情 報 学 基 礎 61－7
自 然 言 語 処 理 142－7
（2001. 3. 5）

参照共起分析のWebディレクトリへの適用

原田 昌紀 風間 一洋 佐藤 進也
NTT 未来ねっと研究所
東京都武蔵野市緑町 3-9-11

WWWの急速な普及に伴い，Webディレクトリの構築と維持に要するコストは増大しつつあり，登録や更新作業に大幅な遅延をもたらしている．そこで我々はハイパーリンクの参照共起関係に基づく関連Webページ発見アルゴリズムを用いて，Webディレクトリを自動的に拡大する手法を提案する．多数のカテゴリから成る実際のWebディレクトリに対して，4種類のアルゴリズムごとに提案する手法を適用し，その有効性とアルゴリズムによる差違を示す．

Automated Web Directory Expansion using Co-citation Analysis of Hyperlinks

Masanori HARADA, Kazuhiro KAZAMA and Shin-ya SATO
NTT Network Innovation Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo

With the rapid growth of the web, it is getting harder to build and maintain web directories and there are significant delays in registering or updating information. To meet the situation, we propose a new method to automatically expand a web directory using related web finding algorithms based on co-citation analysis of hyperlinks. We apply the proposed method with four algorithms respectively to a web directory in real use which consists of hundreds of categories to show effectiveness of the method and differences among the algorithms.

1 はじめに

近年，インターネット上に公開されたWebページ数は増加の一途を続けており，WWW情報検索サービスの役割はますます大きくなってきている．

WWW情報検索サービスの基本的な実現方式は**サーチエンジン**と**Webディレクトリ**の二種類がある．前者は，ロボットあるいはスパイダーと呼ばれるソフトウェアを用いて大量のWebページを収集し，それらに対する全文検索機能を提供するサービスであり，Google*に代表される．一方，Webディレクトリは Webサイトをその主たるトピックに従って階層的なカテゴリに分類して提示するサービスであり，Yahoo!†が有名である．

Webディレクトリはサーチエンジンと比べ，次の点が優れていると考えられる．

- 通常，サーチエンジンはWebページを検索の単位としているが，Webディレクトリは**Webサイト**単位の検索ができる．ここでWebサイトとは同一の作者によって作成された，意味的なまとまりを持った複数のWebページ群とする．
- 一定水準以上の完成度と信頼性を持ったWebサイトのみが登録される．

*http://www.google.com/

†http://www.yahoo.com/

- Webサイトが階層的に分類されているため，情報要求が不明瞭で，適切な検索条件式を入力できない場合でも，ブラウジングによる検索ができる．

しかし，Webディレクトリでは Webサイトの収集，審査，分類等の作業を人手でおこなうため，その構築と保守に要するコストが大きいことが問題となる．そこで，我々はサーチエンジンのロボットが収集した大量の Webページの中から，Webディレクトリの各カテゴリに関連した Webサイトを自動的に発見することで，Webディレクトリの管理を省力化する方法を提案する．この方法はテキストの自動分類とは異なり，任意の Webサイトをいずれかのカテゴリに分類することはできないが，各カテゴリに登録する価値のある重要な Webサイトが発見されやすいという特長がある．

本研究では，ハイパーリンクによる参照共起関係を用いて，関連 Webサイトを発見するアルゴリズムを4種類実装し，実際の Webディレクトリに適用して，その精度を評価する．

以下，第2節では，関連研究を述べる．第3節では，提案する Webディレクトリ拡大手順を述べる．第4節では，関連 Webサイト発見アルゴリズムの詳細を述べる．第5節では，評価実験について述べる．最後にまとめを述べる．

2 関連研究

本節では，Webディレクトリ構築の自動化に関連する先行研究を，ハイパーリンクによる参照共起関係を利用した方法と，テキストの内容に基づく自動分類を拡張した方法に大別して述べる．

2.1 リンクによる参照関係の分析

Kleinberg による HITS[1] の提案以来，Webページ間のハイパーリンクによる参照関係の分析と，その WWW 情報検索システムへの応用が盛んに研究されている．本節では 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を **Web グラフ**と呼ぶ．

