ウェブコミュニティチャートとウェブディレクトリの比較に関する一考察

豊田正史[†] 吉田聡[†] 喜連川優[†]

ウェブ上には，同じ話題に興味を持つ人々によって作成されたページの集合が数多く存在する．それらの多くはハイパーリンクで密に結合されることでウェブコミュニティを形成しているため，リンク解析を用いた様々な抽出手法が提案されてきている．我々はこれまでに，大規模なウェブのアーカイブからすべてのウェブコミュニティを抽出するとともに，それらの間の関連度を算出し，ウェブコミュニティチャートと呼ばれる連関図を作成する手法を開発してきた．この手法は膨大なページを自動的に分類できるが，妥当な質が得られているかどうかを評価するのは容易ではない．また，ウェブコミュニティの性質についての，大規模かつ詳細な調査はいまだに行われていない．本論文では，この評価へ向けての第一歩として，我々のウェブコミュニティチャートと，人手で分類されたウェブディレクトリとの詳細な比較を行い，分類の類似点および相違点について検討する．

Comparing the Web Community Chart with a Web Directory

Masashi Toyoda, Satoshi Yoshida[†] and Masaru Kitsuregawa[†]

Recent research on link analysis has shown the existence of numerous web communities on the Web. A web community is a collection of web pages created by individuals or any kind of associations that have a common interest on a specific topic. We have developed a web community chart that connects related communities, and allows us to navigate the Web through related topics. This chart can classify numerous web pages. However, its accuracy of classification is still unclear. In this paper, we compare our web community chart with a web directory, and clarify differences between them.

1. はじめに

ウェブ上には，同じ話題に興味を持つ人々によって作成されたページの集合が数多く存在する．それらの多くはハイパーリンクで密に結合されることでウェブコミュニティ☆を形成しているため，リンク解析を用いた様々な抽出手法が提案されてきている．既存のリンク解析手法[1)~5)]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により，ウェブコミュニティを抽出する．リンク解析を用いると，テキスト解析では捕らえることが難しいページの関係を抽出できる．一例をあげると，大企業などの有名なページでは，画像が多用され，特徴を表すキーワードがページにほとんど現れないことが多いため，テキスト解析を適用するのは難しい（たとえば，SONY のトップページにはコンピュータというキーワードは現れない）．しかし，ウェブ上に多数存在する公開ブックマークおよびリンクリストでは，有名なページは話題ごとに分類されたうえでリンクが張られている．このため，同じ分類をしている（同じパターンでリンクを張っている）ページが多数存在することを検出すれば，信頼できる分類を得ることができる．

我々はこれまでに，大規模なウェブのアーカイブからすべてのウェブコミュニティを抽出するとともに，それらの間の関連度を算出し，ウェブコミュニティチャート[6),7)]と呼ばれる連関図を作成する手法を開発してきた．この手法は膨大なページを自動的に分類できるが，妥当な質が得られているかどうかを評価するのは容易ではない．また，既存のリンク解析に関する研究においてもウェブコミュニティの性質についての，大規模かつ詳細な調査はいまだに行われていない．

本論文では，大規模なウェブコミュニティ群の評価へ向けての第一歩として，我々のウェブコミュニティ

† 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo

☆ 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．

チャートと，人手で分類されたウェブディレクトリとの比較を通して分類の類似点および相違点について検討する．ウェブディレクトリは，ウェブページを人手により木構造に分類したものであり Yahoo!や Open Directory などが例としてあげられる．これらは，多数のページに人間が目を通して分類を行っているため，ある程度分類精度が保証される大規模な比較対象として利用できる．

今回の比較には，国内 4,500 万ページからなるウェブアーカイブから抽出したウェブコミュニティチャートを使用し，比較対象となるウェブディレクトリとしては国内最多の分類ページ数を持つ Yahoo! Japan を使用した．まずチャートおよびウェブディレクトリの類似度を導入して両者の比較を行い，全体的にどの程度，分類が類似しているかを示した．さらに類似度の低い部分については，分類の観点が異なっているためか，チャートにおける分類が誤っているためかを明らかにするためにサンプリング調査を行い，相違点を調査した．

本論文の構成は以下のとおりである．2 章では関連研究について述べ，本論文の位置付けを明らかにする．3 章では，実験に用いたデータセットを紹介し，4 章で，チャートと Yahoo! Japan のページ集合の相違を示す．5 章ではコミュニティとウェブディレクトリの比較結果について述べ，6 章ではその中で類似度の低いコミュニティに関して手作業で分類を行った結果を述べる．7 章でまとめと今後の課題を述べる．

