データベースシステム　122－40
(2000. 7. 27)

WWWにおける関連コミュニティ群の発見

豊田 正史

東京大学生産技術研究所
〒106-8558 東京都港区六本木 7-22-1
電話番号: 03-3402-6231(ext.2358)
E-mail: mtoyoda@acm.org

あらまし　本稿ではWWWにおいて、ユーザから与えられたwebページ(シードページ)から、それに関連するコミュニティ群を発見する手法を提案する。ここで言うコミュニティとは、同じトピックに関心をもつ人々によって作成されたwebページの集合を指す。例えば、ある野球チームを応援するホームページの集合を、そのチーム応援ページのコミュニティと呼ぶ。本手法は、与えられたwebページを含むコミュニティ、および、関連するトピックに関する複数のコミュニティを検索することができる。例えば、ある野球チームの応援ページを1つ与えると、まずその野球チームの応援ページコミュニティを、次に各野球チームの公式ページコミュニティを発見する。この検索は与えられたwebページの周辺における、ハイパーリンクの解析を基に行う。シードに関連するコミュニティを1つだけ出力する従来手法に比べて、本手法はシードに関連する複数のコミュニティをユーザに提供することができる点で優れている。

キーワード　WWW、コミュニティ、リンク解析

Finding Related Communities on the Web

Masashi Toyoda

Institute of Industrial Science, University of Tokyo
7-22-1 Roppongi Minato-ku Tokyo, 106-8558, JAPAN
Phone: 03-3402-6231(ext.2358)
e-mail: mtoyoda@acm.org

Abstract　We propose a new web search technique, which finds related communities from a given URL. A community is a set of web pages written by authors who have a common interest on a specific topic, such as fan pages of a professional baseball team. Our technique finds a community that includes a given URL, and communities on related topics, using hyperlink analysis. For example, when a fan page of a baseball team is given, our technique finds a fan community of the team, and a community of official homepages of baseball teams. Our technique allows the user to perform new type of navigation through the web. It provides additional ways not only to related pages, but also to related communities.

key words　WWW, community, link analysis

1 はじめに

本稿ではWWWにおいて、ユーザから与えられたwebページ(シードページ)から、それに関連するコミュニティ群を発見する手法を提案する。ここで言うコミュニティとは、同じトピックに関心をもつ人々によって作成されたwebページの集合を指す。例えば、ある野球チームを応援するホームページの集合を、そのチーム応援ページのコミュニティと呼ぶ。本手法は、与えられたwebページを含むコミュニティ、および、関連するトピックに関する複数のコミュニティを検索することができる。例えば、ある野球チームの応援ページを1つ与えると、まずその野球チームの応援ページコミュニティを、次に各野球チームの公式ページコミュニティを発見する。以下は本手法を用いて実際に検索を行った例である。千葉ロッテマリーンズのファンページをシードとして、マリーンズファンのコミュニティ、および、マリーンズの公式ページを中心としたプロ野球チームの公式ページコミュニティが発見されている。

シードページ:
http://www2.justnet.ne.jp/~marinesw/index.htm
千葉ロッテマリーンズのファンページ

コミュニティ1 (マリーンズファンのコミュニティ):
http://www2.justnet.ne.jp/~marinesw/index.htm
http://www.mars.dti.ne.jp/~tazaki/
http://www.tatsuya.pos.to/marines/
http://member.nifty.ne.jp/maru_no/
http://www.netlaputa.ne.jp/~marines/
http://www2s.biglobe.ne.jp/~motoi-/clm/yokochou.htm
http://marines.pos.to
http://www.marines.co.jp/

コミュニティ2 (プロ野球チームの公式ページコミュニティ):
http://www.marines.co.jp/
http://www.buffaloes.co.jp/
http://www.orix.co.jp/BW/
http://www.seibu-group.co.jp/lions/index.html
http://www.fighters.co.jp/
http://www.hawkstown.com/
http://www.hanshin.co.jp/tigers/
http://ifcnet.ne.jp/baystars/

この検索は与えられたwebページの周辺における、ハイパーリンクの解析を基に行う。シードに関連するページのリストを1つだけ出力する従来手法に比べて、本手法はシードに関連する複数のコミュニティをユーザに提供できる点で優れている。