HITS は，WWW に特化した検索方法であり，与えられたキーワードによってサーチエンジンで検索を行ない，その検索結果上位の近傍の Web グラフから，あるトピックに関する**オーソリティ**とハブを抽出する．ここでオーソリティとは Web グラフ中の多くのハブから参照されている Webページであり，高い評価が得られている Webページに相当する．一方，ハブとは Web グラフ中で多くのオーソリティを参照する Webページであり，リンク集などに相当する．HITS では，反復的な計算によって，Web グラフ中の各 Webページのオーソリティスコアとハブスコアを計算する．

HITS にはトピックドリフトが起こりやすいという問題がある．トピックドリフトとは，被参照数が非常に大きく，トピックとあまり関連しない Webページが Web グラフ中にあった場合に，元々のトピックとあまり関係ないオーソリティとハブが現れる現象のことである [2]．

ARC はアンカーテキストを利用することで HITS の精度を改善した方法であり，トピックによっては，Web ディレクトリに登録されている人手で選別された Webサイトに匹敵する品質の Webサイトを発見できる [3]．しかし，Webディレクトリの詳細に分類されたカテゴリの中には，トピックを端的なキーワードで表現することが難しいカテゴリも多い．

一方，トピックを表わすキーワードからオーソリティを検索するのではなく，トピックのオーソリティとなる Webページのアドレスを入力し，その近傍の Web グラフから関連したオーソリティを発見する方法として，Dean らによる Companion と豊田による Companion+ がある [4][5]．本研究ではこの二つのアルゴリズムを拡張し，Webディレクトリの各カテゴリに登録されている Webサイト群に関連したオーソリティを発見することで，キーワード等を一切用いることなく，詳細なカテゴリ分類を持つ Webディレクトリを自動的に拡大する方法を検討する．

本研究と類似したアプローチとして，村田によるコミュニティ発見方法の研究がある [6]．村田の方法では，あるトピックに関してオーソリティとなる Webページ群をシードセットとして入力し，それらすべてを同時に参照する Webページをサーチエンジンを利用して検索し，それらすべてが参照している Webページを再びシードセットとする．この操作を反復して得られる完全二部グラフを，あるトピックについてのコミュニティとみなす．だが，この方法では，シードとなる Webページ数が多い場合 (たとえば6以上) や，被参照数が小さい Webページがシードセットに含まれる場合には，コミュニティを一つも発見できないことが多い．また，HITS 同様のトピックドリフトが起きる．こうした理由から，Webディレクトリの拡大には適していない．

2.2 ハイパーテキストの自動分類

Webディレクトリの拡大を自動化する手段としては，テキストの自動分類の利用も考えられる．テキストの自動分類とは，テキストをあらかじめ決められたカテゴリに分類する，あるいはテキストに複数のカテゴリを付与することをいう [7]．テキストおよびカテゴリが含む語句の生起頻度を変数として，ある一定のモデルの下で類似度を計算する方法が一般的である．

しかし，Webページにはさまざまな言語，表現が用いられ，語彙が統制されていない上に，一般に個々のWebページのテキストサイズは小さいため，従来の方法だけで，高い精度で詳細に分類することは困難である．

そのため，ハイパーテキストの自動分類の精度を改善する方法として，近傍のテキストの分類結果をインクリメンタルに適用する研究がある [8][9]．しかし，これらの研究では少数の大まかなカテゴリへの分類しか評価されておらず，Webディレクトリの多数の詳細なカテゴリへの分類の効果は明らかではない．

そこで本研究では，語句の生起頻度などを用いない，参照関係のみを利用したWebディレクトリの拡大方法を検討し，その有効性を明らかにすることを目的とする．

3 Webディレクトリの拡大

3.1 本研究におけるWebグラフの扱い

本研究ではWebディレクトリの処理単位に合わせて，Webサイトを一つの点と見なしたWebグラフを作成する．Webサイトを明確に定義することは難しいので，本稿では同一サーバの同一パス上に存在するWebページ群は一つのWebサイトを構成するものとする．すなわち，WebページのURLの最後の/より前の部分文字列が等しい場合に同じWebサイトに属するとする．たとえば`http://www.ntt.co.jp/product/index.html`と`http://www.ntt.co.jp/product/`は同じ点と見なされる．

