ウェブコミュニティとウェブディレクトリの比較に関する一考察

吉田　聡[†]　　豊田　正史[†]　　喜連川　優[†]

†東京大学生産技術研究所　〒153-8505 東京都目黒区駒場 4-6-1

E-mail:　†{achika, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし 近年，ウェブのリンク構造を解析することにより同じ話題を持つページの集合を抽出する研究が盛んである．この集合はウェブコミュニティと呼ばれている．我々は大規模なウェブのアーカイブから全てのコミュニティを抽出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，その精度の検証は容易ではない．一方 Yahoo!のようなウェブディレクトリは人手で分類を行うため高い精度が期待できるが，分類できるページ数は少なくなる. 本論文では自動的に抽出されたコミュニティの分類にどのような傾向があるかを評価すべく，第一歩として Yahoo! Japan とコミュニティの比較を行い，類似点や相違点を元に考察を行った.

キーワード　ウェブ，ウェブコミュニティ，リンク解析，ウェブディレクトリ

Comparison between Web Communities and Web Directory

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

† Institute of Industrial Science, The University of Tokyo　4-6-1 Komaba Meguro-ku, Tokyo, 153-8505 Japan

E-mail:　†{achika, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract In recent years, various methods for extracting a set of web pages with a common interest on a specific topic have been developed. The set of web pages is called as web community. We have been developed a method to extract web communities from web archives. Although these methods enable us to classify a huge amount of web pages automatically, it is difficult to verify the accuracy. On the other hand, since web directories like Yahoo! Japan classify web pages manually, we can expect them to have high accuracy. But they can classify fewer pages than web communities. In this paper we compare a set of web communities with a web directory and to clarify differences between them.

Keyword Web, Web Community, Link Analysis, Web Directory

1. はじめに

近年, WWW(World Wide Web)は爆発的な成長を続けてきた．日々大量のウェブページが作成されていくため，利用者が膨大な情報の山から必要とする情報だけを探し出すには多大な労力が必要となる．よって，ウェブページを内容に応じて分類することは利用者の労力の軽減につながり，有益である．大量のウェブページを分類する方法として主にウェブコミュニティとウェブディレクトリの二種類を挙げることができる.

WWW を，ウェブページをノード，その間に張られたハイパーリンクをエッジとするグラフと見なし，そのグラフ構造を解析することによってウェブコミュニティを抽出する研究が最近盛んに行われ，様々な手法が提案されてきた[6]〜[13]．ここでいうウェブコミュニティとは，同じ話題に関心を持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．ウェブコミュニティの例としては，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．我々は大規模な日本のウェブスナップショットを元にウェブコミュニティを抽出する手法を開発してきた[5]が，その分類精度を検証することは，量の多さゆえ容易なことではない.

一方，ウェブページを人手によりツリー状に分類するウェブディレクトリは，人間が一つ一つのページに目を通したうえでその内容により分類されるためその精度は高いことが期待される．しかし，その処理能力は人間の能力による部分が多くウェブコミュニティによる分類に比べれば分類できるページ量は少なくなる．ウェブディレクトリの例としては，Yahoo![1]や Open Directory[3]等が挙げられる.

ウェブコミュニティはウェブディレクトリより分類しているページの量が多いため，ウェブディレクトリとウェブコミュニティを比較することで，類似する部分に対しては

ウェブコミュニティを参考にしてウェブディレクトリの登録ページを増やすことができると思われる．また，両者の分類の相違点を調査することでウェブディレクトリに対し新しい観点に基づくディレクトリの推薦，離れたディレクトリ間にシンボリックリンクを張る提案，分類の粗い部分に対しての細かい分類の推薦等を行うための情報が得られると期待される．これらの情報はウェブディレクトリの管理者にとってディレクトリを新鮮に保ち，分類を改善するために有用である．

そこで本論文では，ウェブコミュニティとウェブディレクトリの分類の傾向を分析すべく，まず両者が共通して含む URL を取り出した．そして類似度を導入して両者の比較を行い，全体的な類似度を調べた．最後に類似度の低いウェブコミュニティを取り出してウェブディレクトリとの相違点を調査すると共に類似度の低いディレクトリを抽出してウェブコミュニティとの相違点を調査した．

比較対象としてウェブコミュニティはウェブコミュニティチャート[5]を利用し，ウェブディレクトリは国内最多のページ量を持つ Yahoo! Japan[2]を利用した．以下でコミュニティと書いた場合はウェブコミュニティを意味する．