以降、第2節では、関連研究について述べ、第 2節、第4節では、関連コミュニティ群発見手法の詳細について解説する。第5節では、ユーザテストによる評価実験について説明し、第6節で、まとめと今後の課題について述べる。

2 関連研究

シードページからそれに関連するページを検索する手法は、すでにいくつか提案されている。これらの手法は、各webページを点、ハイパーリンクを辺と見立てた有向グラフとしてWWWをとらえ、グラフの構造を分析する。

Klienbergが提案したHITS[3, 2]は、WWWの適当な部分グラフの中からハイパーリンクで密に結合されたauthorityおよびhubページを抽出するアルゴリズムである。Authorityとは、多くのページからハイパーリンクを張られている著名なページを表す。Hubとは、リンク集およびブックマークなど、多くの著名なページへハイパーリンクを張っているページを表す。

HITSは、Altavistaを用いた検索の改善手法として提案された。キーワードによる検索結果の上位200ページ、およびそれらのページに(辺の向きを無視して)隣接しているページ、を含む部分グラフを与えると、各webページのauthorityおよびhubスコアを計算する。直観的にはauthorityスコアは良いhubから多く指されている程高くなり、hubスコアは良いauthorityを多く指している程高くなる。これらのスコアは簡単な繰り返し計算によって求められる。

HITSは、シードページに関連するページを検索するのに利用することもできる。シードページから(辺の向きを無視して)距離2にあるページを含む部分グラフを与えると、高いauthorityスコアを持つページをシードに関連しているページとみなすことができる。

Dean、HenzingerによるCompanion[1]は、シードページに関連するページを探す目的のためにHITSを特化したアルゴリズムである。(1)ページ内のリンクの順番を考慮する、(2)リンクにウェイトを付加する、ことにより精度を向上させている。

我々は、これらの関連ページ発見手法が出力する上位のauthority、hubの集合を、WWWにおけるコミュニティであると定義する。本論文で提案する関連コミュニティ群発見手法は、関連ページ発見手法を拡張することで、シードページ周辺において、お互いに関連しあう複数のコミュニティを発見することを可能にする。

Kumarらによる Trawling[4]は、テラバイト規模のwebアーカイブから、コミュニティを抽出する手法である。コミュニティの中には、幾つかのauthorityとhubからなる完全2部グラフが含まれると仮定し、大規模なwebグラフから完全2部グラフを抽出する問題に帰着している。この手法では、大規模なコミュニティのリストを得ることができるが、コミュニティ間の関連については考慮されていない。

3 関連ページ発見アルゴリズム

我々の関連コミュニティ群発見手法は、第2節で述べた関連ページ発見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、HITS、Companion、および、Companion+、の3種類の関連ページ発見アルゴリズムを一般化して解説し、これらの相違を示す。Companion+は我々が、Companionを改良したものである。

関連ページ発見アルゴリズムは、シードページ(複数でも可)を入力とし、上位N個のauthorityおよびhubページを出力する。アルゴリズムの手順を以下に示す。各手順は、アルゴリズムの種類によって異なるため、以下の節ではその詳細について解説する。

1. 近傍グラフの作成
2. ミラーページの削除
3. 各リンクのウェイト決定
4. authorityおよびhubスコアの計算

3.1 近傍グラフの作成

シード周辺の部分グラフを作成する。各手法とも、webサーバ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、各手法において以下のように異なる。

HITSでは、リンクの向きを無視して、シードから距離2にあるページ全て、および、それらのページの間に張られているリンクを全て近傍グラフに含める。

Companionでは、シード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)、および、シード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)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。

すべての手法において、あるページからの逆方向リンク数が$MaxIn$を越える場合には、ランダムに$MaxIn$個のリンクを選び、それだけをたどる。

第5節の評価実験では、$R=10$、$MaxIn=2000$を採用している。これは、経験的に決定したもので、本来はユーザテストによって決定すべきである。

3.2 ミラーページの削除

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。ミラーかどうかは、2つのページがリンクをどの程度共有しているかで判断する。第5節の評価実験では、2つのページが80%以上リンクを共有している場合にはミラーと判断し、片方を削除している。