2. 関 連 研 究

我々のウェブコミュニティチャートは，Kleinberg [8] 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作成されている．オーソリティとは，あるトピックについて良質な内容を持つページを指し，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，あるトピックに関するリンク集やブックマークページを指し，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．Kleinberg は，この循環した定義に基づいてウェブの部分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 HITS [8] を提案した．Gibson らは HITS で得られるオーソリティの集合をウェブコミュニティと見なし，その性質について調査を行っている [1]．さらに 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 [2),9),10)]．また Dean らは，与えられたシードページに対し関連するページ群を結果として返す，Companion [11] という関連ページアルゴリズム（RPA）に HITS を応用している．Companion は，シードページ周辺のグラフからオーソリティを抽出する．我々はチャート作成に，この Companion を改良したものを用いている．これらの研究では，数十の実例を使用した主観評価が行われているが，大規模な結果の質に関する評価はいまだに行われていない．

Kumar らは，2 億ページ以上の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 万個を超えるコミュニティのコアを発見した [3]．部分グラフから特定のトピックを取り出す HITS と異なり trawling はウェブグラフ全体からコアをリストアップする．コアとはオーソリティとハブ☆からな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．Kumar らは，発見したウェブコミュニティの質の検証に Yahoo!を用いている．サンプリングした 400 個のコアの 7 割が Yahoo!上に何らかの形で存在したという結果を示している．しかし，コアの質についての評価はほとんど行われていない．

我々のチャート手法 [6),7)] は，HITS に基づいているが，ウェブ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は trawling と同じである．まずウェブグラフ全体から被リンク数の多いページ群を選んでシードセットとする．次に，すべてのシードページに対して改良した Companion アルゴリズムを適用して，各シードが他のシードを上位 N 個のオーソリティとして導出する関係を表す有向グラフを作成する．最後に，この導出グラフ内で密に結合されたシードの集合をコミュニティとして抽出することで，コミュニティを頂点，コミュニティ間の関連度を重み付き辺とするグラフをチャートとして出力する．関連度としては，2 つのコミュニティにまたがるシードどうしの導出関係の数を用いる．手法の詳細については，文献 6)，7) を参照されたい．文献 7) では，Yahoo! Japan との単純な比較を用いて，上記のパラメータ N を変化させてもコミュニティの分類が安定していることを示した．本論文では，カテゴリの階層やコミュニティ間の関連度を考慮して，より詳細な質の調査を行う．

また，異なるウェブディレクトリどうしの比較については市瀬ら [12] による研究がある．これは，ウェブディレクトリを階層的な知識と見なし，異なるウェブディレクトリどうしで知識を補完しあう手法を提案し

☆ 文献 3) では，オーソリティおよびハブではなく，センターおよびファンという用語が使用されているが，基本的な意味は同じである．ここでは混乱を避けるため前者で統一した．

ている．具体的には，2つの階層間に，共有するインスタンス数を基にした類似度を導入し，類似する階層の間でインスタンスを交換しあう．本研究でも共有インスタンス数を基にした類似度を用いているが，市瀬らの研究が階層構造どうしの比較になっているのに対し，本研究はグラフ構造と階層構造との比較となっている．また我々は，類似しない部分の情報にも利用価値があると考えており，本論文では相違点についても比較検討を行っている．

3. データセット

3.1 ウェブアーカイブおよびウェブグラフ

データセットとしては，2002年2月に収集した4,500万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．本アーカイブから作成した全体のウェブグラフは，約8,400万ページおよび約3.7億のリンクを含む．8,400万ページの内訳は，アーカイブ内部4,500万ページ，および，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3,900万ページとなっている．

3.2 ウェブコミュニティチャート

本調査では，上述した日本のウェブアーカイブから作成したウェブコミュニティチャートを比較対象として用いた．チャート作成アルゴリズムへの入力となるシードページの集合（2章参照）としては，3つ以上のサーバからリンクされているページを使用し，シードページの総数は，約160万ページであった．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ータ N（2章参照）としては，10を用いた．$N = 10$ の周辺において，コミュニティの粒度が安定しており，分類できるページ数も十分多いことからこの値を選択した．N の変化によるコミュニティ粒度の安定性などについては文献7)で調査しているので参照されたい．

このコミュニティチャートでは，2ページ以上を含むコミュニティの数は約17万5千個あり，コミュニティを形成しない孤立したシードページの数は約57万ページであった．コミュニティに含まれるページ数を p とすると，ページ数 p を持つコミュニティの数 C_p の分布は，ほぼべき乗則に従う．すなわち C_p は $1/p^k$ に比例する．k の値は約2.9であった．