本論文の構成は以下の通りである．第 2 節ではコミュニティの抽出と詳細及びウェブディレクトリについて述べる．第 3 節ではコミュニティとディレクトリの比較結果について述べ，第 4 節ではその中で類似度の低いコミュニティとディレクトリに関して手作業で分類を行った結果を述べる．第 5 節で得た結果について議論を行い，そして第 6 節でまとめを行う．

2. コミュニティとウェブディレクトリの特徴

本節では，本実験に用いた約 18 万個のコミュニティの抽出方法と特徴，および約 3 万個のディレクトリと約 20 万個の URL を含む Yahoo! Japan の特徴について記す．

2.1. コミュニティについて

2.1.1. コミュニティの抽出方法

本実験で利用したコミュニティはウェブコミュニティチャートとして作成されている．ウェブコミュニティチャートとは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きのエッジを張ったグラフである．

ウェブコミュニティチャートの作成に，我々は関連ページアルゴリズムを利用している．関連ページアルゴリズムは，１つのシードページを入力として与えると，シードページの近傍のウェブグラフから，良い authority ページ及び良い hub ページを抽出し，上位の authority 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 authority とは多くの良い hub からハイパーリンクが張られている著名なページを表す．良い hub とは，リンク集及びブックマークなど，多くの良い authority へハイパーリンクを張っているページを表す．これらの循環した定義により，密に結合した hub と authority が抽出され，それらがよく関連したページを表すことが[4],[5]で示されている．

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け取り，チャートを結果として出力する．シードページとしては，外部のサーバからｎ本以上リンクが来ているページなどある程度有名なページを集める．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．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話題を共有することを[5]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．2 つのコミュニティのメンバ間に導出関係があるとき，それらのコミュニティの間にエッジが張られる．

2.1.2. コミュニティの特徴

実験には 2002 年 2 月に収集した日本のウェブページの約 4 千 5 百万ページからなるウェブスナップショットを利用した．日本のウェブページとは，.jp ドメイン内のページである．まずウェブスナップショットから URL と URL 間のリンクからなるウェブグラフを生成した．このウェブグラフの中にはスナップショットに含まれるウェブページからリンクを張られているスナップショットに含まれていない URL も含まれているため，約 8 千 4 百万 URL まで URL 数が増加している．また，jp ドメインではない URL も含まれている．

チャート生成アルゴリズムに与えるシードとしては 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選択した．結果としてシードセットの大きさは 150 万 URL ほどになった．そして，シードセットを入力として，2.1.1 節で述べた方法で 18 万個のコミュニティから構成されるコミュニティチャートを作成した．

図 1 がコミュニティの大きさの分布を示している．

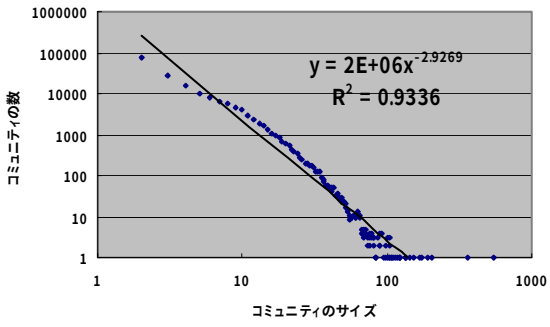

図 1　コミュニティのサイズの分布

トップディレクトリ	ディレクトリ	URL	平均URL数
Arts	1607	13449	8.37
Business_and_Economy	12132	71487	5.89
Computer_and_Internet	867	6888	7.94
Education	404	3769	9.33
Entertainment	5073	27579	5.44
Government	387	3996	10.33
Health	794	5978	7.53
News	625	3715	5.94
Recreation	3750	22483	6.00
Reference	100	1691	16.91
Regional	1805	5588	3.10
Science	1575	9386	5.96
Social_Science	723	3636	5.03
Society_and_Culture	1221	20902	17.12
合計	31072	200541	6.45

表 1　Yahoo! Japan のトップディレクトリごとの登録 URL 数

2.2. Yahoo! Japan の特徴

Yahoo! Japan は世界最大のウェブディレクトリである Yahoo!の日本語版であり，日本最大のウェブディレクトリである.比較に使用するデータとして 2002 年 9 月に収集を行った.