3.3 各リンクのウェイト決定

近傍グラフに含まれる各リンクに、0〜1の範囲でauthorityウェイトおよびhubウェイトを与える。Authority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authorityスコアを与えるか、に影響する。Hubウェイトは、リンク基のページに対してどの程度のhubスコアを与えるか、に影響する。

HITSでは、リンクに対すしてウェイトを定義していないため、すべてのリンクのauthorityおよびhubウェイトを1とする。

Companionでは、基本的にはauthorityおよびhub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authority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hub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。

Companion+では、あるページの中にシードページを指すリンクl_sがある場合、l_sのauthorityウェイトを1とし、ページ内でl_sに近い位置にあるリンクのauthorityウェイトを$(R-l_s$からの距離$)/R$とする。シードを指すリンクが複数ある場合には、最大のウェイトを選択する。これはシードを指すリンクから離れる程影響を少なくする効果を持つ。その他のリンクのauthorityウェイトは0とする。また、hubウェイトはすべて1とする。この上で、Companionと同様にサーバによるウェイト調整を行う。

3.4 authorityおよびhubスコアの計算

あるページnのauthorityスコア(auth(n))とhubスコア(hub(n))は、以下の手順で算出する:

全てのページのauth(n)、hub(n)を1とする。
auth(n)、hub(n)が収束するまで以下を繰り返す。
$auth(n) = \sum_{m(m \Rightarrow n)} authority\_weight(m,n) hub(m)$
$hub(n) = \sum_{m(n \Rightarrow m)} hub\_weight(n,m) auth(m)$
($x \Rightarrow y$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。

最後に、auth(n)、hub(n)それぞれについて上位Nページを出力する。

4 コミュニティ群発見アルゴリズム

コミュニティ群発見アルゴリズムは、シードページを入力とし、authorityのリストを複数、コミュニティとして出力する。本手法の基本的なアイデアは、シードに関連ページ発見アルゴリズムを適用して得たauthorityそれぞれを、新たにシードとして関連ページ発見アルゴリズムに適用することで、選択的に近傍グラフの範囲を広げて行くことにある。こうして広げた近傍グラフから複数のコミュニティを発見する。以下に、コミュニティ群発見アルゴリズムの詳細を述べる。

1. シードページに対して関連ページ発見アルゴリズムを適用する。A_0を上位N個のauthorithyを含むリストとする。
2. A_0に含まれるURL$(a_1, ..., a_N)$をそれぞれ別々にシードとして関連ページ発見アルゴリズムに与えることで、N個の関連ページリストを得る。各々の結果から上位N個のauthorityおよびhubを取り出し$A_1, ..., A_N$、$H_1, ..., H_N$とする。
3. $a_1, ..., a_N$を幾つかのグループに分類する。目的は似かよった関連ページを生成するシードを1つにまとめることである。本手法では、$A_n \bigcup H_n$と$A_m \bigcup H_m$が閾値T個を越えてURLを共有している場合、a_nとa_mを同じグループに分類する。まず、閾値を1として分類を行い、グループが複数できない(1つにまとまってしまう)場合には、閾値を1増やして再度分類を行う。これをグループが複数できるまで繰り返す。こうして得られたURLのグループを$G_1, ..., G_M$とする。
4. 各G_xについて、G_xに含まれる(複数の)URLをシードとして関連ページ発見アルゴリズムを適用する。G_xに含まれるシード、および、関連ページ発見アルゴリズムによる上位のauthorityを併せて合計でN個取り出し、コミュニティとして出力する。この際、どれがシードページか分かるようにする。

このアルゴリズムは、A_0の中に、他とは異なる関連ページを生成するURLが入っている場合に、複数のコミュニティを発見できる。第1節で示した例では、A_0の中に、マリーンズのファンページの他に、マリーンズの公式ページが含まれていた。公式ページは、多くのhubページにおいて他チームの公式ページと密に繋がっているため、これをシードとするとマリーンズのファンは現れず、野球チームの公式ページコミュニティが現れる。この結果、このコミュニティはファンコミュニティとは切り離されることになる。これが、コミュニティ群発見の基本的なメカニズムである。

5 評価実験

本手法の有効性を評価するため、大学関係者を対象として2つのユーザテストを行った。テストの内容は、HITS、Companion、Companion+の精度比較、およびコミュニティ群発見アルゴリズムの性能評価である。