3.3 Yahoo! Japan

Yahoo! Japanは，人手で分類されているものの中では日本最大のウェブディレクトリである（以降Yahoo!はYahoo! Japanと同じ意味で使用する）．比較に使用するデータとして，2002年9月時点の内容を使用した．ただし，「地域情報/日本の地方」以下のカテゴリは，他のカテゴリにおいてすでに分類されているページを地域によって分類し直したものであり，ページの話題による分類ではないことから，今回は分析の対象とはしなかった．

表1 Yahoo! Japanの各トップカテゴリのURL数
Table 1 The number of URLs in each top category of Yahoo! Japan.

トップカテゴリ	カテゴリ数	URL数	平均URL数
芸術と人文	1,607	13,449	8.37
ビジネスと経済	12,132	71,487	5.89
コンピュータとインターネット	867	6,888	7.94
教育	404	3,769	9.33
エンターテインメント	5,073	27,579	5.44
政治	387	3,996	10.33
健康と医学	794	5,978	7.53
メディアとニュース	625	3,715	5.94
趣味とスポーツ	3,750	22,483	6.00
各種資料と情報源	100	1,691	16.91
地域情報	1,805	5,588	3.10
自然科学と技術	1,575	9,386	5.96
社会科学	723	3,636	5.03
生活と文化	1,221	20,902	17.12
合計	31,072	200,541	6.45

Yahoo!には約3万1千個のカテゴリの中に約20万ページが登録されている．ただし，複数のカテゴリに登録されているページが約2万3千ページ存在するため，実際のページ数は約17万7千ページである．Yahoo!のトップカテゴリごとの状況を**表1**にまとめた．URL数では「ビジネスと経済」カテゴリが多く，企業ページか否かが主要な分類基準になっていることが分かる．また，Yahoo!では，すべてのカテゴリが複数のページを含むとは限らない．約7千のカテゴリが単独のページからなっている．

ここで用いるYahoo! Japenの内容は，我々のウェブアーカイブの収集時期より後のものであるため，アーカイブから作成したウェブグラフに存在しないページを約2万7千含んでいる．これらのページは，ウェブアーカイブ収集後に作成されたページ，および，収集時に何らかの理由で収集ができなかったページ☆を含んでいる．結果として比較対象となるYahoo!のページ数は，約15万ページとなる．

4. ウェブコミュニティチャートとYahoo!におけるページ集合の相違

本章では，分類の比較を行う前に，ウェブコミュニティチャートとYahoo!に含まれるページ集合の相違点を述べる．チャートは約160万ページからなるシードページ集合を元に作成されており，そのうち約103万ページが孤立せずに2ページ以上からなるコミュニティに分類されている．Yahoo!に含まれる約15万

☆ 一時的に不通だったページ，クローラによるアクセスを拒否しているページ，収集期間内に到達できなかったページなど．

ページと比較すると，チャートはシードページ数では10倍以上，孤立していないページ数では7倍近くの大きさとなっている．

チャートは，Yahoo!よりも多数のページを自動分類しているが，Yahoo!を完全に包含してはおらず，含まれているページの質は異なっている．チャートのシードページ集合とYahoo!とに共有されているページ数は，約11万4千ページである．これは，チャートが最大でYahoo!の約76%をカバーできることを意味する．Yahoo!の残り約3万6千ページは，3つ以上のサーバからリンクされていないためチャートの分類対象から外れたことになる．また，2ページ以上のコミュニティに含まれるページの集合とYahoo!との共有ページ数は約8万1千ページである．孤立している3万3千ページのほとんどは，パラメータ N（2章参照）を大きくするなどの操作を行うことで，コミュニティを形成するようになる．しかし，カバーできる範囲が大きくなる分，関連のないページが結合される可能性が高くなるため，今回の実験ではこの操作を行っていない．

以上からYahoo!に登録されているページすべてが，十分な被リンク数を持つ著名なページではないことが分かる．この原因は，Yahoo!がページ作者による申請に基づいてページの登録をしていることにある．登録審査はページの内容によって行われるため，著名でないページが多数登録されており，Yahoo!からリンクされているだけのページも数多く存在する．一方でチャートには，十分な被リンク数を持ち著名であるがYahoo!には登録されていないページが多数含まれている．

これらの事実は，チャートとYahoo!とが互いの弱点を補いあえる可能性を示唆している．この相互補間は我々の最終的な目的であるが，そのためにはまず共通部分における分類の相違点を明らかにする必要がある．以降の章では，チャートとYahoo!の共通部分において，分類がどの程度類似しているのかを定量的に比較し，どのような場合に分類が異なるのかを調査検討する．