Yahoo! Japan には約 3 万 1 千個のディレクトリの中に約 20 万ページが掲載されている．ただし，複数のディレクト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! Japan のトップディレクトリごとの状況を表 1 にまとめた．URL 数では圧倒的に/Business_and_Economy ディレクトリが多く内部のディレクトリ数も多い．また，/Society_and_Culture ディレクトリでは，他のディレクトリに比べて 1 ディレクトリ当たりの URL 数が大きいが, これは「個人ホームページ(/Society_and_Culture/ People/ Personal_Home_Pages/)」というディレクトリにおいて索引別と総リストで 2 重に登録されている URL の数が約 1 万個あるためである.

図 2 はディレクトリサイズの分布を示している．個人ホームページの総リストとして約 1 万ものページが掲載されているディレクトリが一つ存在していることがグラフ 2 の右下から読み取れる．また，分類は人間が行っているにも関わらずサイズの分布はほぼべき乗則に従っている.また，コミュニティと異なり URL を 1 個しか含まないディレクトリも多い．なお，URL の数え方は単純に中間ノードも末端ノードもそのノードに含まれている URL 数のみを数えている.

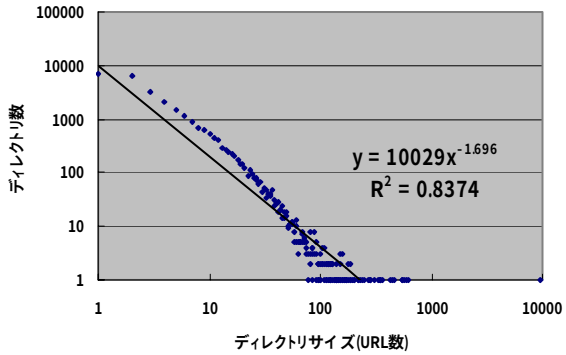

図 2　Yahoo! Japan における各ディレクトリのサイズの分布

3. コミュニティと Yahoo! Japan の比較

この節では，ウェブコミュニティチャートと Yahoo! Japan を比較し，その結果を示す．最初に記した様に，ウェブコミュニティチャート全体を評価することは不可能に近い．そこで，ウェブコミュニティチャートに含まれ，なおかつ Yahoo! Japan にも含まれる URL を取り出して, 類似点及び相違点を調査した．最初に単純なコミュニティとディレクトリの比較を行い，それから Yahoo! Japan の階層を考慮した比較を行った.

まず，この節で用いる記号を以下に示す.

u_i : URL

$c_1, c_2, \ldots$: コミュニティ． c_i は URL の集合.

$y_1, y_2, \ldots$: Yahoo! Japan のディレクトリ.

y_i は URL の集合.

C : ウェブコミュニティチャートに含まれる URL の集合.

$$C = (c_1 \cup c_2 \cup c_3 \cup \ldots \cup c_n)$$

Y : Yahoo! Japan に含まれる URL の集合.

$$Y = (y_1 \cup y_2 \cup y_3 \cup \ldots \cup y_m)$$

c_i' : コミュニティ c_i と Yahoo! Japan の共有部分. $c_i' = c_i \cap Y$

y_i' : ディレクトリ y_i とウェブコミュニティチャートの共有部分． $y_i' = y_i \cap C$

3.1. コミュニティと Yahoo! Japan の共有 URL

Yahoo! Japan 側に存在し，なおかつ 2.1.2 で説明したウェ

ブグラフ側にも存在する URL 数は約 15 万 1 千個で，ウェブグラフには含まれない URL は約 2 万 6 千個だった．ウェブグラフに含まれないが Yahoo! Japan に存在するウェブページの一例としては，ウェブグラフの生成から Yahoo! Japan の収集を行うまでの 7 ヶ月間で発生したページや，Yahoo! Japan 以外からあまりリンクの張られていないページが挙げられる．後者の例としては規模の小さい企業のサイトにそのような傾向が見られた．

また，ウェブコミュニティチャート上に存在し，Yahoo! Japan にも存在する URL の数 $|C \cap Y|$ は約 8 万 1 千個だった．約 7 万個のページが Yahoo! Japan とウェブグラフに存在するがコミュニティに存在しないことになる．これは，Yahoo! Japan は基本的に登録を依頼しなければ登録されないため，リンクが多く張られているページでも Yahoo! Japan に掲載されていないということがあると同時に登録制であるために有名でないページも登録される可能性があるためである．以下では，ウェブコミュニティチャートと Yahoo! Japan の双方に存在する URL を取り扱う．

図 3 は，各コミュニティが Yahoo! Japan と共有する URL 数の分布図である．逆に図 4 は各ディレクトリがコミュニティと共有する URL 数の分布図である．どちらもべき乗則に従う分布になった．後者のグラフより 1000 以上の URL をコミュニティと共有するディレクトリが存在することがわかる．これは 2.3 で説明した「個人ホームページ」ディレクトリである．