5.1 被験者

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WWWを日常的に利用している。

5.2 データセット

Webロボットを使って収集したwebページを、データセットとして用いた。データの詳細は以下の通りである。

ページ数: 約17,000,000ページ(HTML文書のみ、90GB)

収集期間: 1999年7月～9月

収集条件: jpドメインにあるページ。他ドメインにあっても日本語を含めば収集する。

収集の起点: http://www.yahoo.co.jp/

収集の戦略: 幅優先

収集したすべてのwebページからハイパーリンクを抽出し、webのグラフデータベースを作成した。これを用いると、URLをキーとして、キーページからリンクを張られているページのURL、および、キーページにリンクを張っているページのURLを高速に検索できる。

グラフデータベースには、収集したページのURLに加え、収集されていないがリンクを張られていることが分かっている約21,000,000ページのURLも登録してある。合計で約38,000,000のURLが検索可能となっている。また、ハイパーリンクはwebサーバ間に張られているもののみ、約23,000,000を登録している。

5.3 シードセット

被験者から、「過去にあるトピックに沿って集めたことのあるwebページ」または「関連する情報を集めたいと思うwebページ」を募集した。各被験者から、1～4個の別々なトピックに関するシードページを回収し、合計24ページを得た。

URL	HITS				Companion				Companion+			
	○	△	−	精度	○	△	−	精度	○	△	−	精度
archives.math.utk.edu/popmath.html	4	4	0	0.60	5	2	1	0.67	4	3	0	0.55
home.att.ne.jp/green/asj	5	1	1	0.61	6	2	0	0.70	6	1	0	0.65
islamcenter.or.jp/	6	1	0	0.65	6	1	0	0.65	4	3	0	0.55
lang.nagoya-u.ac.jp/~matsuoka/Japan...	0	0	0	0.00	0	0	0	0.00	4	3	2	0.69
plaza.harmonix.ne.jp/~kamao/	0	0	0	0.00	0	2	0	0.10	6	1	1	0.72
weather.is.kochi-u.ac.jp/	2	4	0	0.40	2	3	0	0.35	9	0	1	1.00
www2e.biglobe.ne.jp/~TKG/	0	0	1	0.00	0	0	0	0.00	7	1	2	0.94
www2j.biglobe.ne.jp/~tatuta/	0	0	0	0.00	0	0	0	0.00	6	3	0	0.75
www3.famille.ne.jp/~s370902/camera/...	2	6	0	0.50	1	0	1	0.11	1	3	0	0.25
www.alc.co.jp/nihongo/nihongo1.html	5	0	1	0.56	7	3	0	0.85	8	0	0	0.80
www.i-kochi.or.jp/prv/kochi/	0	1	0	0.05	1	1	0	0.15	8	0	2	1.00
www.isp.ne.jp/~nakajima/index.html	0	0	0	0.00	0	2	0	0.10	4	4	2	0.75
www.japan.msf.org/	5	3	1	0.72	4	4	2	0.75	4	5	1	0.72
www.maccentral.or.jp/pokemon/	1	1	0	0.15	4	5	0	0.65	5	5	0	0.75
www.mahjong.or.jp/	0	2	1	0.11	0	0	0	0.00	5	4	1	0.78
www.mars.dti.ne.jp/~o-shin/	2	2	0	0.30	3	5	0	0.55	5	5	0	0.75
www.ops.dti.ne.jp/~glass/	1	4	2	0.38	2	5	2	0.56	9	1	0	0.95
www.panda.org/	2	7	0	0.55	3	7	0	0.65	3	6	1	0.67
www.peugeot.co.jp/	10	0	0	1.00	10	0	0	1.00	10	0	0	1.00
www.red-hell.com/	10	0	0	1.00	10	0	0	1.00	10	0	0	1.00
www.spice.or.jp/~mt0711/index.html	9	0	1	1.00	8	0	2	1.00	10	0	0	1.00
www.tintin.com/	9	0	0	0.90	9	0	1	1.00	8	0	2	1.00
www.triathlon.or.jp/	8	1	1	0.94	9	1	0	0.95	9	1	0	0.95
www.watch.impress.co.jp/pc/index...	1	4	0	0.30	1	5	0	0.35	5	4	0	0.70
平均	3.4	1.7		0.45	3.8	2.0		0.51	6.3	2.2		0.79