また，異なるWebサーバ上に存在するWebサイト間のハイパーリンクのみを辺として用いる．これは同一Webサイト内では，参照元と参照先の内容上の関連以外の理由からリンクが作成されることが多いことと，同一Webサイト内からの参照数が大きくてもWebサイトの価値が高いとは言えないためである．

さらに辺に付随する情報として，ハイパーリンクのWebページ中での生起順序も保存する．これは第4節で述べる参照共起関係を調べるために使用する．

3.2 Webディレクトリ拡大手順

Webディレクトリの拡大手順は次の通りである．

ステップ1 大域Webグラフを作成する．

ステップ2 カテゴリごとに関連Webサイト発見アルゴリズムを適用する．

ステップ3 発見されたWebサイトから重複を除き，カテゴリごとに関連度の高いものを出力する．

3.3 ステップ1: 大域Webグラフの作成

Webディレクトリが検索対象とする範囲に存在するWebページを，ロボットを用いてなるべく多く収集し，それらの参照関係をデータベースに格納し，Webグラフとして表現する．これを**大域Webグラフ**と呼ぶ．

3.4 ステップ2: 関連Webサイトの発見

Webディレクトリのカテゴリごとに，登録されているWebサイト群をシードセットとして，第4節で述べる関連Webサイト発見アルゴリズムを適用し，大域Webグラフの中から関連度の高いWebサイトを発見する．

関連度とはHITSにおけるオーソリティスコアと同様の指標であり，カテゴリの主なトピックに適合するWebサイト群の中で，高い評価を得ているWebサイトほど大きくなるように，関連Webサイト発見アルゴリズムごとに定義する．ただし，オーソリティスコアとは異なり，別々のシードセットを元に計算された値が比較できるようにする．

3.5 ステップ3: 重複の除去

複数のカテゴリに重複しているWebサイトは，関連度が最大となったカテゴリのみに残す．また，登録済みのWebサイトが発見された場合には削除する．

これは一般的なWebディレクトリでは，原則として一つのWebサイトは一つのカテゴリのみに登録するためである．また，トピックドリフトが発生して，被参照数の大きいWebサイトが多くのカテゴリで発見されても，その影響が限定される効果がある．

4 関連 Web サイト発見アルゴリズム

関連 Web サイト発見アルゴリズムは，一つ以上の Web サイトをシードセット S として受け取り，S 中の Web サイトと**参照共起関係**にある Web サイトを関連 Web サイト p を発見し，その関連度 $R(p)$ を計算する．ここで，ある二つの Web サイトが参照共起関係にあるとは，その両者を参照する Web ページが存在し，その Web ページにおける二つの Web サイトへのハイパーリンクの生起順序の差があらかじめ定められた定数 L 以内となることとする．

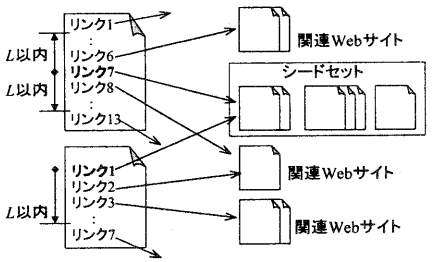

図 1: 参照共起関係による関連 Web サイト発見

従来，与えられた Web ページに関連する Web ページを発見するアルゴリズムとして，Dean らによる Companion と Cocitation，豊田による Companion+ などが提案されてきたが [4][5]，シードセットを入力する方法は明確でなく，また，別々のシードを元に計算された関連度を比較することができなかった．

そこで，本研究ではこれらを拡張した関連 Web サイト発見アルゴリズムとして，**Companion+**，**Companion++**，**Cocitation++**，**MultiCocitation** の 4 つを比較検討する．

4.1 Companion+

Dean らによる Companion は，HITS を関連 Web ページ発見用に特化させたアルゴリズムであり [4]，豊田による Companion+ は Companion の精度を改善したアルゴリズムである [5]．本節で述べる Companion+ ではオリジナルの Companion+ を複数のシードへ対応させたほか，ストップリストの導入，オーソリティスコアの関連度への変換という変更を加えており，次の 5 ステップからなる．