3.2. コミュニティとウェブディレクトリの単純な比較

3.2.1. 単純類似度の定義

次に，あるコミュニティに含まれる URL がどの程度同一の Yahoo! Japan のディレクトリに含まれているのか，ということを調べた．以下で定義する類似度は一つのコミュニティと Yahoo! Japan の分類がどの程度類似しているかを示すものである．

あるコミュニティ c_i の Yahoo! Japan に対する類似度 $R(c_i)$ を以下に定義する．

$$R(c_i) = \frac{|c_i' \cap y_k'|}{|c_i'|}$$

ただし y_k' は $|c_i' \cap y_k'|$ が最大となる k をとる．

また，ある Yahoo! Japan のディレクトリ y_i のウェブコミュニティチャートとの類似度 $R(y_i)$ を以下に定義する．

$$R(y_i) = \frac{|c_k' \cap y_i'|}{|y_i'|}$$

ただし c_k' は $|c_k' \cap y_i'|$ を最大とする k を取る．以下では，この方法で求めた類似度 $R(c_i), R(y_i)$ のことを単純類似度と呼ぶ．

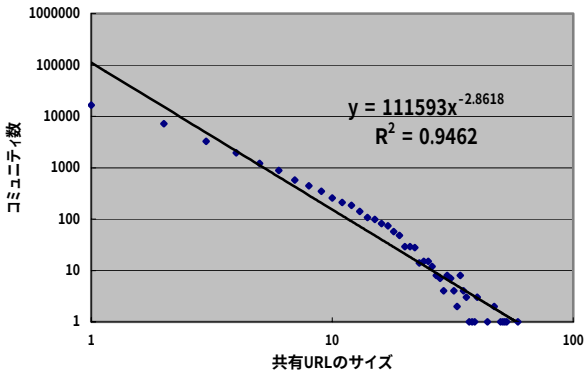

図 3　各コミュニティの Yahoo! Japan との共有 URL 数

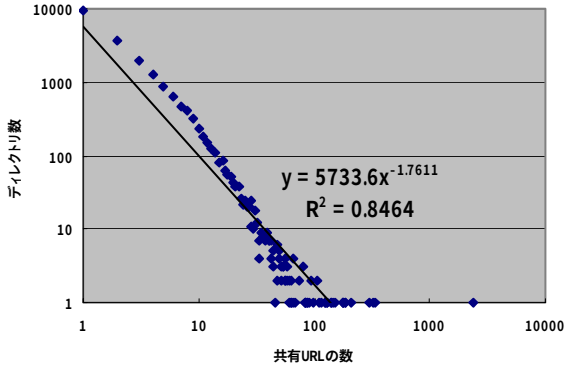

図 4　各ディレクトリのコミュニティとの共有 URL 数

3.2.2. 単純類似度による比較

図 5 の一番上の帯グラフは $5 \leq |c_i'|$ となる約 4 千コミュニティについて単純類似度を調べたものである．全体の約 40%のコミュニティが類似度 0.6 以上となった．それらのコミュニティでは Yahoo! Japan と共有する URL のうち 60%以上が一つの Yahoo!Japan のディレクトリに含まれている．

図 5 の下の帯グラフは逆に $5 \leq |y_i'|$ となる約 5 千ディレクトリについて類似度を調べたものである．こちらは類似度が 0.6 以上となったのは全体の 20%に過ぎず，0.4 未満が全体の 60%を占めている．この様なディレクトリでは共有する URL が様々なコミュニティに分かれてしまっている．

全体的に $R(c_i)$ が $R(y_i)$ より高い，つまりコミュニティの大部分が yahoo!の特定のディレクトリに含まれていることが多いと言う結果は Yahoo! Japan の方がコミュニティと比べて分類が大きいということを示している．これは $5 \leq |c_i'|$

となる c_i' の平均サイズが 8.13 であるのに対し，$5 \leq |y_i'|$ となる y_i' の平均サイズが 12.84 であることが一つの理由である．