5. ウェブコミュニティとYahoo!カテゴリとの類似度による比較

5.1 単純類似度による比較

本章では，ウェブコミュニティチャートによる分類がどの程度妥当かを調べるため，チャートとYahoo!との類似度を比較する．チャートはグラフであり，Yahoo!は木構造と分類構造が異なるため，妥当な類似度を導出するのは難しい．そこで，まず個々のウェブコミュニティとカテゴリを1対1で比較する最も単純な類似度を導入して比較を行う．この単純類似度は，文献7)で導入したものと同じであるが，論文の完全性のため再掲する．文献7)ではパラメータ N を変化させたときの平均類似度の変化に着目したが，本論文では類似度の値によるコミュニティ数の分布に着目する．

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．ただし，チャートおよびYahoo!の構造は変わらないものとして扱う．今回のデータセットにおいて，共通部分におけるページ数は81,177ページであった．

以下において，この共通部分のみからなるチャートを W，Yahoo!を Y で示す．W はウェブコミュニティの集合 $(c_1, \ldots, c_n)$ であり，Y はカテゴリの集合 $(d_1, \ldots, d_m)$ である．コミュニティおよびカテゴリはページの集合である．また，Yahoo!は，階層構造内の中間カテゴリにもページを分類しており，それらは末端カテゴリのページよりも一般的な内容を持つ．そこで，ここでは中間カテゴリも独立した1つのカテゴリとして扱う．

個々のコミュニティ（c）のYahoo!（Y）に対する単純類似度 $Sim(c, Y)$，および，個々のカテゴリ（d）のチャート（W）に対する単純類似度 $Sim(d, W)$ 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1 に両類似度の分布を示した．共通部分においてサイズが5以上のコミュニティは約4,000個存在する．この約4,000のコミュニティ中，約45%のコミュニティが0.6以上の単純類似度（$Sim(c, Y)$）を持っており，Yahoo!の特定のカテゴリに対応付けられることが分かる．残り55%のコミュニティはYahoo!においてより細かく分類されていることになる．一方，共通部分においてサイズが5以上のカテゴリは約5,000個あるが，単純類似度（$Sim(d, W)$）が0.6以上となったのは約5,000個中20%であった．これは残り80%のカテゴリが，チャートにおいて細分化されていることを示している．また，共通部分での平均サイズを見ると，コミュニティが8.13ページであるのに対しカテゴリは12.84ページであり，カテゴリのサイズが比較的大きいことが分かる．

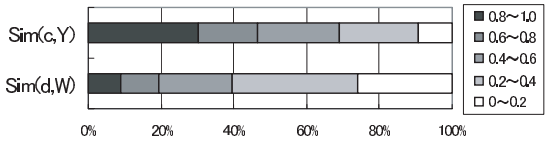

図 1 単純類似度の分布
Fig. 1 Distribution of one to one similarity.

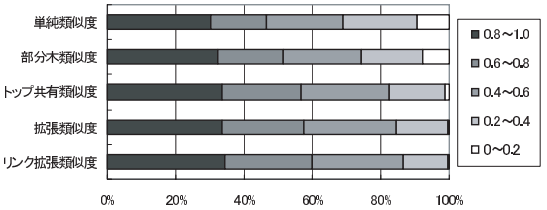

図 2 拡張類似度 $Sim_{ext}(c, Y)$ の分布
Fig. 2 Distribution of expanded similarity, $Sim_{ext}(c, Y)$.

5.2 カテゴリ階層を考慮した比較

Yahoo!は木構造を分類に利用しているため，木構造において近い場所にあるカテゴリは類似した話題を持つと考えられる．このため，1 つのウェブコミュニティに含まれるページが複数のカテゴリに分布していても，それらのカテゴリが近い位置にある場合，それを考慮した類似性を計算することでより実態を反映することができる．そこで，個々のコミュニティ（c）と Yahoo!（Y）との単純類似度を拡張し，カテゴリの距離を考慮した拡張類似度を定義する．

まず，コミュニティ c に含まれる各ページが Y において所属するカテゴリのリストを $(d_1, \ldots, d_k) \subset Y$ とする．k は，c に含まれるページ数に等しく，リストには同じカテゴリが複数含まれうることに注意されたい．またリスト中で，c と最も多くのページを共有するカテゴリを d' とする．このとき c の Y に対する拡張類似度 $Sim_{ext}(c, Y)$ を以下のように定義する．

$$Sim_{ext}(c, Y) = \frac{\sum_{i=0}^{k} 1/D(d', d_i)}{k}$$

ただし，$D(d', d_i)$ は，カテゴリ間の距離を表し，d' から d_i までカテゴリの木構造をたどって到達するまでのステップ数とする．このステップ数は 1 から数えるため $D(d_i, d_i) = 1$ である．