ステップ 1. 近傍 Web グラフの作成　大域 Web グラフにおいて，シードセット S 中の Web サイトから，ハイパーリンクを逆方向に一回たどり，そこから S 中の Web サイトと参照共起関係にある Web サイトへのハイパーリンクのみを順方向に一回たどる．ただし，シードの被参照数が定数 B を超える場合には，ランダムに選択された B 個のリンクを逆にたどる．そして，この間に通る Web サイトおよびハイパーリンクを点と辺として，近傍 Web グラフ G を作成する．すなわち，G は二部グラフとなる．実験では $L = 5$，$B = 2000$ とした．

ステップ 2. ミラーサイト等の削除　近傍 Web グラフ G において，2 つの Web サイトのリンク先が 80% 以上一致する場合には，被参照数が小さいものを削除する．これはミラーサイトや，雛形をもとに機械的に生成された Web ページによって，それらのリンク先となる Web サイトのオーソリティスコアが不正に高くなるのを防ぐためである．

また，大域 Web グラフにおいて被参照数が大きい上位 100 個の Web サイトをストップリストとし，このリストに含まれる Web サイトは近傍 Web グラフから削除する．これは被リンク数が極めて大きい Web サイトは，サーチエンジンなど多くのトピックに関連する Web サイトであることが多く，トピックドリフトの原因となるためである．

ステップ 3. 各リンクのウェイト決定　近傍 Web グラフに含まれるハイパーリンクに，0 ～ 1 の範囲でオーソリティウェイトおよびハブウェイトを与える．まず，シード s を参照するハイパーリンク l_s のオーソリティウェイトは 1 とし，それと生起順序の差が L 以内にあるハイパーリンクのオーソリティウェイトは $(L - |l_s$との生起順序の差$|)/L$ とする．それ以外のハイパーリンクのオーソリティウェイトは 0 とする．一方，ハブウェイトはすべて 1 とする．

ただし，ある Web サイトが同じサーバ上にある n 個の Web サイトから参照されている場合には，それらのハイパーリンクのオーソリティウェイトを $1/n$ 倍する．同様に，ある Web サイトが同じサーバ上にある n 個の Web サイトを参照している場合，それらのハイパーリンクのハブウェイトを $1/n$ 倍する．

ステップ 4. オーソリティスコアとハブスコアの計算　オーソリティスコアのベクトル A およびハブスコアのベクトル H を 1 で初期化し，次の手順を収束するまで繰り返す．ここで $W_a[p,q]$ は Web サイト p から Web サイト q へのハイパーリンクのオーソリティウェイト，$W_h[q,p]$ は q から p へのハイパーリンクのハブウェイトである．

1. $A[q] := \sum_{p \in \{x|x \to q\}} W_a[p,q] \times H[p]$
2. $H[q] := \sum_{p \in \{x|q \to x\}} W_h[q,p] \times A[p]$
($p \to q$はpからqへのハイパーリンクを示す．)
3. Aの要素の二乗和が1になるよう正規化する．
4. Hの要素の二乗和が1になるよう正規化する．

ステップ5. オーソリティスコアの変換　ステップ4で得られたオーソリティスコアは，近傍Webグラフが大きいほど小さい値となる．これを次式により近傍Webグラフの大きさを反映するよう変換して関連度とする．

$$R(p) = A[p]^2 \times |G|$$

4.2 Companion++

Companion++では，シードセットの要素となるWebサイト一つずつに，前節のCompanion+アルゴリズムを適用し，それぞれ関連Webサイトを求める．そして，Webサイトごとに関連度の総和を求め，カテゴリ全体での関連度とする．

Companion+と比較すると，一つのシードあたりの近傍Webグラフのサイズは小さくなるため，トピックドリフトが発生しにくいと期待される．また，仮にトピックドリフトが発生しても，カテゴリ全体の関連度に与える影響は小さくなる．