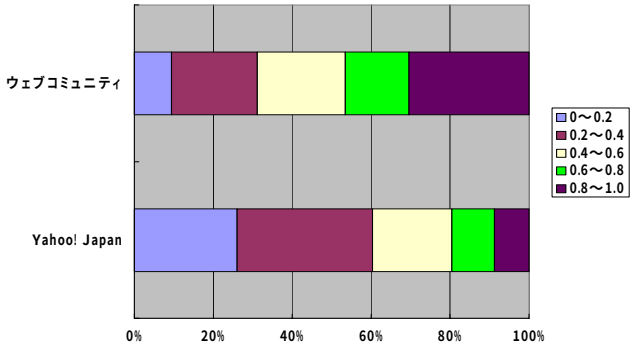

図 5　ウェブコミュニティと Yahoo! Japan の単純適合度の比較

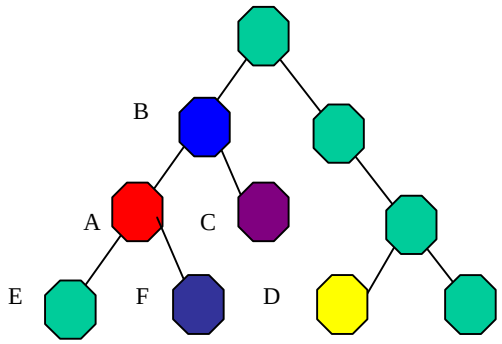

図 6 ディレクトリツリーの例

3.3. ディレクトリ階層を考慮した比較

3.3.1. 拡張類似度の定義

Yahoo! Japan 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ディレクトリは類似した話題を持つと考えられる．これを利用してあるコミュニティ c_i' が Yahoo! Japan とどの程度類似しているのかを計るための拡張類似度を定義する．

ここで，Yahoo! Japan のディレクトリ y_i と y_r の間の距離 $D(y_i, y_r)$ を y_i から y_r までディレクトリツリーをたどって到達するまでのステップ数と定義する．ただし，ステップ数は 1 から数える．つまり，$D(y_i, y_i) = 1$ となる．また，Yahoo! Japan ではディレクトリとディレクトリをつなぐシンボリックリンクが数多く存在するが，本論文では簡単のためその存在を無視する．

c_i' の定義より $c_i' = c_i \cap Y = (u_1, u_2, ..., u_n)$ となる．そして各 URL が Yahoo! J で所属するディレクトリを $(y_1, y_2, ..., y_n)$ とする．ここで $u_i \in y_i$ である．

まず，中心点となるディレクトリ y_k を $y_1, y_2, ..., y_n$ の中から決める． y_k は $\max(|c_i' \cap y_k'|)$ をとる y_k をとる．

そして，ディレクトリ距離を考慮した拡張類似度 $R_h(c_i')$ を以下に定義する．

$$R_h(c_i') = \frac{\sum_{i=1}^{n} (1/D(y_i, y_k))}{n}$$

拡張類似度 $R_h(c_i')$ は 0 から 1 までの値をとり，値が高いほどこのコミュニティは Yahoo! Japan に類似していることになる．

ここで，拡張類似度は距離を任意のディレクトリのペアに対して定義するが，トップディレクトリや親ディレクトリを通過して他のディレクトリに到達する場合，話題が類似する保証は無い．そこで距離に制限を加えた 2 種類の類似度の定義を行う．

- **トップ共有類似度**

この類似度は，トップディレクトリが共通していないディレクトリは類似した話題を持つ可能性が低いであろうという考えを元にしている．拡張類似度の計算を行う際，ディレクトリ間の距離の計算を中心ディレクトリと同一のトップディレクトリに分類されているもののみに行い，そうでないものは距離を無限大として扱った場合の類似度である．例えば図 6 のようなディレクトリ構造を考える．この場合 A が中心ディレクトリである時，A と D の間の距離は無限大となる．

- **部分木類似度**

この類似度は，距離の計算を中心ディレクトリから上がるだけ，または下がるだけで到達することが出来ることができるディレクトリに限定し，それ以外のディレクトリに対しては距離を無限大として扱った場合の類似度である．例えば図 6 において，A が中心ディレクトリである時，B,E,F に対しては距離の計算が行われるが C,D に対しては計算がなされずに無限大として扱われる．

3.3.2. ディレクトリ階層を考慮した比較結果

図 7 の下側 3 つのグラフが 3.3.1 節で定義した拡張類似度を用いて計算した類似度の分布である．拡張類似度では全体の約 55%が類似度 0.6 以上となっている．単純類似度では数値が悪くても，近くに URL が集まっているコミュニ