○: トピックが同じ URL の数、△: 関連するがトピックは異なる URL の数、−: 消滅した URL の数 (上位 10 位以内)
精度: ○ を 2 点、△ を 1 点としたときの満点に対する割合。ただし消滅したページは考慮しない。

表 1: HITS、Companion、Companion+の比較

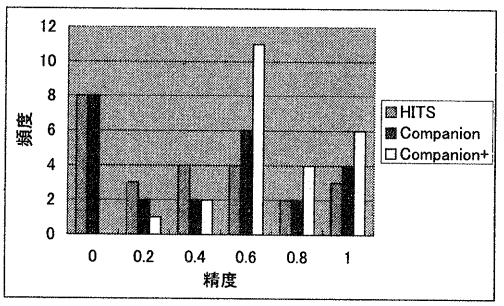

図 1: 精度のヒストグラム

5.4 HITS、Companion、Companion+の精度比較

まず、シードセットの各 URL に対して、HITS、Companion、Companion+を適用し、上位 10 の authority リストを作成した。シードを提供した被験者に対応するリストを、CGI を用いたアンケートフォームの形で渡し、以下のような指示で主観評価を依頼した。

- リストに含まれる各ページを実際にブラウズして、シードページにどの程度関連しているかを主観によって評価して下さい。
- 評価基準は、「トピックまで同じ」、「関連するがトピックは異なる」、「関連しない」の 3 段階です。
- ページが消えてしまっているときは、ロボットで収集したときのページを見て判断し、それもないときには「消えている」を選んで下さい。

表 1 に、主観評価をまとめたリストを示す。○ は「トピックまで同じ」を選んだ数を表し、△ は「関連するがトピックは異なる」を選んだ数を表す。−は消えてしまったページの数を表す。精度は、○ を 2 点、△ を 1 点と数えたときの満点 (全ページが存在するときは 20 点) に対する割合を表している。図 1 には、この精度のヒストグラムを示した。

この結果は、3 手法のうちで Companion+がもっともコミュニティ発見手法に適していると判断するに十分なものである。Companion+は、HITS、Companion に比べて平均精度が約 0.3 上回っている。○、△ の個数においても、平均では HITS、Companion を上回っている。ヒストグラムにおいても精度のピークが比較的高いほうにあることが分かる。

Companion+に比べ、HITS、Companion の精度が悪い原因は、Yahoo!等の非常に有名なページに authority スコアの大半を奪われる、トピックドリフトと呼ばれる現象に依るところが大きい。図 1 において、精度 0〜0.2 の範囲

に HITS、Companion のピークがあるのはこのためである。一方、Companion+は、ほとんどトピックドリフトを起こしていない。

この違いは近傍グラフの取り方に原因がある。HITS では、距離 2 以内のページを全て集めるため、その中には有名なページが含まれやすく、トピックドリフトを起こす可能性が高い。Companion では、シードページが Yahoo! などの有名なページを指している場合に FB 集合の中に Yahoo!を指す hub が多く含まれるため、そこでトピックドリフトを起こす。

Companion は、本来 Yahoo!などの非常に有名なサイトをストップ 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、今回の実験ではストップ URL リストを使用していない。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と推測できる。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優れていると言える。

トピックドリフトを起こしていないケースでは、Companion と Companion+は同程度の精度を示している。Companion+ではリンクのウェイト計算を Companion と変えているが、これによる精度の差は現れなかった。したがって、Companion+の精度を良くしている要因は、ほぼ近傍グラフの取り方に依るものであると判断できる。

5.5 コミュニティ群発見アルゴリズムの評価

関連ページ発見アルゴリズムの精度比較の結果から、コミュニティ群発見アルゴリズムの評価では、Companion+を使用した。シードセットの各シードを基に $N = 10$ としてコミュニティ群を生成し、被験者による主観評価を行った。質問項目は以下の通り。

- 各コミュニティのシードになっているページは、最初のシードになんらかの意味で関連していますか？
- 関連していると答えたコミュニティに対して、名前を付けてください。
- コミュニティに含まれる各ページは、そのコミュニティのシードにどの程度関連していますか？(関連ページの評価と同じ基準で)