4.3 Cocitation++

Cocitation++は，Deanらによる Cocitation[4]を，シードセットを入力とするように拡張したアルゴリズムである．

まず，Companion+のステップ1とステップ2により近傍WebグラフGを得る．ここで，Gにおいて，あるWebサイトpが，Webサイトrを介してシードsと参照共起関係にあることを$r \Rightarrow (p,s)$と書くことにする．Companion++はシードセットS中の一つ以上のシードと参照共起関係にあるWebサイトpを関連Webサイトとし，その関連度$R(p)$を次のように定義する．

$$R(p) = \sum_{s \in S} |\{r|r \Rightarrow (p,s)\}|$$

4.4 MultiCocitation

MultiCocitationはCocitation++と同様，Gにおいてシードセットと参照共起関係にあるWebサイトを関連Webサイトとするが，多くの異なるシードと参照共起関係にあるWebサイトほど関連度を高くする．これはCocitation++では，シードセットに被参照数が突出して大きいものが含まれていた場合に，そのWebサイトのみと関連したWebサイトでも高い関連度を得てしまうという問題を回避するためである．

ただし，参照共起関係にあるシードの異なり数が等しい場合には，参照共起の起きる回数の総和が大きいほうが高い関連度を得るとする．つまり，MultiCocitationでは関連度$R(p)$は次のように定義する．

$$R(p) = |\{s|r \Rightarrow (p,s)\}| + \alpha \times \sum_{s \in S} |\{r|r \Rightarrow (p,s)\}|$$

αは定数であり，実験では$\alpha = 0.1$とした．

5 評価実験

実際のWebディレクトリとロボットで収集したWebページのデータを用いて，提案するWebディレクトリ拡大手順の有効性を，4種類の関連Webサイト発見アルゴリズムをそれぞれ採用した場合について評価した．

5.1 Webディレクトリ

拡大対象のWebディレクトリとして，Netscape Communications社のOpen Directory[‡]の一部を利用した．Open Directoryはボランティアとして参加した多数の編集者の共同作業によって構築されており，カテゴリ名，Webサイトの URL，表題，要約などのデータを公開している．

実験ではこのデータから，日本語Webサイトを対象とした/World/Japanese以下のカテゴリのみを抜粋し，一つのWebディレクトリと見なして利用した．2000年12月現在，これらのカテゴリのうち，一つ以上のWebサイトが登録されているカテゴリは702個あり，登録されているユニークなWebサイトは6,143URL であった．この中には専任の編集者が存在し，重要なWebサイトが多数登録されているカテゴリがある一方，編集者が存在せず，あまり重要でないWebサイトばかりが登録されているカテゴリもある．

5.2 大域Webグラフ

2000年12月末にロボットによるWebページの収集をおこなった．前節で述べたWebディレクトリのデータ6,143URL を起点とし，JPドメインに

[‡]http://dmoz.org/

表 1: 大域 Web グラフの概要

取得された Web ページ:	11,268,680
抽出されたハイパーリンク:	120,299,318
異なる Web サーバ間のリンク:	20,957,220
終点が取得済のリンク (= 辺):	13,522,961
辺の起点となる Web サイト:	805,004
辺の終点となる Web サイト:	1,101,987

存在する Web ページ，あるいは漢字・ひらがな・カタカナを含むアンカーテキストによってリンクされている Web ページを収集対象とした．収集された Web ページを元にして作成した大域 Web グラフの概要を表1に示す．

5.3 実験 1: 精度の評価

実験 1 では Web サイトが適切なカテゴリで発見されるかどうかを評価した．

まず，Web ディレクトリの 702 個のカテゴリのうち，登録されている Web サイト数が 4 以上のカテゴリ 474 個から，Web サイトを 1 つずつ無作為に選択して，評価用 Web サイトとし，それらを除いた Web ディレクトリを対象とした．