ティが約 10%存在したということになる．トップ共有類似度や部分木類似度を用いても，50%以上のコミュニティが類似度 0.6 以上となる．

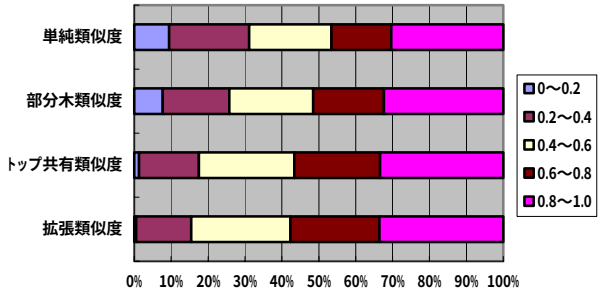

図 7　コミュニティから見た Yahoo! Japan ディレクトリの類似度の変化

4. 類似度の低いコミュニティとウェブディレクトリの調査

この節では，類似度が低いコミュニティと Yahoo! Japan のディレクトリをサンプリングして，コミュニティと Yahoo! Japan の相違点を調べた結果を記す．

4.1. コミュニティのサンプリング調査

単純類似度と拡張類似度，どちらの類似度を利用しても Yahoo! Japan との類似度が 0.6 未満になったコミュニティをランダムに 100 個サンプリングした．そしてサンプルコミュニティの中身を一つ一つ確認し，Yahoo! J 上では別々のディレクトリに分類されている各ページが実際には類似したトピックを持つのかどうかを検証した．

サンプルのコミュニティに話題があるかどうかを主観により判断し，その話題に合う URL 数が 6 割を超えるものが 44 個あった．その 44 個のコミュニティを中身により分類すると 3 通りに分かれた．

以下ではその 3 種類の分類について説明する．

なお，以下では/Business_and_Economy ディレクトリを B&E，/Society_and_Culture ディレクトリを S&C，/Business_and_Economy/ Business_to_Business ディレクトリを B&E/B2B, /Business_and_Economy/ Shopping_and_Service ディレクトリを B&E/S&S と省略することがある．

- **主観による分類で類似度が高いと見ることもできるコミュニティ　26 個**

類似している話題を扱っているにもかかわらず別のディレクトリに分類されているようなコミュニティをここに分類した．例えば企業のサイトと個人の趣味のサイトで類似している話題を扱っていても企業サイトは B&E，個人のサイトは Recreation に分けられるといったものである．文末にある表 4 で，類似した話題を扱っているにも関わらず別のトップディレクトリに分類された場合，どのトップディレクトリの組み合わせが多く出てくるのか，というのを示した．ただし，B&E は他のトップディレクトリに比べて URL が非常に多いので B&E/B2B,B&E/S&S,その他に分け，また Entertainment の中の「お勧めリンク(Cool_links)」も Entertainment とは別の集計にした．その中で 4 回出現した頻出の例を示す．

Arts と Entertainment の場合では，「星界の紋章」という作品に関するコミュニティで，Arts 側にこの作品の小説の著者である森岡浩之氏に関するページの分類があり，Entertainment 側では星界の紋章に関係のあるファンタジー系のイラストのページが分類されているという例があった．

B&E/B2B と S&C では，子供に関するコミュニティが存在した．B&E/B2B では NHK の「週間子供ニュース」という TV 番組のページがあり，S&C では子供を持つ親のための情報のページが存在した．

B&E/B2B と Science の場合では，天文学のコミュニティで B&E/B2B に企業が運営する「星座の博物館」と題した星座や天文学を説明するサイトがあり，Science に一般の天文学を扱うサイトが分類されている場合が存在した．

S&C と Science では，森林に関するコミュニティで，Science では森林科学から見たページが存在し，S&C では環境保護の観点から森林を見たページが存在していた．なお，最後の例では森林科学のディレクトリから林業のディレクトリへのシンボリックリンクが存在していた．このことにより，Yahoo! Japan もこれらのディレクトリの類似性を認識していることが分かる．

- **途中まで同じパスを持つコミュニティ　10 個**

Yahoo! Japan 上で所属するディレクトリが途中まで同一のパスを持ち，そこから下が異なるため単純類似度が低くなっているものの中で，同一のパスの部分だけで十分に話題が限定できるようなものがこの分類にあたる．この場合では，一致していない部分が長いと拡張類似度を利用してもほとんど類似度は改善されない．以下に例を示す．