$Sim_{ext}(c, Y)$ は 0 から 1 の間の値をとり，値が高いほど c は Y の分類と類似していることを示す．この拡張類似度は，異なるカテゴリ間の距離をすべて ∞ とすると，単純類似度と等しくなるように定義してある．また基本的には，d' に近いカテゴリにページが存在すればするほど，高いスコアが加算される．このため，単純類似度との差を比較しやすい類似度になっている．

拡張類似度は，距離を任意のカテゴリ間について定義するが，トップカテゴリや親カテゴリを通過して，他のカテゴリに到達する場合，話題が保たれている保証はない．またカテゴリ間のシンボリックリンクも考慮して距離を計算することも可能である．ここでは，木構造およびシンボリックリンクを考慮して距離を計算する範囲を変更することで以下の 4 通りの拡張類似度を使用する．これらは距離の制限の厳しいものから緩いものへ順番にならべてある．

部分木類似度 Yahoo!において，ほぼ確実に関連のあるページのみが含まれている範囲で類似度計算を行う．すなわち，カテゴリ間の距離の計算を d' 以下のサブカテゴリおよび d' の祖先カテゴリについてのみ行う．それ以外のカテゴリ間の距離は ∞ となる（d' の兄弟カテゴリも無視する）．

トップ共有類似度 部分技類似度より範囲を広げ,「芸術と人文」以下,「ビジネスと経済」以下など，Yahoo!において同じトップカテゴリに含まれれば距離計算の考慮対象とする．これにより，兄弟および，いとこカテゴリも考慮した類似度となる．具体的には，カテゴリ間の距離の計算を d' と同じトップカテゴリ内に存在するカテゴリについてのみ行い，それ以外は距離を ∞ として扱う．

拡張類似度 距離の計算の際の制限をなくし，カテゴリ間の距離の計算をすべてのカテゴリについて行う．

リンク拡張類似度 Yahoo!においては関連するカテゴリどうしがシンボリックリンクで結合されているため，結合されているカテゴリ間の距離は短いと見なせる．この類似度では，カテゴリ間の距離の計算をすべてのカテゴリについて行ううえ，シンボリックリンクをたどることも許す．シンボリックリンクをたどる場合もステップは 1 と数える．

図 2 に，単純類似度とそれぞれの拡張類似度の分布を示す．この結果から，コミュニティチャートと Yahoo!の構造に類似性のあることが分かる．まず，一番条件の厳しい部分木類似度においても，0.6 以上の値を示すコミュニティが 50%に達しており，単純類似度の場合より 5%増加している．これは，全体の約 5%のコミュニティについては，単純類似度が低くてもそのメンバは Yahoo!上の関連のあるカテゴリに分布していることを示している．トップ共有類似度および拡張類似度では，さらに 0.6 以上の値を示すコミュニティが増加して約 55%となる．これは，d' の兄弟カテゴリにコミュニティのメンバがしばしば含まれているこ

とを示している．トップ共有類似度と拡張類似度の間では，値にほとんど変化はなく，「トップカテゴリを共有している」という条件は影響が少ないことが分かる．シンボリックリンクを考慮するとさらに類似度は増加して 0.6 以上の値を示すコミュニティの数は約 60%となる．木構造上では距離が遠くても，シンボリックリンクを考慮すると近い位置にコミュニティのメンバが分布しているケースが多いことが分かる．

5.3 コミュニティ間の関連度を考慮した比較

ウェブコミュニティチャートでは，関連するコミュニティどうしを辺で結んでおり，これを考慮することで 1 つのカテゴリに含まれるページが，チャートにおいても近くに分布しているかを調べることができる．そこで，個々のカテゴリ（d）とチャート（W）との単純類似度を拡張し，コミュニティ間の関連を考慮した拡張類似度を定義する．