そして，第4節で述べた 4 種類のアルゴリズムを適用し，各カテゴリにおいて関連度が高い Web サイト最大 N 個に対して，次のように定義される精度を求めた．

$$\text{精度} = \frac{|\text{元のカテゴリで発見されたサイト}|}{|\text{発見された評価用 Web サイト}|}$$

ただし，ここでいう精度は元々人手によって登録されていた Web サイトが発見された場合のカテゴリの正しさしか保証していない．

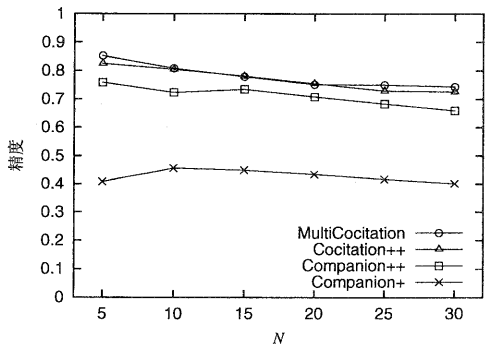

図 2: 関連 Web サイト上位 N 件の精度

$N = \{5, 10, 15, 20, 25, 30\}$ とした場合の精度を図2に示す．MultiCocitation と Cocitation が優れており，各カテゴリで 10 個の Web サイトを発見した場合，0.8 以上の精度でカテゴリに適した Web サイトが発見されている．これは分類先のカテゴリ数が 700 個以上に及ぶことから，良好な結果であると考えられる．また，元のカテゴリとは異なるカテゴリで発見される場合でも，「ビジネス / 食品 / 飲料 / 酒類」に分類されていた Web サイトが「レクリエーション / グルメ・ドリンク / 酒類」で発見されるなど，関連したカテゴリで発見される傾向があった．

しかし，Companion+ は高い精度を得ることができなかった．これは複数のシードを起点とした場合，近傍 Web グラフのサイズが大きくなることから，元々のトピックより一般化されたトピックに関するオーソリティが発見されやすくなり，トピックドリフトが起きるためと思われる．Companion++ では，トピックドリフトの影響は抑えられたものの，より単純なアルゴリズムである Cocitation を上回る精度は得られなかった．

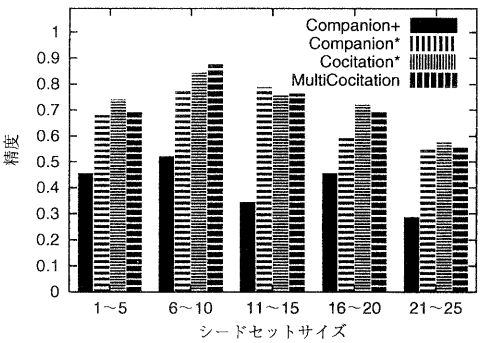

図 3: シードセットのサイズと精度 ($N = 20$)

次にシードセットの大きさによる精度の違いを図3に示す．Companion++, Cocitation++, MultiCocitation ではシード数が 10 前後のカテゴリで精度が最大となり，シード数がさらに増加すると，逆に精度が低下する傾向にある．この原因は，これらのアルゴリズムでは，シード数が大きいカテゴリほど高い関連度が得られやすいためと思われる．対策としては，Web ディレクトリ拡大手順のステップ 3 を変更し，シード数の比に合わせて，関連 Web サイトをカテゴリに割り当てる方法が考えられる．

5.4 実験 2: 適合度と重要度の主観評価

実験 2 では各カテゴリで発見された関連 Web サイトが，カテゴリに対してどの程度適合しているか，また Web ディレクトリに登録するだけの価値

を有しているかを被験者に判断してもらった．被験者はネットワーク分野の研究者8名であり，日常的にWWWを利用している．

まず，被験者に702個のカテゴリから，それぞれよく知っている分野のカテゴリを2～4個挙げてもらった．それらから重複を除き，表2に挙げるカテゴリを評価用とした．

表 2: 評価に用いたカテゴリ

略号	カテゴリ名 (シードセットの大きさ)
A	アート / 音楽 / 海外.../ イギリス / ビートルズ (3)
B	ショッピング / アウトドア用品 (4)
C	科学 / 自然科学 / 天文と宇宙 / 天体写真と画像 (5)
D	スポーツ /.../ オリンピック /2000_シドニー (5)
E	地域 / 地方自治体 / 神奈川 (5)
F	健康 / 食事と栄養 (8)
G	アート / 映画 / 洋画 (9)
H	レクリエーション / グルメ.../ 酒類 / ワイン (10)
I	家庭 / 料理 / 食材 (10)
J	各種資料 / 辞書・事典 (10)
K	社会 / 時事 / 自然災害 (11)
L	ビジネス / 情報産業 /.../ 携帯電話と PHS(13)
M	ゲーム / ビデオゲーム / アドベンチャー (15)
N	ニュース / 新聞 (19)
O	レクリエーション / 車・バイク (28)
P	コンピュータ /.../WWW/ ホームページ検索 (32)