	ケース数
全コミュニティが関連している (過剰に分割されている)	18(5)
一部のコミュニティが関連していない	4
最初のコミュニティ以外は関連していない	2
合計	24

表 3: コミュニティの関連に関する集計

表 2 に評価の結果を示す。URL の欄には、シードページの URL とそのページのタイトルを示した。関連の欄は、出力されたコミュニティの数に対して、被験者が関連すると答えたコミュニティの数を示している。閾値の欄は、コミュニティ群発見アルゴリズムにおいてコミュニティの分割が起きた際の閾値を示す。コミュニティ名の欄は、被験者が付けた名前である。その他の欄は、各コミュニティの精度評価である。

表 2 から、ほぼ全てのケースで関連するコミュニティ群が生成されていることが分かる。コミュニティ名を見ると多くのケースにおいて、シードに関連し、かつ観点の異なるコミュニティ群が生成されており、興味深い結果となっている。表 3 には、コミュニティ群の関連に関する頻度を示した。全 24 ケースのうち、全てのコミュニティが関連しているケースが 18 ある。これは、リンク解析のみを用いていることを考慮すると十分に良い精度である。

一方、これら 18 のケースのうち、本来は 1 つにまとまるべきコミュニティが過剰に分割されているケースが 5 つある (ケース 12、17、18、19、23)。これら 5 つのケースのうち 4 つについては、閾値が 3 以上となっている。過剰な分割を防ぐ対策としては、閾値に上限を設けることが考えられる。今回の結果からは閾値の上限は 2 が適当であると思われる。ただし上限を設けると、ケース 17、18 においては本来分割されるべきコミュニティが分割されなくなる可能性がある。これはリンク解析のみを用いているための限界であり、本質的な改善を行うにはキーワード解析などを導入する必要があると思われる。

残り 6 つのケースで関連しないコミュニティが出ている原因は、主に最初の Companion+の適用の際に関連しないページが出てきているためである。Companion+の精度は 0.8 程度であり、これもリンク解析の限界に依るものである。

6 まとめと今後の課題

ユーザが与えたシードページから、関連コミュニティ群を発見するアルゴリズムを提案した。本アルゴリズムはリンク解析のみを用いて、意味的に関連のあるコミュニティ

群を発見することができる。また、ユーザテストを行い、本手法の有効性を示した。

ユーザテストの結果から、リンク解析しか用いていないための限界がいくつか観察された。今後の課題としては、キーワードなどのコンテンツ解析を導入することが挙げられる。また、関連ページ発見アルゴリズムの繰り返しによる近傍グラフの拡大を現在は1段しか行っていないが、これをさらに広げる実験を行う予定である。

参考文献

[1] Jeffrey Dean and Monika R. Henzinger. Finding related pages in the World Wide Web. In *Proceedings of the 8th World-Wide Web Conference*, 1999.

[2] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring Web Communities from Link Topology. In *HyperText98*, 1998.

[3] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[4] Sridhar Rajagopalan Ravi Kumar, Prabhakar Raghavan and Andrew Tomkins. Trawling the Web for emerging cyber-communities. In *Proceedings of the 8th World-Wide Web Conference*, 1999.