まず，カテゴリ d に含まれる各ページがチャート（W）において所属するコミュニティのリストを $(c_1, \ldots, c_k) \subset W$ とする．k は，d に含まれるページ数に等しく，リストには同じコミュニティが複数含まれうることに注意されたい．またリスト中で，d と最も多くのページを共有するコミュニティを c' とする．このとき d の W に対する拡張類似度 $Sim_{ext}(d, W)$ を以下のように定義する．

$$Sim_{ext}(d, W) = \frac{\sum_{i=0}^{k} 1/D(c', c_i, w)}{k}$$

ただし，$D(c', c_i, w)$ は，コミュニティ間の距離を表し，c' から c_i まで関連度 w 以上の辺のみをたどって到達するときのステップ数とする．関連度 w は，チャートにおけるコミュニティ間の関連度である（2 章参照）．ただし 5.2 節と同様にステップ数は 1 から数える．また，c' から c_i に到達不可能な場合，距離は ∞ とする．

$Sim_{ext}(d, W)$ は 0 から 1 の間の値をとり，値が高いほどカテゴリ d はチャート W に分類が類似していることを示す．この拡張類似度は，すべてのコミュニティ間の距離を ∞ とすると，単純類似度と値が等しくなるように定義されており，単純類似との差の比較が可能になっている．

図 3 に，単純類似度と閾値 w の値を変化させたときの拡張類似度の分布を示す．閾値 w が 8 になると，0.6 以上の拡張類似度を示すカテゴリが全体の約 30%となり，単純類似度と比べて 10%の増加が見られる．閾値 w を 1 とすると，全体の約 50%が 0.6 以上の拡張類似度を示す．これは，1 つのカテゴリに含まれるページは，チャートにおいても近い距離に分布していることを示す．

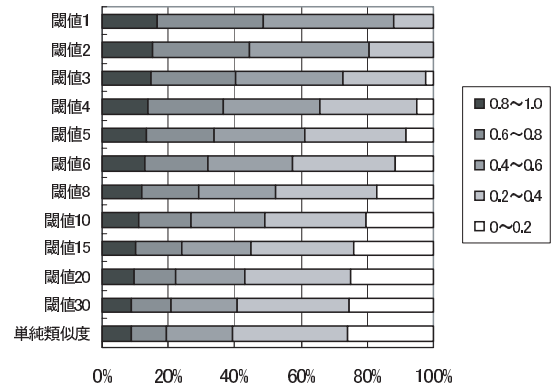

図 3 拡張類似度 $Sim_{ext}(d, W)$ の分布
Fig. 3 Distribution of expanded similarity, $Sim_{ext}(d, W)$.

表 2 サンプリングしたコミュニティの分類
Table 2 The classification result of sample communities.

分類	個数
共通の話題があるが，分類観点が異なる	36
共通の話題があるが，Yahoo!の方が詳細	8
共通する話題が存在しない	56
合計	100

6. 類似度の低い部分の比較検討

6.1 ウェブコミュニティのサンプリング調査

本章では，Yahoo!との類似度の低いウェブコミュニティが実際には共通する話題を持つかどうか，および，各ページがどのように Yahoo!において分散しているかを調査した結果を示す．具体的には，5.2 節で使用したどの拡張類似度を用いても Yahoo!との類似度が 0.6 未満となるコミュニティをランダムに 100 個選び，内容の調査を行った．

まず，各コミュニティに含まれるページを実際に確認し，共通する話題を持つコミュニティをリストアップした．判断基準は，6 割以上のページが共通する話題を持つこととした．この結果，100 個のコミュニティのうち，44 個に共通する話題が見られ，残りの 56 個については共通の話題が見られなかった．さらに，これら 44 個のコミュニティは，Yahoo!の方が分類が詳細なケース，および分類観点が異なるケースに分類できる．分類したコミュニティの個数を**表 2** に示す．Yahoo!の方が詳細なケースは比較的少なく，これは，Yahoo!の方が全体的に分類が大きいという結果に合致する．以下に各分類における具体例を示す．

分類観点の相違 分類の観点が異なるために，各ページが Yahoo!の異なるカテゴリ以下に分類されているコミュニティである．たとえば，企業のペー

ジと個人のページで共通の話題を扱っていてもYahoo!において企業サイトは「ビジネスと経済」，個人サイトは「趣味とスポーツ」に分類されてしまう．ウェブコミュニティは，ウェブ上のリンク集に頻繁に現れる分類を基にしていることから，Yahoo!にはウェブ上の多くのリンク集とは分類が解離しているケースがあることが分かる．1つのコミュニティに含まれるページが異なるトップカテゴリに分散している場合に，頻出するトップカテゴリの組合せは以下のとおりである．

- 「ビジネスと経済」「自然科学と技術」
- 「ビジネスと経済」「生活と文化」
- 「ビジネスと経済」「エンターテインメント」
- 「芸術と人文」「エンターテインメント」
- 「生活と文化」「自然科学と技術」