被験者に関連Webサイト発見アルゴリズムの違いを意識させないため，評価用のカテゴリごとに，4種類のアルゴリズムで関連度が高い10個のWebサイトを求め，それらをマージしたリストを作成した．リストでは，Webサイトの表題や要約文は提示せず，Webサイトを実際に閲覧して判断してもらった．指示内容は次の通りである．

- このカテゴリにはリスト1に挙げたURLのWebサイトが登録されています．また，リスト2に挙げたURLのWebサイトが登録候補となっています．
- あなたの仕事はそれぞれのリストから，このカテゴリに関連した内容を含み，Webディレクトリに登録するに値するWebサイトを選別することです．
- 登録すべきかどうかは知名度，信頼性，情報の豊富さ，オリジナリティ，デザインなどから判断します．
- まず，リスト1のWebサイトを一つずつ閲覧し，カテゴリに登録するべきWebサイトであるかどうかを「重要でぜひとも登録すべきサイトである」「どちらかといえば重要で、登録すべきサイトである」「どちらかといえば重要ではなく、登録すべきでないサイトである」「重要ではなく、登録すべきでないサイトである」の4段階で評価してください．
- 次に，リスト2のWebサイトを一つずつ閲覧し，Webサイトの内容がカテゴリに関連しているかを，「関連している」「どちらかといえば関連している」「どちらかといえば関連していない」「関連していない」の4段階で評価してください．また，「関連している」あるいは「どちらかといえば関連している」Webサイトについては，リスト1と同様に価値を評価してください．
- アクセスできないWebサイトや，リストに既出のWebサイトの一部となるWebサイトは判断不能を選択してください．

まず，発見されたWebサイトとカテゴリの関連の4段階評価をそれぞれ+2点,+1点,−1点,−2点，判断不能を0点とし，カテゴリごとに関連度上位10個のWebサイトの平均点を適合度として，集計した結果を図4に示す．

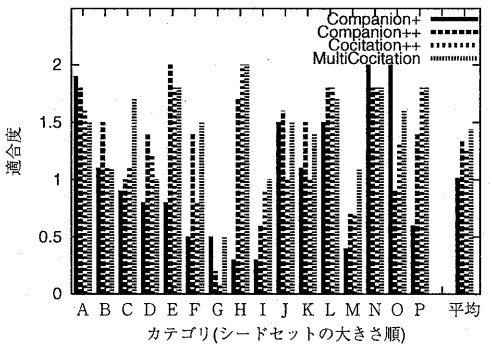

図 4: 関連度上位10サイトの適合度

全カテゴリの平均では，実験1と同様にCompanion+を除く3つのアルゴリズムが高い適合度を達成し，MultiCocitationが約1.44で最高となった．しかし，これら3つのアルゴリズム間の差違は比較的小さく，むしろカテゴリによって，適合度が異なる傾向が見られる．

一つの要因として，専任の編集者が存在しないなどの理由で，被参照数の大きいWebサイトが登録されていないカテゴリでは，近傍Webグラフが小さくなり，関連Webサイト10個を発見するのに必要な情報が不足することが考えられる．図5にMultiCocitationを用いた場合の，近傍WebグラフGのサイズと適合度の関係を示す．適合度が低くなるのは，近傍Webグラフが小さいカテゴリに限られることがわかる．

別の要因としては，ハブとなるリンク集における分類基準と，Webディレクトリのカテゴリのミスマッチが考えられる．たとえば，「アート / 映画 / 洋画」カテゴリの適合度が低いのは，邦画と洋画を区別しないリンク集が多いためと思われる．

次に，シードセットと関連Webサイトの重要さの評価結果を，やはり4段階評価をそれぞれ+2点,+1点，−1点，−2点，判断不能を0点として集計し，その平均点を**重要度**とした．その結果，全カテゴリのシードセットの重要度の平均は1.0となり，関連Webサイトの重要度の平均は，Companion+で0.96，Companion++で0.88，Cocitationで0.86，

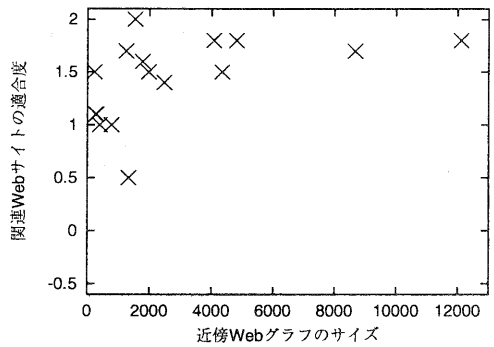

図 5: 近傍 Web グラフのサイズと適合度 (MultiCocitation)

MultiCocitation で 0.74 となった.

