手法を提案し，現実のデータセットに対して実験を行った．Yahoo! Japan との比較から，平均してコミュニティ内の約半数のページが同じカテゴリに含まれ，コミュニティによる分類は Yahoo!よりも粒度が小さいことが判明した．またチャートの辺と関連度についても Yahoo!を用いて調査を行った．この結果，辺の関連度が高い程，対応するカテゴリ間の距離が短くなる可能性が高く，辺の関連度が低いと対応するカテゴリ間の距離がランダムに近くなることが判明した．現在，大規模なチャートを閲覧するためのブラウザを開発するとともに，時間によるチャートの発展過程の抽出手法を開発している．

謝辞

本研究の一部は，文部科学省科学研究費特定領域研究 (13224014) によるものである．

文　　献

[1] K. Bharat and M. Henzinger. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In *Proc. ACM SIGIR '98*, 1998.

[2] S. Chakrabarti. Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation. In *Proc. 10th WWW Conference*, 2001.

[3] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. 7th International WWW Conference*, 1998.

[4] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. In *Proc. 8th WWW Conference*, 1999.

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

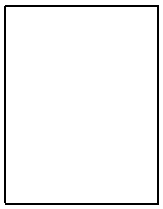
[6] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web Communities from Link Topology. In *Proc. HyperText98*, 1998.

[7] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[8] R. Lempel and S. Moran. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. In *Proc. 9th WWW Conference*, 2000.

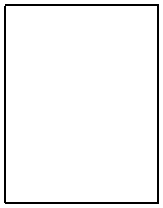
[9] S. R. Ravi Kumar, Prabhakar Raghavan and A. Tomkins. Trawling the Web for emerging cybercommunities. In *Proc. 8th WWW Conference*, 1999.

（平成 x 年 xx 月 xx 日受付）



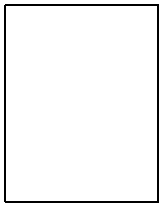
豊田　正史

1994 東京工業大学理学部情報科学科卒．1996 同大学大学院情報理工学研究科修士課程修了．1999 同大学大学院情報理工学研究科博士後期課程修了．博士 (理学)．同年，科学技術振興事業団計算科学技術研究員．2001 東京大学生産技術研究所学術研究支援員．現在，同大学同研究所産学官連携研究員．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に興味を持つ．ACM, IEEE CS, 情報処理学会，日本ソフトウェア科学会各会員．



吉田　　聡

2001 東京工業大学電気電子工学科卒業．2003 東京大学大学院情報理工学系研究科電子情報学専攻修士課程修了．



喜連川　優　（正員）

1978 東京大学工学部卒．1983 同大大学院工学系研究科情報工学博士課程了．工学博士．同年同大生産技術研究所講師．現在，同教授．2003 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，ウェブマイニングに関する研究に従事．現在，情報処理学会理事，日本データベース学会理事，平成 11-14 年 ACM SIGMODJapan Chapter Chair 平成 9，10 年本学会データ工学研究専門委員会委員長．VLDB Trustee(97-02)，IEEE ICDE，PAKDD，WAIM などステアリング委員