「ビジネスと経済」カテゴリとの組合せの数が多いが，これは Yahoo!において企業ページであるかどうかが主要な分類基準になっているためである．対してウェブコミュニティは，企業かどうかに関係なくページを分類しているケースが多いことが分かる．いくつかの組合せについて具体的な例を示す．まず「ビジネスと経済」と「自然科学と技術」の場合では，天文学に関するウェブコミュニティが例としてあげられる．このコミュニティでは，企業が運営する「星座の博物館」と題した星座や天文学を説明するページが「ビジネスと経済」以下に存在し，一般の天文学を扱うサイトは「自然科学と技術」以下に分類されていた．次に「生活と文化」「自然科学と技術」の組合せでは，森林に関するウェブコミュニティが例としてあげられる．このコミュニティでは，森林科学から見たページが「自然科学と技術」以下に分類され，環境保護の観点から森林を見たページが「生活と文化」以下に分類されていた．

Yahoo!の方が詳細 コミュニティに含まれる各ページの所属する Yahoo!のカテゴリが途中まで同一のパスを持ち，それ以下が異なるため類似度が低くなっているものである．原則として深さ3以上の同一パスを持つものをこの分類とした．以下は，雑貨屋に関するコミュニティを例として，それに含まれる各ページが所属する Yahoo!のカテゴリのリストを示したものである．

ビジネスと経済/ショッピングとサービス/雑貨
- /輸入雑貨/アジア/インドネシア
- /輸入雑貨/アジア/タイ
- /輸入雑貨/ラテンアメリカ/メキシコ
- /テイスト，モチーフ/動物/イルカ
- /バラエティ

この例では,「ビジネスと経済/ショッピングとサービス/雑貨」まではパスが一致しており，Yahoo!内ではさらに扱う商品によって細分化されている．

6.2 Yahoo!カテゴリのサンプリング調査

本節では，ウェブコミュニティチャートとの類似度の低い Yahoo!のカテゴリをサンプリングし，その相違を調査した．具体的には，閾値を1としたときの拡張類似度が 0.6 未満のカテゴリを 100 個ランダムにサンプリングし，各ページがどのようにチャート内に分散しているかを調査した．この結果，各カテゴリを**表 3** に示すように3通りに分類した．各分類のカテゴリ個数はほぼ同数となっている．以下に各分類における具体例を示す．

分類観点の相違 ウェブコミュニティチャートと Yahoo!で分類の観点が異なっていたと判断したカテゴリである．これもまた Yahoo!とウェブ上のリンク集とでは分類に解離があるケースがあることを示している．たとえば,「ビジネスと経済/企業間取引/商社/総合商社/丸紅」カテゴリは，丸紅社が運営している会社の一覧が掲載されているのに対し，チャートでは丸紅社に関係なく福祉，不動産，総合商事，ファッションなど事業ごとに各社が分類されていた．

チャートの方が詳細 カテゴリとページを共有するコミュニティがいくつかのより詳細なトピックに分類されており，そのカテゴリをより細かく分類する余地が残っているということを示す．たとえば,「自然科学と技術/生物学/植物学/植物」は，植物について扱ったページを集めたカテゴリであるが，チャート上ではその中に含まれるページが植物図鑑のページ，個人が製作したガーデニングに関するページ，法人が製作したガーデニングに関する

表 3 サンプリングしたカテゴリの分類
Table 3 The classification result of sample directories.

分類	個数
分類観点の相違	30
チャートの方が詳細	34
チャートによる細分化が不明確	36
合計	100

ページ，および，植物写真に関するページに分類されていた．

チャートによる細分化が不明確 チャートにおいて細分化されているカテゴリのうち，その分類理由が不明確なものである．たとえば，「自然科学と技術/物理学/高エネルギー物理学，素粒子物理学」カテゴリ内のページは，チャート上では5つのコミュニティに分割されていたが，分割される理由は判然としなかった．これらのページ周辺においてリンクの密度が薄く結合されるのに十分なリンク数がないことが原因として考えられる．

7. まとめと今後の課題

ウェブコミュニティを抽出する様々な手法が開発されているが，妥当な質が得られているかの大規模な評価は，いまだに行われていない．本論文では，その第一歩として，我々がこれまでに開発してきたウェブコミュニティチャートによる分類の妥当性を検討すべく，日本最大のウェブディレクトリ Yahoo! との比較調査を行った．本論文は，1つのウェブディレクトリと我々のチャート，それぞれの事例の比較にすぎず，ウェブディレクトリとウェブコミュニティの一般的な比較を試みているわけではない．したがって，得られた結果も限定されたものではあるが，我々の知る限りにおいて同様の実験は報告されていない．