Companion+ で高い重要度が得られたのは，大域 Web グラフにおいて被参照数が大きい Web サイトほど高い関連度を得やすいためと思われる．これらはトピックドリフトの原因となる反面，トピックに適合している場合には，重要な Web サイトとなる．

一方，MultiCocitation で高い重要度が得られなかったのは，シードセットに重要でない Web サイトが多く含まれる場合，それらを参照するハブは，やはり重要でない Web サイトを多く参照することが多いためと推測される．MultiCocitation を用いた場合の，シードセットの重要度と関連 Web サイトの重要度の関係を図6に示す．この図からシードセットの重要度には大きなばらつきがあることがわかる．また，正の相関 (相関係数 0.64) があり，重要な Web サイトが登録されているカテゴリほど，発見される Web サイトの重要度が高い傾向がある．

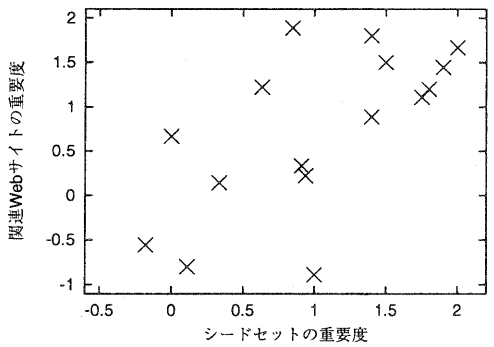

図 6: 重要度の相関 (MultiCocitation)

ただし，今回の実験では被験者による判断基準の違いを補正していない．より厳密な評価は今後の課題である．

6 おわりに

ハイパーリンクの参照共起関係を利用した Web ディレクトリの自動的な拡大方法を提案した．また，4 種類の関連 Web サイト発見アルゴリズムを，多数のカテゴリを持つ Web ディレクトリに適用して，それらを比較した．

今後は，パラメータの調整や，カテゴリの階層関係の利用によって適合度と重要度の向上を図る．また，Web ディレクトリ拡大の完全自動化に向けて，Web サイトの簡潔な要約を提供する方法についても検討していきたい．

謝辞

Open Directory のデータを公開されている Netscape Communications 社と，ボランティア編集者の方々，そして評価実験にご協力いただいた皆様に感謝いたします．

参考文献

[1] J. Kleinberg: "Authoritative sources in a hyperlinked environment," Proc. of 9th ACM SIAM Symposium in Discrete Algorithms, pp.668–677, 1998.

[2] Krishna Bharat, et al.: "Improved Algorithms for Topic Distillation in a Hyperlinked Environment," Proc. of the 21st ACM SIGIR Conf., pp.104–111, 1998.

[3] Soumen Chakrabarti, et al.: "Automatic resource compilation by analyzing hyperlink structure and associated text," Proc. of the 7th World Wide Web Conf., pp.65–74, 1998.

[4] Jeffrey Dean, et al.: "Finding Related Pages in the World Wide Web," Proc. of the 8th World Wide Web Conf., 1999.

[5] 豊田正史: "WWW における関連コミュニティ群の発見," 情報処理学会データベースシステム研究会報告 DBS122-40, pp.307–314, 2000.

[6] 村田剛志: "Web におけるコミュニティの発見," 第 47 回人工知能学会知識ベースシステム研究会資料, SIG-KBS-9904, pp.79–84, 2000.

[7] 徳永健伸: "情報検索と言語処理," 言語と計算 5, 東京大学出版会, 1999.

[8] Soumen Chakrabarti, et al.: "Enhanced Hypertext Categorization using Hyperlinks," Proc. of the Intl. Conf. on ACM SIGMOD'98., 1998.

[9] Hyo-Jung Oh, et al.: "A Practical Hypertext Categorization Method using Links and Incrementally Available Class Information," Proc. of the 23rd ACM SIGIR Conf., pp.264–271, 2000.