  - /B&E/S&S/General_Merchandise/Imported/Asia/Indonesia/ (輸入雑貨・インドネシア)
  - /B&E/S&S/General_Merchandise/Imported/Asia/Thailand/ (輸入雑貨・タイ)
  - /B&E/S&S/General_Merchandise/Imported/Latin_America/Mexico/ (輸入雑貨・メキシコ)
  - /B&E/S&S/General_Merchandise/Thematic/Animals/Dolphins/ (雑貨・イルカグッズ)
  - /B&E/S&S/General_Merchandise/Variety/ (雑貨・バラエティ)

この例では/B&E/S&S/General_Merchandise/までパスが一致している．このパスは雑貨屋という分類を示し，これだけで十分に話題が限定できると判断した．

- **話題は存在するが, ウェブディレクトリ側に該当するディレクトリが存在しないコミュニティ　8 個**

これらのコミュニティには，確かに一致する話題を発見できるのだが，Yahoo! Japan 側に当てはめるべきディレクトリが存在しないというものである．以下にその話題のリストを記す．

- 立命館大学大学院各専攻の公式ページ群
- 岡山大学各学科の公式ページ群
- コンピュータ系の各イベントのページ群
- 広島の公共施設のホームページ群
- 箱根にある博物館・美術館のページ群
- 京都の公共施設のページ群
- 群馬県の美術館等のページ群
- 北海道から東北の博物館・美術館

ただし，立命館大学については現時点(2003 年 1 月)において新しく「立命館大学」のディレクトリが作られ，そこに分類されていた．収集を行った 2002 年 9 月の時点でそのディレクトリが存在しなかったことは確認した．この分類について我々は Yahoo! Japan を先取りしていたことになる．

4.2. ディレクトリのサンプリング調査

単純類似度が 0.6 未満であるディレクトリを，その類似度が低い理由を調査するためにランダムに 100 個サンプリングした．そしてサンプルディレクトリの中身を一つ一つ人手により確認した．

サンプリングしたディレクトリを分類した結果としては 3 種類に分けることができた. 以下で各項目の説明を行う．

- **コミュニティの方が詳細　36 個**

ディレクトリと URL を共有するコミュニティが複数に分かれ，各々はディレクトリより細かい話題を持っているものがここに分類された．例えば、http://dir.yahoo.co.jp/Science/Biology/Botany/Plants/ は植物について扱ったページを集めたディレクトリであるが、コミュニティ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ページ、植物写真に関するページに分類されていた．

- **分類観点が異なる　30 個**

ディレクトリ側とコミュニティ側で分類する観点が異なると判断されたものがここに分類された．例えば、http://dir.yahoo.co.jp/Business_and_Economy/Business_to_Business/Trade/General_Trading_House/Marubeni_Corp_/ は総合商事の丸紅社のディレクトリであるが、ディレクトリ側では丸紅社が運営しているページの一覧が掲載されているのに対し、コミュニティ側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分類されていた。

- **コミュニティの分類精度が悪い　34 個**

コミュニティ側が同じ話題を持つ複数のコミュニティに分裂しているため，結果としてディレクトリの類似度が低くなってしまったと判断したものがここに分類された．例えば、http://dir.yahoo.co.jp/Science/Physics/High_Energy_and_Particle_Physics/ は高エネルギー物理学・素粒子物理学に関するページが含まれているディレクトリであるが、コミュニティ側では５つのコミュニティに 1 ページずつ含まれているという状況である。そして、各コミュニティを見てもそのコミュニティの間で明らかな相違が見つからないという状態であった。このため、この部分においてはコミュニティ側の分類が悪いと判断した。

5. 議論

今節ではこれまでに示した実験結果について議論を行う．

3.2.2 節で，コミュニティの約 50%は Yahoo! Japan に対し 0.6 以上の類似度を持つが，逆に Yahoo! Japan で 0.6 以上の類似度を持ったディレクトリは全体の 20%に過ぎなかったという結果を示した．これはコミュニティの大部分が yahoo!の特定のディレクトリに含まれていることが多いことを示している．この結果により Yahoo! Japan の方がコミュニティと比べて分類が広いということがいえる．

4.1 節で「話題は存在するが Yahoo! Japan に該当するディレクトリが存在しないコミュニティ」という部分で立命館大学のディレクトリが新しく作られていたことを示した．このようなコミュニティを抽出して，Yahoo! Japan に対し新しい分類の推薦が出来る可能性がある．

コミュニティの分類と Yahoo! Japan の分類が相違を持つ理由の一つとして，Yahoo! Japan ではウェブページを製作した組織や個人についての分類が主となっている部分があることが挙げられる．一方ウェブコミュニティチャートは多くの hub, つまりリンク集の分類をまとめた結果であり，また普通のリンク集では特定の話題を扱うサイトに対しそのサイトが個人のものか企業のものかといったことはあまり気にされないことが多い．これは，どちらの分類が良いとかという問題ではなく両者の観点の違いによるものである．