URL	関連	閾値	コミュニティ名	○	△	−	精度
1. archives.math.utk.edu/popmath.html POPMathematics	1/2	1	Mathematica	4	4	0	0.60
2. home.att.ne.jp/green/asj (社) アフリカ協会	3/5	1	アフリカ開発援助	3	1	0	0.35
			日本政府	10	0	0	1.00
			フランス語圏情報	2	7	1	0.61
3. islamcenter.or.jp/ イスラミックセンター・ジャパン	3/5	1	イスラムと中東	5	2	2	0.75
			大使館	1	3	0	0.25
			外国の事情	1	0	4	0.17
4. lang.nagoya-u.ac.jp/~matsuoka/Japan.html Japan, my Japan! – A Guide to Japan –	3/3	1	Japan	4	2	2	0.62
			Traveling Japan	4	6	0	0.70
			Japanese Kids	4	5	1	0.72
5. plaza.harmonix.ne.jp/~kamao/ 個人利用のためのバーチャルドメイン・サービスのリンク集	3/3	1	ドメインとホスティング	8	0	0	0.80
			検索サービス??	4	0	0	0.40
			HP 素材集	1	8	1	0.56
6. weather.is.kochi-u.ac.jp/ 高知大学気象情報頁	2/2	1	気象	9	0	1	1.00
			官庁	10	0	0	1.00
7. www2e.biglobe.ne.jp/~TKG/ TKG Home Page	2/2	1	パズル	5	1	4	0.92
			Web 上で行うゲーム	4	1	3	0.64
8. www2j.biglobe.ne.jp/~tatuta/ フリーマーケットへ行こう	3/4	1	Fleamarket	8	1	1	0.94
			個人の情報交換サイト	2	7	0	0.55
			懸賞関連が多い?	1	0	0	0.10
9. www3.famille.ne.jp/~s370902/camera/board/index.html カメラの広場	1/4	1	写真	1	7	0	0.45
10. www.alc.co.jp/nihongo/nihongo1.html にほんごセンター	2/2	5	日本語教育	6	2	2	0.88
			better education ? online education ?	8	2	0	0.90
11. www.i-kochi.or.jp/prv/kochi/ Web 高知ホームページ	4/4	1	高知	8	0	2	1.00
			地方新聞	9	0	1	1.00
			県	10	0	0	1.00
			高知の学校	4	0	6	1.00
12. www.isp.ne.jp/~nakajima/index.html いいことがたくさんありますように (映画情報編)	3/3	1	映画	5	4	1	0.78
			映画	1	6	0	0.40
			映画	6	2	1	0.78
13. www.japan.msf.org/ 国境なき医師団 MSF JAPAN	4/4	2	紛争地域などに医療チームを派遣する団体	2	6	2	0.62
			第 3 世界の子供を対象とした NGO	5	3	2	0.81
			医薬団体	3	7	0	0.65
			トルコ地震援助	5	5	0	0.75
14. www.maccentral.or.jp/pokemon/ ポケモンだいすきクラブ	2/2	2	ポケモン	6	4	0	0.80
			ゲーム会社	2	8	0	0.60
15. www.mahjong.or.jp/ Mahjong Walker	2/2	2	麻雀	5	4	1	0.78
			–	1	3	0	0.25
16. www.mars.dti.ne.jp/~o-shin/ RDB 研究館	3/3	1	RDBMS	4	6	0	0.70
			ベンダー	2	0	0	0.20
			Visual Basic	1	5	0	0.35
17. www.ops.dti.ne.jp/~glass/ お気楽株式投資倶楽部	5/5	3	相場情報	8	2	0	0.90
			相場情報	10	0	0	1.00
			新聞社経済情報	10	0	0	1.00
			投資術	10	0	0	1.00
			投資情報	10	0	0	1.00
18. www.panda.org/ WWF International	4/4	3	環境保護団体	4	5	1	0.72
			環境保護団体	3	7	0	0.65
			自然保護団体	1	8	0	0.50
			自然保護団体	2	6	2	0.62
19. www.peugeot.co.jp/ Official Peugeot Japan Web Site	2/2	7	外車	10	0	0	1.00
			外車	10	0	0	1.00
20. www.red-hell.com/ RED-HELL Urawa Red Diamonds Unofficial Site	2/2	1	Urawa	10	0	0	1.00
			J-League	1	9	0	0.55
21. www.spice.or.jp/~mt0711/index.html 初心者のための海外自由旅行への道	2/4	2	Hawaii	10	0	0	1.00
			Travel	4	0	2	0.50
22. www.tintin.com/ 海外旅行好き、集まれ!	3/3	1	海外旅行情報	8	0	2	1.00
			旅行、飛行機情報	8	2	0	0.90
			旅行情報	8	1	1	0.94
23. www.triathlon.or.jp/ Triathlon World	3/3	3	トライアスロン	8	2	0	0.90
			トライアスロン	8	0	2	1.00
			トライアスロン	8	0	2	1.00
24. www.watch.impress.co.jp/pc/index.htm PC Watch	2/2	1	コンピュータ関連ニュース	5	4	0	0.70
			一般ニュース or 報道機関	1	9	0	0.55

表 2: コミュニティ発見の結果