我々は1つのウェブコミュニティと Yahoo!，および，1つのカテゴリとチャートを比較する類似度を用いて，比較を行った．単純類似度においては，チャート内のコミュニティのうち，約 45%が Yahoo!に対して 0.6 以上の類似度を持ち，特定のカテゴリと対応付けられることが判明した．対して，0.6 以上の類似度を持ちコミュニティに対応付けられるカテゴリの割合は約 20%であった．また，ウェブディレクトリの階層構造，およびチャートのグラフ構造を考慮した拡張類似度を提案し，その結果，コミュニティとカテゴリとを 50%以上対応付けられることが判明した．さらに類似度の低い部分についてはサンプリング調査を行い，ウェブコミュニティと Yahoo!の間には分類観点の異なる部分，および，どちらかにおいて詳細な分類が得られる部分があることを示した．

我々は，本研究を発展させて，ウェブコミュニティチャートとウェブディレクトリの差分情報を，ウェブディレクトリの保守および更新に応用する予定である．具体的には，ウェブディレクトリに対して，(1) 未登録ページを登録する，(2) 新しい観点に基づくカテゴリを作成する，および (3) 分類の粗い部分を詳細に分類する，などの推薦を行う．これらの目的を達成するためには，コミュニティにおける分類の Yahoo!に対する適合性の検証，コミュニティの分類精度の検証，および推薦目的に適した類似度の検討などを行っていく必要がある．また，本論文で用いた類似度は，全体的な概念構造の違いをとらえるにはまだ不足している．概念構造の詳細な比較検討については今後の課題としたい．

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．

参 考 文 献

1) Gibson, D., Kleinberg, J. and Raghavan, P.: Inferring Web Communities from Link Topology, *Proc. HyperText98*, pp.225–234 (1998).
2) Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D. and Kleinberg., J.: Automatic resource compilation by analyzing hyperlink structure and associated text, *Proc. 7th International WWW Conference*, pp.65–74 (1998).
3) Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the Web for emerging cyber-communities, *Proc. 8th WWW Conference*, pp.403–415 (1999).
4) Lempel, R. and Moran, S.: The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect, *Proc. 9th WWW Conference*, pp.387–401 (2000).
5) Flake, G.W., Lawrence, S. and Giles, C.L.: Efficient Identification of Web Communities, *Proc. KDD 2000*, pp.150–160 (2000).
6) Toyoda, M. and Kitsuregawa, M.: Creating a Web Community Chart for Navigating Related Communities, *Proc. Hypertext 2001*, pp.103–112 (2001).
7) 豊田正史，吉田 聡，喜連川優：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，電子情報通信学会論文誌 D-I，Vol.J87-D-I, No.2, pp.256–265 (2004).
8) Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment, *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pp.668–677 (1998).
9) Bharat, K. and Henzinger, M.: Improved Algorithms for Topic Distillation in a Hyperlinked Environment, *Proc. ACM SIGIR '98*, pp.104–111 (1998).
10) Chakrabarti, S.: Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation, *Proc. 10th WWW Confer-*

ence, pp.211–220 (2001).
11) Dean, J. and Henzinger, M.R.: Finding related pages in the World Wide Web, *Proc. 8th WWW Conference*, pp.389–401 (1999).
12) 市瀬龍太郎，武田英明，本位田真一：階層的知識間の調整規則の学習，人工知能学会論文誌，Vol.17, No.3, pp.230–238 (2002).

(平成 15 年 12 月 20 日受付)
(平成 16 年 4 月 6 日採録)

（担当編集委員　定兼 邦彦）

豊田　正史（正会員）

昭和 46 年生．平成 6 年東京工業大学理学部情報科学科卒業．平成 11 年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）．同年東京大学生産技術研究所博士研究員．平成 16 年より同所特任助教授．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，日本ソフトウェア科学会各会員．

吉田　　聡

昭和 52 年生．平成 13 年東京工業大学工学部電気電子工学科卒業．平成 15 年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（正会員）

昭和 30 年生．昭和 53 年東京大学工学部電子工学科卒業．昭和 58 年同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．同年同大学生産技術研究所第 3 部講師．現在同教授．平成 15 年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 SIGMOD Japan Chapter Chair，平成 9 年，10 年本学会データ工学研究専門委員会委員長．VLDB Trustee（1997 年〜2002 年），IEEE ICDE，PAKDD，WAIM 等ステアリング委員．