4.1 節の「主観による分類で類似度が改善するコミュニティ」の中で，トップディレクトリが異なる部分に分類されている例をいくつか挙げたが, 3.3.1 節で記した様に Yahoo! Japan にはディレクトリ間をつなぐシンボリックリンクが数多く存在し，その中には異なるトップディレクトリの間をつなぐものも存在する．そのため，実際にはごく近い場所にあるにも関わらず，ディレクトリパスだけを見て全く別の場所に存在していると判断してしまうという例も有り得る．森林科学から林業へとシンボリックリンクが張られていた例も実際に存在した．逆に，類似した話題を持つにも関わらずお互いの間にシンボリックリンクが張られていないようなディレクトリも多く，新しくシンボリックリンクを張るという提案を行うのに使える可能性が存在する．

4.2 節の「コミュニティの方が詳細」という部分で Yahoo! Japan の分類が粗い部分を示したが，このような分類の粗

い部分を，類似度の低い部分を調査することにより発見できる可能性が示された．また,「分類観点が異なる」という部分で Yahoo! Japan とウェブコミュニティチャートの分類の観点が異なる部分を示したが，この事はコミュニティ，Yahoo! Japan 双方に別の視点による分類の示唆を行っていると考えられる．

6. まとめと今後の課題

本論文では，コミュニティとウェブディレクトリを比較することによって双方の分類の傾向を調査した．両者が共通して含む URL を取り出してウェブディレクトリとコミュニティのお互いに対する類似度を調べ，コミュニティの約 40%, ウェブディレクトリの約 20%が相手側と類似しているという結果を得た．また，拡張類似度を導入するとコミュニティの約 55%が Yahoo! Japan に類似しているという結果を得た．そして類似度の低いコミュニティ及びディレクトリに対しサンプルを抽出してその内容を調査した. 100 個中 44 個のコミュニティに対しては, 別の観点から見て類似度が高いという結果を得た．

今後の課題としては，本論文では導入した拡張類似度においてシンボリックリンクの存在を無視していたが，実際にはシンボリックリンクを考慮することにより 2 つの Yahoo! Japan ディレクトリの距離がより近づく事例が存在する．そこで，各ディレクトリをノード，ディレクトリ間のリンクをエッジとしたグラフを作成し，その中で幅優先探索を行うことでディレクトリ間の距離をより正確に求めることを考えている．また，今回は拡張類似度をコミュニティ側にのみ導入したが，ウェブコミュニティチャートのエッジを利用してウェブディレクトリ側にも導入することを検討している．

文　献

[1] Yahoo! http://www.yahoo.com

[2] Yahoo! Japan http://www.yahoo.co.jp

[3] Open Directory Project http://dmoz.org/

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

[5] M. Toyoda and M. Kitsuregawa, Creating a Web Community Chart for Navigating Related Commu- nities. In Proc. Hypertext 2001, pages 103-112, 2001.

[6] 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.

[7] K.Bharat and G. A. Mihaila. When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics. In Proc. 10th WWW Conference, 2001.

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

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

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

[11] R. Kumar, P.Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In Proc.25th VLDB Conference, 1999.

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

[13] S. R. Ravi Kumar, Prabhakar Raghavan and A. Tomkins. Trawling the Web for emerging cyber-communities. In Proc. 8th WWW Conference, 1999.

	Arts	B2B	Other	S&S	C&I	Cool	Edu.	Ente.	Gov.	Hea.	News	Pers.	Rec.	Ref.	Reg.	S&C	Sci.	Soc.
Arts		1	1		1	1		4			1	1	2	1	2	2		1
B&E/B2B				2		1		3		1		1			1	4	4	
B&E/Other				1				1								1	1	
B&E/S&S							2	2	1	1		1	2			3	3	
C&I						1								1				1
Cool_links														1		2	2	1
Education									1	1						1		
Entertainment											2	2	1		3	3	1	
Government										1				1		1		1
Health														1		3	1	1
News												1				1		
Personal																	1	
Recreation															2	1		
Reference																1		2
Regional																1		
S&C																	4	1
Science																		
Social_Science																		

表 4 主観による分類で類似度が高いと見ることも出来るウェブコミュニティが含む URL が複数のトップディレクトリに分類されていた場合のその組み合わせの度数